DEVELOPMENT AND APPLICATION OF NOVEL BIOINFORMATICS TOOLS FOR PROTEIN FUNCTION PREDICTION

A thesis submitted for the degree of

Doctor of Philosophy

by

Danielle Allison Brackenridge

Faculty of Life Sciences
School of Biological Sciences
University of Reading

May 2021

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Danielle Allison Brackenridge

Date: 9th May 2021

Acknowledgement

First and foremost, I'd like to thank God, my lord, my saviour who answers all my prayers in one way or another.

I would like to express my deep and sincere gratitude to my supervisor, Professor Liam McGuffin for giving me the opportunity to achieve a dream I had since watching Jurassic Park as a seven-year-old and finding the science utterly fascinating. Thank you for your invaluable guidance throughout the past six years. Your motivation and commitment to your students is truly inspiring and has been the best part of my PhD journey.

Thank you to Dr D Roche who taught me the methodology to carry out my research and pointed me in the right direction a fair few times, especially when I was a newbie to bioinformatics and had no idea what was going on. Dr R Adiyaman you are the whizz at fixing problems. Thank you for your patience. Nicholas Edmunds thank you for your valuable input into my research and for responding to panicked Google Hangout messages during the weekends.

Helen "Original" Jones, not quite sure how I managed to convince you to read my thesis but a heartfelt thanks for correcting my many typographical and grammatical mistakes when I couldn't see the wood for the trees and for being a part of this journey with me.

I am extremely grateful for my mum. Your love, prayers, care and sacrifices have made me the woman I am today. None of this would be possible without you and I truly wish your mother, my Grandmother, was here to see the woman you raised and to share this joy with us.

The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the true goal of education

Martin Luther King Jr

Table of Contents

List of Tables

List of Figures

List of Equations

Abstract

Proteins are essential molecules with a diverse range of functions. Current experimental methods such as X-ray crystallography and Nuclear Magnetic Resonance are costly and time consuming. Additionally, the gap between sequence and structure stands at 1/500th, as sequences of proteins are easier to deduce then the overall tertiary structure. As a result, *in silico* methods of structure prediction can assist in bridging this gap. Identifying the role of proteins *in silico* starts with the amino acid sequence, the blue print for all proteins. In turn, amino acid sequence is pivotal for determining the three-dimensional structure, which is vital for the protein to function adequately. Once the structure of a protein is known, ligands which interact with the protein can be deduced by investigating ligand-binding interactions and ultimately function can be determined. However, *in silico* methods for function prediction are not without problems which are (1) can function be predicted from structure and (2) can prediction of ligands and ligand-binding site residues provide insight into a protein's function. Finally, there is a need to validate prediction results to assess the state-of-the-art and separate the known from the unknown.

FunFOLD3 developed by the McGuffin group, is a template-based method for protein-ligand prediction and works on the premise that proteins with similar folds are likely to bind the same ligands. FunFOLDQ, also developed by the McGuffin group provides insight into protein function by prediction of Gene Ontology (GO) terms. Two experimental competition methods will be used to objectively measure the results from FunFOLD3 and FunFOLDQ, the biennial Critical Assessment of Techniques for Protein Structure Prediction (CASP) and Critical Assessment of protein Function Annotation (CAFA), respectively. The ninth CASP competition in 2010 saw the introduction of ligand-binding site prediction, with this category subsequently becoming a function prediction category in the 10th CASP competition in 2012. The 11th, 12th and 13th CASP competitions in 2014, 2016 and 2018, respectively provided an extensive source and range of proteins in the prediction of ligands and ligand-binding site residues. The third CAFA competition benchmarked FunFOLDQ in the prediction of GO terms. The 3D structure models obtained from CASP and GO terms predicted from CAFA will assist with the benchmarking of FunFOLD3 and FunFOLDQ to determine what improvements needed be made (i.e. identifying the problem) and will also provide an objective measure of the predictions. If closely aligned 3D structure models can be obtained, then FunFOLD3 could be used in the determination of novel or poorly annotated proteins.

In order to objectively measure the binding site quality of FunFOLD3, two scoring metrics will be used; Matthew's Correlation Coefficient (MCC) and Binding Distance Test (BDT) score developed by the McGuffin group. Matthew's Correlation Coefficient is a special case of Pearson Correlation Coefficient and provides a value between -1 to 1, with -1 being a total negative correlation, 0 is no correlation and 1 is a total positive correlation based on the observed and predicted ligand-binding site residues. Scores of 0.40 to 0.69 are strong positive relationships and 0.70 and higher are strong positive relationships. The downside of MCC is that it does not take into consideration the overall 3D structure of the protein model. Therefore, BDT will also be utilised as this score, which is also scored from -1 to 1, to take into consideration the 3D structure. Both MCC and BDT are only possible to produce when there is an observed (actual) structure available with bound ligands to compare against the predicted structure and hence why MCC and BDT are objective measures of ligand-binding site prediction. The average MCC and BDT score from CASP11 was 0.42 and 0.51, respectively. CASP12 saw the prediction of ligands for low annotation level proteins with no known ligands, demonstrating the potential use of FunFOLD3 in novel protein prediction. The average MCC and BDT score from CASP13 was 0.47 and 0.53. CAFA3 showed FunFOLDQ can be used in the prediction of GO terms, however further refinements are needed to increase specificity of the term predictions. The development option this thesis has explored is the use of docking (preferred orientation of interacting partners) with AutoDock Vina to improve the accuracy of ligand-binding residues by FunFOLD3, as the problem with TBM methods can be that predicted ligand(s) from a similar template will be forced to fit within the ligand-binding pocket. However, with docking, the aim of this method is to predict the preferred orientation of the ligand within the ligand-binding space. Utilisation of docking has also added to the novelty of this research, as different grid box calculations around the ligand-binding space was explored, with varying degrees of success with each grid box calculation. Examples of two CASP targets which had improvements in MCC and BDT score following docking were CASP11 target T0783 (2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase) the MCC and BDT scores by FunFOLD3 were 0.17 and 0.21, respectively. Following docking the MCC and BDT scores increased to 0.63 and 0.45, respectively. CASP13 target T1016 (alpha-ribazole-5'-P phosphatase) had MCC and BDT scores of 0.556 and 0.646 by FunFOLD3, respectively. Following docking the MCC and BDT increased to 0.85 and 0.91, respectively.

Lastly, CASP_Commons, a community-wide experiment to find the consensus structures, explored the role of FunFOLD3 with predicting ligands and ligand-binding sites for the novel protein and proteins domains of SARS-CoV-2. The protein domains were non-structural proteins 2, 4 and 6, open reading frames 3a, 6, 7b, 8 and 10, membrane protein and papain-like protease. FunFOLD3 predicted ligands for ten of the protein domains, of which there were a total of 32 targets due to domains being split into smaller residues and subsequent rounds of 3D modelling improvement.

Increased understanding of protein structures can provide further insight into a protein’s function, particularly if ligands are bound and identified, an example in this thesis is the prediction of chlorophyll A for non-structural protein 4 (nsp4). Chlorophyll A, like haemoglobin is a porphyrin ring and templates related to nsp4 show a role in blood clotting.

Therefore, whilst chlorophyll A might not be the exact ligand, similarities between haemoglobin and chlorophyll A can clearly be determined and assist in understanding the role of nsp4 in the pathology of COVID-19. Identification of GO terms can provide more detailed understanding into the function or functions of proteins and, in proteins with limited annotation information this can assist with comprehending their role.

This thesis has focused on improving and developing a function prediction method, FunFOLD3, to better understand the role and function of proteins. The new method of FunFOLD3 which utilises docking will be integrated into the McGuffin group prediction servers and will be benchmarked in subsequent CASP competitions, to critically assess the performance of the developed method.

Chapter 1: Introduction

1.1 Protein structure

Proteins are essential molecules involved in a wide variety of essential intra- and inter-cellular activities. The particular activities include, but are not limited to; maintaining cellular defences; enzymatic catalysis; metabolism and catabolism; maintenance of the structural integrity of cells and signalling within and between cells (Du *et al.*, 2016). A protein can be identified on each level of its structure and every protein will contain at least a primary, secondary and tertiary structure with only multimeric proteins (e.g. haemoglobin) having a quaternary structure (Sanvictores & Farci, 2020). Examples of the four different levels of protein structure is shown in Figure 1.1 below.

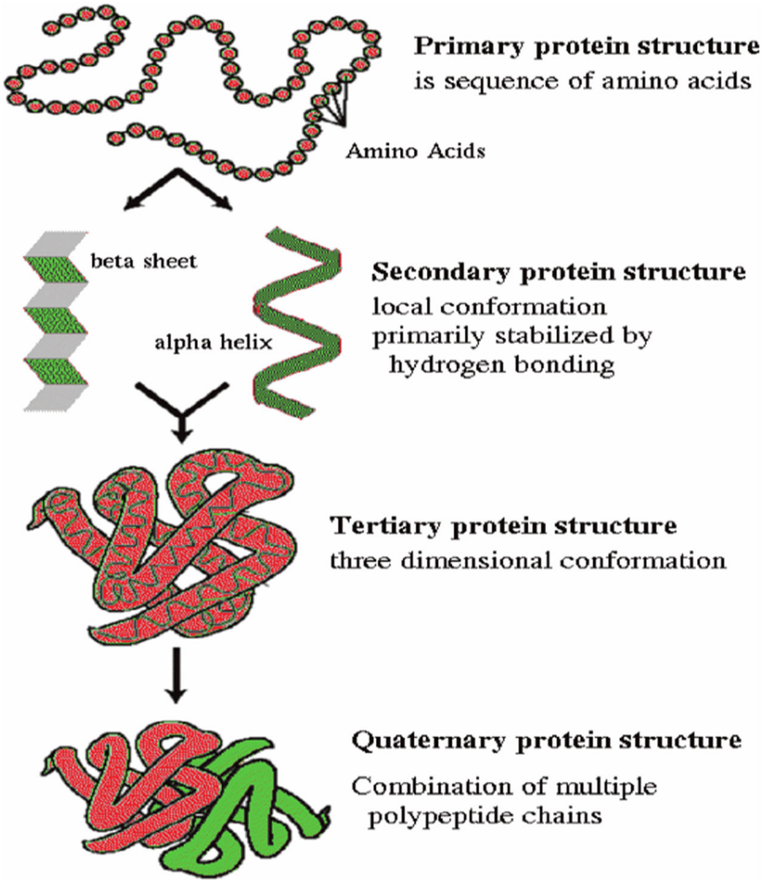

Primary protein structure
is sequence of amino acids
Amino Acids
beta sheet
alpha helix
Secondary protein structure
local conformation
primarily stabilized by
hydrogen bonding
Tertiary protein structure
three dimensional conformation
Quaternary protein structure
Combination of multiple
polypeptide chains

Figure 1.1. Levels of protein structure
The four different levels of protein structure are depicted above. Briefly, primary structure is the amino acid sequence. Secondary structure is the local conformation. Tertiary structure is the 3D structure and quaternary structure is the combination of independently formed tertiary structures. Figure taken from Patel & Shah, 2013.

Protein structure is hierarchical and polypeptide chains fold locally to form α-helices and β-strands which then combine to form tertiary structures with these tertiary structures having the potential to form complexes or quaternary structures (Moutevelis & Woolfson, 2009). A protein can be simply thought of a building block, starting with the amino acid sequence, and then order in assembly, however this masks the potential complexity of protein structures and there are a large number (order of thousands) of possible ways to fold protein chains into stable tertiary and quaternary structures (Blundell & Johnson, 1993; Alexandrov & Gō, 1994; Chothia, 1992).

As previously mentioned, amino acids are the building blocks of proteins and is always a linear sequence and relates to how a protein is named, starting from the amino-terminal (-NH2) end to the carboxyl-terminal also referred to as carboxylic acid functional group (-COOH) (Alberts et al., 2002). Of the 300+ amino acids, only 20 of them serve as building blocks of proteins (Wu, 2009) and chains of amino acids assemble via amide bonds known as peptide linkages. The unique properties of each amino acid is due to the difference in the side-chain group or R-group (Alberts et al., 2002) and the uniqueness of different proteins is determined by which amino acids it contains, how these amino acids are arranged in a chain, and further complex interactions the chain makes with itself and finally the environment (Alberts et al., 2002). The 20 amino acids needed to make all the proteins in the human body are all L-isomer, alpha-amino acids and all of them, except for glycine, contain a chiral alpha carbon and are R-absolute configuration except for glycine and cysteine (S-absolute configuration, because of the sulphur-containing R-group) (Alberts et al., 2002). A list of the 20 standard amino acids and their abbreviations is given below in Table 1.1.

Table 1.1 List of the 20 standard amino acids, their abbreviations and type
Table adapted from Rovira et al., 2008

Amino acid	3 letter	1 letter	Type
Alanine	Ala	A	Nonpolar, neutral
Arginine	Arg	R	Polar, basic
Asparagine	Asn	N	Polar, neutral
Aspartic acid	Asp	D	Polar,acidic
Cysteine	Cys	C	Polar, neutral
Glutamic acid	Glu	E	Polar, acidic
Glutamine	Gln	Q	Polar, neutral
Glycine	Gly	G	Nonpolar, neutral
Histidine	His	H	Polar, basic
Isoleucine	Ile	I	Nonpolar, neutral
Leucine	Leu	L	Nonpolar, neutral
Lysine	Lys	K	Polar. basic
Methionine	Met	M	Nonpolar, neutral
Phenylalanine	Phe	F	Nonpolar, neutral
Proline	Pro	P	Nonpolar, neutral
Serine	Ser	S	Polar, neutral
Threonine	Thr	T	Polar, neutral
Tryptophan	Trp	W	Nonpolar, neutral
Tyrosine	Tyr	Y	Nonpolar, neutral
Valine	Val	V	Nonpolar, neutral

1.1.1 Primary structure

As mentioned in Section 1.1, amino acids are the building blocks of proteins. In cells, DNA contains the code to synthesise proteins and the nucleotide sequence of a protein-encoding gene is transcribed into mRNA, which synthesises the sequence of amino acids to form a protein (Sanvictores & Farci, 2020). The gene corresponding to the protein is unique to that protein and defines the overall 3D structure and function of the protein (Alberts et al., 2002). Crucial insight into the sequence-structure relationships of proteins was made by Anfinsen who showed that all the information about the native structure of a protein is encoded in its amino acid sequence (Anfinsen, 1973).

Amino acids are linked together by joining the amino group of one amino acid with the carboxyl group of the adjacent amino acids through peptide bonds (Sanvictores & Farci, 2020). This linkage forms the polypeptide chain and polypeptides of more than 50 amino acids are known as proteins (Engelking, 2015). The characteristics of the amino acids e.g. acidic, basic, polar uncharged or non-polar will determine specific characteristics of the protein such as solubility in water or lipids and optimal physiological conditions for protein function (Sanvictores & Farci, 2020).

Insulin was the first protein to have the primary structure determined by Frederick Sanger in 1951 using hydrolytes and chromatography (Sanger & Tuppy, 1951). Current methods for determining the primary structure are Edman degradation, tandem mass spectrometry and DNA sequencing (Deutzmann, 2004).

The importance of the amino acid sequence in overall protein function is shown by the following disease examples:

1. **Huntingdon's disease**

Huntingdon's disease is a neurodegenerative disorder caused by a DNA trinucleotide repeat expansion of equal to or greater than 40 CAG repeats within the gene Huntingtin (Nopoulos, 2016)

2. **Sickle cell anaemia**

Sickle cell anaemia is caused by mutations in the HBB gene which encodes the haemoglobin subunit beta and is the result of a substitution of glutamic acid to valine amino acid and results in the sickle shape of red blood cells as opposed to the normal biconcave disk of healthy red blood cells (Cai *et al.*, 2018).

3. **Cystic fibrosis**

Cystic fibrosis is caused by a mutation on the CFTR (cystic fibrosis transmembrane regulator) gene on chromosome 7 (Sanvictores & Farci, 2020). Although there are more than 1,000 mutations for the CFTR gene the most common is deletion of Phe508 (Cutting, 2015). These mutations within CFTR gene alter the protein structure and thereby impairing chloride ion transport (Sanvictores & Farci, 2020).

1.1.2 Secondary structure

Secondary protein structure is the backbone of a protein and is subdivided into three categories: α-helix, ß-sheets and coil (Patel & Shah, 2013). With α-helix and ß-strand being the most common types of structure and was first suggested in 1951 by Linus Pauling and colleagues who referred to these structures as 5.1-residue helix and 3.7-residue helix, based on the number of amino acids residues per turn (Pauling, Corey & Branson, 1951). α-helix is considered the default state for secondary structure (Patel & Shah, 2013) and interactions via hydrogen bonding occurring between a carbonyl oxygen atom of a peptide linkage and the hydrogen atom of an amino group of another peptide linkage further along the protein backbone account for the formation of this structure (Stoker, 2016). In comparison, ß-sheets are formed when several ß-strands self-assemble and are stabilised by interstrand hydrogen bonding (Boyle, 2018). ß-sheets can have parallel, antiparallel, or mixed arrangements of the individual strands, however antiparallel is the most natural sheets in proteins (Boyle, 2018). The third class of secondary structure; coil or loop refers to less regular folds and is more unstructured (Pirovano & Heringa, 2010).

Prediction of secondary structure from protein sequence plays a crucial role in establishing the 3D structure as the secondary structure comes together to form the tertiary structure (Atasever *et al.*, 2019). This has importance in understanding the function of proteins and also drug design (Atasever *et al.*, 2019). Secondary structure is important for better understanding of the tertiary structure and more importantly knowledge of secondary structure helps in the prediction of tertiary structure, with structure discovery without sequence similarity in the datasets (Patel & Shah, 2013). Furthermore, accurate secondary structure information is at the core of most *ab initio* methods for the prediction of protein structure (Bradley *et al.*, 2003).

Experimentally, secondary structure can be solved with several spectroscopic methods such as circular dichroism (CD), Raman and infrared (IR) (Pelton & McLean, 2000). Nuclear magnetic resonance chemical shifts may also be used to determine the positions of secondary structure within the primary sequence of a protein (Pelton & McLean, 2000).

Modern algorithms can outline about 80% of the secondary structure based on the primary sequence and these methods include, but are not limited to, PORTER (Pollastri &

McLysaght, 2005), PSIPRED (Buchan & Jones, 2019), SSpro (Magnan & Baldi, 2014) and JPred4 (Drozdetskiy *et al.*, 2015).

1.1.3 Tertiary structure

Tertiary structure refers to the overall 3D arrangement of a polypeptide and is generally stabilised by outside polar hydrophilic hydrogen and ionic bond interactions in addition to, internal hydrophobic interactions between nonpolar amino acid side chains (Engelking, 2015). Based upon the tertiary structure, proteins can be divided into either globular or fibrous types. Fibrous proteins (e.g. α-keratin) have elongated rope-like structures which are strong and hydrophobic. On the other hand, globular proteins are more spherical and hydrophilic (Engelking, 2015). The process for tertiary structure folding begins while the protein is being moulded to its primary polypeptide sequence and is guided by chaperones (Engelking, 2015). The role of chaperones will be discussed later in Section 1.1.5.

Myoglobin was the first protein to have its tertiary structure solved in 1958 by Kendrew and colleagues using X-ray analysis (Kendrew *et al.*, 1958). Current experimental methods to determine tertiary structure are X-ray crystallography, nuclear magnetic resonance (NMR), cryogenic electron microscopy (cryo-EM) and dual polarisation interferometry (see Section 1.2).

1.1.4 Quaternary structure

Quaternary structures are formed from independently folded tertiary structures which associate together to form complexes (Moutevelis & Woolfson, 2009). Each subunit has its own primary, secondary and tertiary structure with the subunits held together by hydrogen bonds and van der Waals forces between the non-polar side chains (Ouellette & Rawn, 2015). The subunits must be arranged specifically for the entire protein to function properly. If alterations occur, this could have marked effects on the biological activity (Ouellette &

Rawn, 2015). The nomenclature for the subunits are monomer for one unit, dimer for two units, trimer for three units and so forth (Alberts et al., 2002).

Examples of proteins having quaternary structure are alcohol dehydrogenase, aldolase, fumarase, haemoglobin and insulin (Ouellette & Rawn, 2015).

Quaternary structure is usually determined by X-ray crystallography, but when crystallographic data are difficult to gather, electron microscopy can be used (Skipper, 2005).

Macromolecular complexes are of special interest in structural biology as direct protein-protein interactions, as well as indirect ones are essential for performance of several cellular processes (Bertoni *et al.*, 2017). The importance of protein-protein interactions is discussed in Section 1.5.

One of the first approaches to model interactions *de novo* was macromolecular docking (predicting the preferred orientation of two macromolecules) with docking approaches being generally more accurate when no significant conformational changes are required for interface formation (Bertoni *et al.*, 2017). If some experimental details of the interactions are available (e.g. EM density maps, crosslinking SAXS or NMR data, co-evolution analysis) then hybrid modelling tools such as ROSETTA3 (Leaver-Fay *et al.*, 2011) or HADDOCK (de Vries, van Dijk and Bonvin, 2010) can be used.

1.1.5 Protein folding

In order for a protein to assume its correct 3D structure, which is essential for function, the protein needs to fold. Proteins fold into their native state when they emerge from the ribosome (Englander & Mayne, 2014) and the folding process often starts when protein translation is not yet completed i.e. the protein N-terminus begins to fold whilst the C-terminus is still being synthesised (Rocco *et al.*, 2008). Protein folding is determined by different bonds between one chain and another (Alberts et al., 2002). These non-covalent bonds are hydrogen bonds, ionic bond and van der Waals forces. Another central force is hydrophobic molecules, which are forced together in an aqueous environment. Thus, an important factor in the folding of proteins is the distribution of nonpolar (hydrophobic) and polar (hydrophilic) amino acids (Alberts et al., 2002).

In 1962, Christian Anfinsen and Edgar Haber postulated that the native structure of a protein is the thermodynamically stable structure, or in other words, the final folded structure adopted by a polypeptide chain is generally the one in which the free energy is minimised (Alberts et al., 2002) and depends on the amino acid sequence and the conditions of solution (Haber & Anfinsen, 1962). Additionally, the native structure does not depend on whether the protein was synthesised using ribosomes, with the help of chaperone molecules, or if the protein was refolded as an isolated molecule in a test tube, with the exception of insulin (the biologically active form is kinetically trapped) (Haber & Anfinsen, 1962). A protein can be denatured following application of certain solvents that disrupt the noncovalent interactions involved with forming the folded chain, this converts the protein to a flexible polypeptide chain that has lost its natural shape. When the solvent is removed, the protein refolds to its original confirmation (Alberts et al., 2002).

Although a protein can fold to its native structure without assistance, protein folding, unfolding and homeostasis is assisted by molecular chaperones (Saibil, 2013). Chaperones are found in all cellular compartments and have little specificity but provide essential assistance to the highly specific protein folding process and members of structurally unrelated chaperones are known as heat-shock proteins (HSPs) (Hartl, Bracher and Hayer-Hartl, 2011). These chaperones are usually classified according to their molecular weight (e.g. HSP40) and the chaperones involved in *de novo* protein folding and refolding are

HSP70, HSP90 and HSP60 (Hartl, Bracher and Hayer-Hartl, 2011). These HSPs typically recognise hydrophobic amino-acid side chains which can become exposed by non-native proteins and may functionally cooperate with ATP-independent chaperones (HSPs are ATP and cofactor-binding dependent) to buffer aggregation (Hartl, Bracher and Hayer-Hartl, 2011). Therefore, chaperone binding to hydrophobic regions transiently blocks aggregation (Hartl, Bracher and Hayer-Hartl, 2011).

Partially folded or misfolded proteins cause problems as they tend to aggregate in a concentration-dependent manner and expose hydrophobic amino acid residues on a protein, which would normally be buried in the native state (Hartl, Bracher and Hayer-Hartl, 2011). A number of diseases can result from protein misfolding events and this ultimately leads to the malfunction of the cellular machinery (Welch, 2004). In the example shown below in Figure 1.2, amyloid fibrils have occurred due to unfolded proteins and this is a common feature of neurodegenerative conditions such as Alzheimer’s disease and Parkinson’s disease (Dobson, 2003).

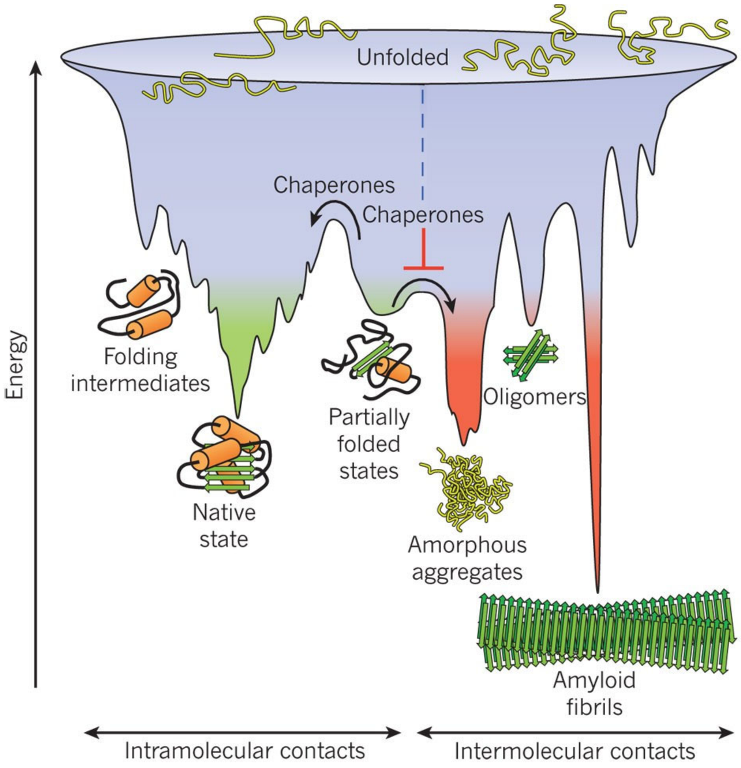


Figure 1.2. Competing reactions of protein folding and aggregation
Schematic of the funnel-shaped free-energy surface that proteins explore as they move towards the native state (shown in green in the energy section). The schematic also shows the importance of chaperones in accelerating the favourable downhill path i.e. lower energy. Amorphous aggregates are formed when several molecules fold simultaneously in the same compartment and the free-energy surface of folding may overlap with that of intermolecular aggregation. Figure taken from Hartl, Bracher and Hayer-Hartl, 2011.

Correct folding of proteins is critical for the biological activities of proteins. Most proteins, such as receptors, fulfil their biological activity only when correctly and completely folded (Rocco et al., 2008). Thereby, highlighting the importance of structure for function. Due to the importance of structure to function and the vital role folding plays in the overall 3D structure, the “protein folding problem” aims to answer the question of how a protein’s amino acid sequence dictates its 3D structure and consists of three problems: (1) what is the folding code? (2) what is the folding mechanism? (3) can structural biologists predict the native structure of a protein from its amino acid sequence? (Dill *et al.*, 2008). With regards to protein function, the additional problems are (1) can function be predicted from structure and (2) can prediction of ligands and ligand-binding residues provide insight into a protein’s function of which the thesis will aim to address. The *In silico* methods to predict structure will be discussed in Section 1.2.2 and *in silico* methods to predict ligands is discussed in Section 1.7.1

1.2 Protein structure determination

As mentioned previously in Section 1.1.3, myoglobin was the first protein to have its 3D structure solved experimentally and over the past six decades structural biologists have experimentally determined the structures of 180,000 proteins in the Protein Data Bank (AlQuraishi, 2021). Protein structure comparison is essential in almost every aspect of modern structural biology, ranging from experimental protein structure determination to *in silico*-based protein folding and structure prediction.

The Protein Data Bank (PDB) is the single global repository of experimentally (see Section 1.2.1 for a description of the experimental methods) determined 3D structures of biological macromolecules and their complexes (Xu & Zhang, 2010; Burley et al., 2017). In 2000, there were <80,000 protein structures in the PDB (Berman et al., 2003) and in 2015 there were more than 90,000 structures, with more than 75% of these having a protein-ligand complex (Burley *et al.*, 2017). In May 2017, the PDB archive housed ~130,000 entries (Salentin *et al.*, 2015) and at the time of writing, January 2021 PDB housed 173,110 with 105 structures released annually (Berman *et al.*, 2000).

GenBank is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation (Burley *et al.*, 2017). In comparison to the PDB, GenBank contains available nucleotide sequences for almost 260,000 formally described species (as of 2013) (Benson *et al.*, 2013) and release 194, produced in February 2013, contained more than 162 million sequences (Benson *et al.*, 2013). As of December 2020, GenBank contained 221 million sequences (Sayers *et al.*, 2019). It has been cited that experimental protein structures are currently available for less than 1/500th of the proteins with known sequences, thus highlighting the knowledge gap and the need to reduce this gap (*GenBank release notes*, 2017). Hence why there has been a need to solve the problem of predicting protein structures by in silico methods to help reduce this knowledge gap and reduce the time and/or resources spent on addressing this problem. The requirement to use in silico methods to solve protein structure is of particular importance with disordered proteins where disordered regions can prevent structure determination entirely by affecting solubility and/or crystalisability (Moult *et al.*, 2016). Additionally, there has been an explosion in the number of protein sequences from genome projects making it essential to have automated methods of prediction (Esnouf *et al.*, 2006).

1.2.1 Experimental methods for protein structure determination

Various experimental techniques can be used to determine protein structure, these are X-ray crystallography, nuclear magnetic resonance (NMR), small-angle X-ray scattering and cryo-electron microscopy (Du *et al.*, 2016).

The underlying principle of X-ray crystallography is that the crystalline atoms cause a beam of X-rays to diffract into many specific directions and by measuring the angles and intensities of these diffracted beams, a 3D image can be produced, detailing the density of electrons within the crystal (Ryu, 2017). From this image, the mean positions of the atoms in the crystal can be determined, as well as chemical bonds and their disorder, for example (Ryu, 2017). The advantages of X-ray crystallography include; provides a two-dimensional view that gives an indication of the three-dimensional protein structure, relatively cheap and simple compared to other techniques, useful for large structures and is not limited by size or atomic weight and can yield high atomic resolution (Brünger, 1997). Disadvantages are the protein must be crystallisable, with membrane proteins and large molecules difficult to crystallise due to their large molecular weight and poor solubility. An organised single crystal must be obtained to produce the desired diffraction and is a non-dynamic method due to preparation of samples and crystallisation so only a static three-dimensional analysis is produced (Brünger, 1997).

Since the 1970s, NMR has been used to study the interplay between biomolecular structure, dynamics and function (Mittermaier & Kay, 2006). Protein dynamics also has a role in ligand-binding, as this involved the entry of molecules into areas that would normally be occluded (Mittermaier & Kay, 2006). Specifically, saturation transfer difference NMR spectroscopy has been used to characterise binding in tightly bound ligand-receptor complexes (Meyer & Peters, 2003). When a protein becomes saturated, ligands that are in exchange between a bound and the free form become saturated when bound to the protein, by exchange that saturation is carried into solution where it is detected (Meyer & Peters, 2003). By subtraction of this spectrum from a spectrum without protein irradiation an NMR spectrum is obtained in which the signals are form molecules that bind to the protein (Meyer & Peters, 2003). A clear advantage being that resonance signals from nonbinders so not show up in the difference spectrum (Meyer & Peters, 2003). Additionally, NMR methods have the advantage of characterising the protein-ligand dynamics over a wide range of timescales, from picoseconds to seconds (Mittermaier & Kay, 2006) and can detect and reveal protein-ligand interactions with a large range of affinities (10^{-9} – 10^{-3} M) (Cala, Guilliere and Krimm, 2013). NMR allows internal motions to be probed with exquisite time and spatial resolution. Methodological advancements in NMR have extended the ability to characterise protein dynamics and will shed new light on the mechanisms by which these molecules function (Mittermaier & Kay, 2006). The disadvantages of protein-observed methods are the experimental time and the need for highly stable and soluble protein (Cala et al., 2013). Additionally, these methods are limited to proteins with low molecular masses (<30 kDa) to avoid great effort with regard to both labelling strategies and resonance assignment (Cala et al., 2013).

For high resolution X-ray crystallography, a homogenous crystal is needed and this results in a reaction needing to be synchronised across the entire crystal (Henzler-Wildman & Kern, 2007). The requirement for a homogenous crystal is relieved when using both cryo-electron microscopy and small-angle X-ray scattering (Henzler-Wildman & Kern, 2007). Small-angle X-ray scattering is a technique where the elastic scattering of X-rays by a sample is recorded at very low angles (typically 0.1-10^{o} measured from the beam axes) (Londoño *et al.*, 2018). This angular range contains information regarding the structure of scatterer entities, like nanoparticles and micro- and macromolecules (Londoño *et al.*, 2018). The advantages of small-angle X-ray scattering is a larger volume of sample can be illuminated when compared to other methods used such as transmission electron microscopy, a fact that leads to the estimation of more precisely average values and disadvantages are despite the small-angle X-ray scatter pattern being obtained from over all particles orientated in all directions, the structural features are determined in an indirect way, an issue that could lead to ambiguous results and incorrect interpretations (Londoño *et al.*, 2018). Additionally, both small-angle X-ray scattering and cryo-electron microscopy does not characterises the timescales of interconversion (Henzler-Wildman & Kern, 2007). In comparison, cryo-electron microscopy involves proteins, or other specimens being prepared by plunge-freezing thin aqueous films into a liquid cryogen (Tivol et al., 2008). Using an electron microscope the structure, of a protein, for example, is visualized (Tivol et al,, 2008). Advantages of cryo-electron microscopy include; small samples can be used to determine structure and a wide range of samples can be characterised which are over 500 kDa (Wang *et al.*, 2015). Disadvantages include time-constraints, cost and fully hydrated specimens can be electron-beam sensitive (Wang, 2015).

1.2.2 Computational methods for protein structure prediction

Predicting 3D structure from its amino acid sequence is an important unsolved problem in both biophysics and computational biology (Deng, Jia and Zhang, 2018) and the goal is to determine the shape (or fold) that an amino acid will adopt (Rangwala & Karypis, 2010). The number of proteins depositing into UniProt and PDB is growing at an exponential rate, particularly in the last two decades, the reason for this is that it is easier to obtain protein sequences than to predict protein structure (Deng, Jia and Zhang, 2018). This has been aided by the development of advanced DNA sequencing technology, which has enabled the sequences of proteins to be rapidly accumulated (Deng, Jia and Zhang, 2018). Computational methods for prediction of protein structure from its amino acid sequence has become increasingly popular (Deng, Jia and Zhang, 2018). In 1973 (Anfinsen, 1973), Anfinsen demonstrated that all the information a protein needs to fold properly is encoded in the amino acid sequence (referred to as Anfinsen's dogma) (Deng, Jia and Zhang, 2018), therefore this makes the determination of a protein's structure possible from a computational perspective.

The problem is divided based on will the sequence adopt a new fold or bear resemblance to an existing template (fold). Fold recognition is easy when the sequence has a high degree of sequence similarity to a sequence with a known structure. The second part of the problem is building the protein structure from scratch and are usually referred to as *ab initio* methods (Rangwala & Karypis, 2010). Computational methods for protein structure prediction are categorised as either template-based modelling (TBM) or template-free modelling based on the above problems and will be discussed in Section 1.2.2.1 and 1.2.2.2, respectively. Table 1.2 on the next page is a list of some of publicly available methods for protein structure prediction.

Table 1.2 A list of some of the publicly available methods for tertiary protein structure prediction
Table adapted from (Farhadi, 2018). TBM = template-based modelling, HMM-HMM = Hidden Markov Model-Hidden Markov Model. BLAST= Basic Local Alignment. PDB=Protein Databank

Method	URL	Summary
HHpred (Zimmermann *et al.*, 2018)	https://toolkit.tuebingen.mpg.de/tools/hhpred	Homology detection and structure comparison by HMM-HMM comparison
IMP (Webb *et al.*, 2018)	https://integrativemodeling.org	TBM method following four stages 1) gathering input data 2) converting input data into a model 3) scoring function 4) alternative model configurations
IntFOLD (McGuffin *et al.*, 2019)	http://www.reading.ac.uk/bioinf/IntFOLD/	Unified TBM resource for automated prediction with built-in estimates of model accuracy
I-TASSER (Yang & Zhang, 2015)	https://zhanggroup.org/I-TASSER/	TBM approach with multiple threading approaches using a hierarchical approach to protein structure prediction
ModPipe (Sánchez & Sali, 1998)	https://salilab.org/modpipe/	Automated software TBM pipeline using template structures and sequence-structure alignments
MODELLER (Webb & Sali, 2016)	http://salilab.org/modeller/	TBM method for homology or comparative modelling using spatial restraints. Additional tasks include de novo modelling of loops
Phyre2 (Kelley *et al.*, 2015)	http://www.sbg.bio.ic.ac.uk/~phyre2	TBM webserver with a suite of tools and uses advanced remote homology detection
Robetta (Kim, Chivian and Baker, 2004)	http://robetta.bakerlab.org	Automated webserver using TBM comparative modelling
ROSETTA (Rohl *et al.*, 2004)	https://www.rosettacommons.org	De novo fragment insertion method for protein structure prediction
SWISS-MODEL (Guex & Peitsch, 1997)	http://swissmodel.expasy.org	Fully automated homology modelling server which utilises BLAST and HHblits
ModWeb (Sánchez & Sali, 1998)	https://modbase.compbio.ucsf.edu/modweb/	Automated TBM web server utilising templates from PDB or SWISS-PROT protein sequences
Chunk-TASSER (Zhou & Skolnick, 2007)	http://cssb.biology.gatech.edu/skolnick/webservice/chunk-TASSER/index.html	*Ab initio* method using supersecondary structure chunks as well as threading templates
LOMETS (Zheng, Zhang, *et al.*, 2019)	https://zhanggroup.org/LOMETS/	TBM method integrating multiple deep learning-based threading methods

The fundamental steps in protein structure prediction are conformation initialisation (using either Template-free or Template-based methods), conformational search, structure selection, all-atom structure reconstruction and structure refinement (Deng, Jia and Zhang, 2018). Figure 1.3 below is a flowchart of protein structure prediction.

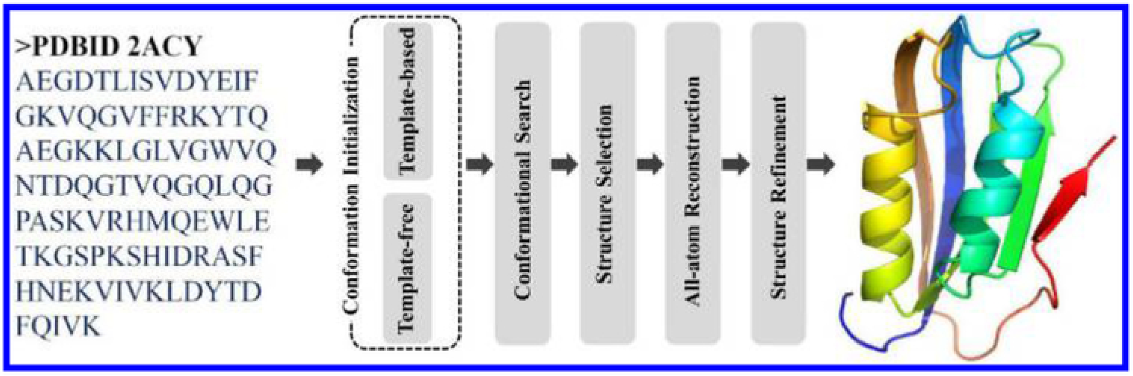


Figure 1.3. The general flowchart of protein structure prediction
The specific details of protein structure determination methods can vary significantly, however there are fundamental steps which are consistent across all methods. Figure taken from (Deng, Jia and Zhang, 2018)

1.2.2.1 Template-based modelling

Template-based structure prediction methods consist of homology modelling (sequence comparison) sometimes referred to as comparative modelling and threading methods (fold-recognition) (Deng, Jia and Zhang, 2018). The basis of homology modelling is that similar sequences from the same evolutionary family often adopt similar protein structures (Deng, Jia and Zhang, 2018). Homology modelling has clear advantages in that when a single structure within a family of homologous proteins has been determined experimentally there is potential to model all proteins within that family (Rangwala & Karypis, 2010)

The most accurate way of predicting protein structure is by taking its homologous structure (i.e. known structure) in PDB and utilising this as a template. When there is no structure with sequence similarity to the investigated protein, threading or fold recognition is used to identify templates (Deng, Jia and Zhang, 2018). Homology modelling usually requires the protein target and the template to share sequence identity of >25% and if not, protein threading can overcome this limitation (Deng, Jia and Zhang, 2018). Fold recognition identifies remote sequence homology via sequence comparison to detect structural similarity (Zhou & Zhou, 2005).

Examples of template-based modelling methods are SWISS-MODEL(Guex & Peitsch, 1997), Modeller (Webb & Sali, 2016) and IntFOLD (McGuffin *et al.*, 2019). The prediction process for SWISS-MODEL consists of template recognition, target-template alignment, model building and model evaluation (Guex & Peitsch, 1997). SWISS-MODEL utilises BLAST and HHblits for template recognition and target-template alignment. The structure of the protein in question is then built by copying the atomic co-ordinates from the template according to target-template alignment (Guex & Peitsch, 1997). Modeller (Webb & Sali, 2016) implements structure modelling by satisfaction of spatial restraints which are derived from target-template alignment and experimental data (e.g. NMR spectroscopy) and other sources such as stereochemistry (Webb & Sali, 2016). IntFOLD is a server with six component methods and the tertiary structure prediction component, IntFOLD-TS works using iterative multi-template modelling using the target-template alignments from 14 alternative methods (McGuffin *et al.*, 2019). Further information about IntFOLD is provided in Section 1.7.3.

It is important to note, that historically most of the successful protein structure modelling methods are template-based (Deng, Jia and Zhang, 2018)

1.2.2.2 Template-free modelling

Template-free modelling, is divided into two categories *de novo* modelling (structure-based) and *ab initio* modelling (modelling from first principles) (Rangwala & Karypis, 2010). Both modelling methods aim to predict protein structure without relying on having a previously solved structure (Deng, Jia and Zhang, 2018). One of the strengths of template-free methods is the structure prediction of hard target proteins of which no satisfactory template can be found (Deng, Jia and Zhang, 2018) and a weakness is success is better on small proteins below 100 residues (Xu & Zhang, 2012). *Ab initio* protein folding has difficulties in force field design and conformational search, of which requires extensive computing demand. There is a lack of decent force fields to accurately describe the atomic interactions which can assist in guiding protein folding simulations (Xu and Zhang, 2012). Rosetta is a template-free modelling protein structure prediction method in which short fragments of known proteins are assembled by a Monte Carlo (repeated random sampling) strategy to yield native-like protein conformations (Rohl *et al.*, 2004). One of the ways template-free methods have increased the efficiency of conformational search is to reduce the level of protein structure representation (Xu & Zhang, 2012). In I-TASSER, each residue is specified by two units of Cα atom and side-chain center of mass (Zhang, 2007). These reductions of structure representation can dramatically reduce the total number of conformations needed for searching (Xu & Zhang, 2012).

Despite TBM methods historically outperforming free modelling methods, AlphaFOLD which was developed by DeepMind, a machine learning method predicted protein structure with near-experimental accuracy and outperformed other Critical Assessment of protein Structure Prediction (CASP) 13 entrants (AlQuraishi, 2021) placing first in the free modelling category, for reference the Zhang group placed first in the TBM category (AlQuraishi, 2019). In CASP13, the median accuracy was 6.6Å and in CASP14, AlphaFOLD2 had improved accuracy to 1.5Å and this is comparable to the accuracy of experimental methods (AlQuraishi, 2021). The relevance of CASP is discussed in Section 1.8.1.

1.3 Proteins and their interacting partners: a biomedical perspective on protein-ligand interactions

Proteins perform their biological functions through direct physical interaction with other molecules, such as proteins, peptides, nucleic acids, membrane, substrates and small molecule ligands e.g. metals (Du *et al.*, 2016). The molecule that is bound to a protein, no matter whether it is a small ion or macromolecule is referred to as a ligand for that protein and comes from the Latin *ligare*, meaning “to bind” (Alberts et al., 2002). Thus, a ligand can be both endogenous and exogenous. The resulting endogenous protein-ligand complex is important for a variety of intracellular processes. Upon binding of a ligand, a conformational change occurs in the protein which in turn starts the initiation of cellular functions. These functions include, but are not limited to, immune defense with the binding of antibodies to antigens, which exhibit precise ligand-binding specificity (Lodish et al., 2000) and catalysts of chemical reactions by enzymes is dependent upon the specificity of ligands (enzyme substrate) (Lodish et al., 2000). Hence, studying protein-ligand interactions is an important step in the functional elucidation of proteins involved in these cellular processes (Marienhagen *et al.*, 2008).

Several protein classes have functional implications as drug targets, these include enzymes along with transportation and signalling proteins, which are further subdivided into transportation molecules, ion channels and receptors (Roche et al., 2011, 2012, 2013).

A key example of an enzyme is Cytochrome P450, which has an essential role in the electron transfer chain, and is therefore ubiquitous in all kingdoms of life. Cytochrome P450s biological function is to support the oxidative, peroxidative and reduction metabolism of both endogenous and xenobiotic substrates (e.g. steroids, fatty acids and environmental pollutants) (Danielson, 2002). Additionally, it is the single most important enzyme for Phase I drug metabolism where the critical role is drug interactions and interindividual variability in drug metabolism (Danielson, 2002). Cytochrome P450 is a haem protein and is part of a subfamily of related but equally distinct proteins (Rang et al., 2015). Cytochrome P450s are organised on the basis of similarities in protein sequence, individual CYP450s within a family are defined as having ≤40% sequence similarity compared with another CYP450s in any other family (Gonzalez & Gelboin, 1992). Families are further divided into subfamilies, with all CYP450s within a subfamily are >55% similar in sequence (Gonzalez & Gelboin, 1992). Figure 1.5A shows Cytochrome P450 bound to the drug N-Benzylformamide; with a detailed view of the ligand-binding site shown in Figure 1.6A. The subfamily of enzymes which makes up this family, are distinct from each other in terms of ligand specificity (Rang et al., 2015). For example, CYP1A1 (PDB ID 418v) is responsible for the metabolism of theophylline (Rang et al., 2015), a drug used to provide symptomatic relief from the asthma as it is a phosphodiesterase inhibitor; whereas, CYP2C9 (PDB ID 4ph9) is responsible for the metabolism of ibuprofen a cyclooxygenase inhibitor (Rang et al., 2015). CYP2C9 bound to ibuprofen can be seen in Figure 1.5B, while Figure 1.6B highlights the protein-ligand interaction focused on the binding site. It is worth noting that, whilst the CYP450 subfamily show ligand specificity, they are not molecule specific and are in fact quite promiscuous. For example, CYP1A1 binds the following compounds; β-napthoflavone (aryl hydrocarbon agonist), cisplatin (DNA replication inhibitor), dopamine hydrochloride (agonist based on endogenous dopamine), 17 β-estradiol (endogenous steroid hormone), and quinidine sulphate (inhibitor of voltage-gated sodium channel), to name but a few (Rang et al., 2015) refer to Figure 1.4 which illustrates the diverse ligand structures. Thus, showing the diversity in the substrates encompassing a broad array of molecular shapes, volumes, geometrics and chemical properties that CYP450s are able to oxidise with exceptional regio- and stereospecificity (Gay, Roberts and Halpert, 2010).

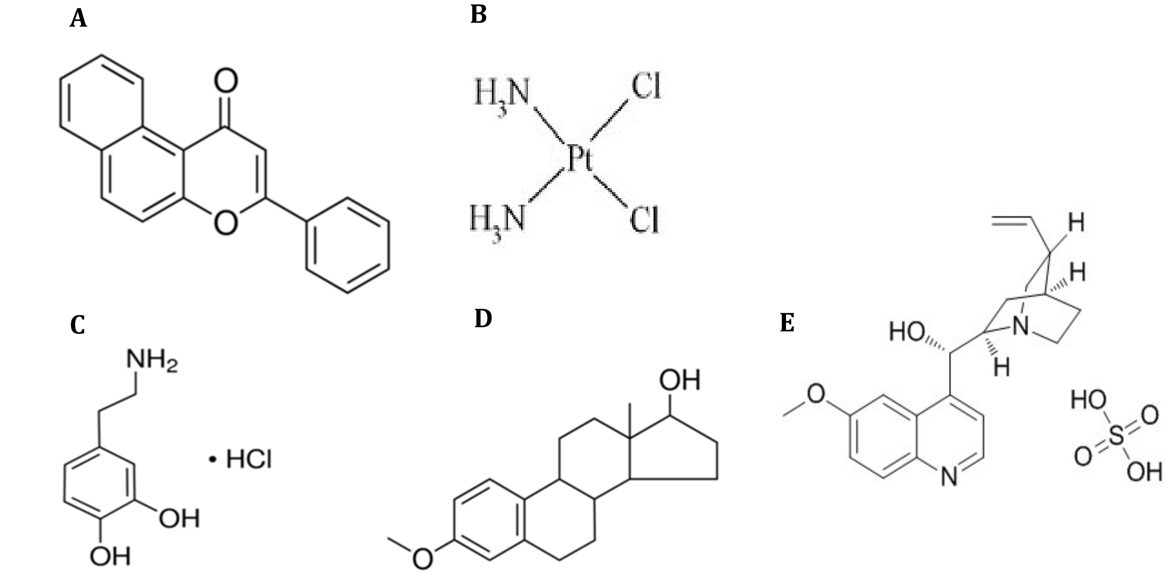


Figure 1.4. Diversity of ligand substrates for CYP1A1
Figure 1.4 above shows the diverse ligand structures which are metabolised by CYP1A1. **(A)** β-napthoflavone **(B)** cisplatin **(C)** dopamine hydrochloride **(D)** 17 β-estradiol and **(E)** quinine sulphate

Crystal structures of CYP450 have captured these enzymes in a variety of conformations, highlighting the ability of CYP450s to adapt their structures to accommodate a wide variety of substrates (Gay, Roberts and Halpert, 2010). For example, CYP2B4 has shown to have the largest degree of structural flexibility of any of the CYP450s to be crystallised. CYP2B4 has been observed in the similar closed, compact conformation when bound to four different molecules but has also been crystallised in three distinct open conformations when bound to large compounds or in the absence of a ligand (Gay, Roberts and Halpert, 2010). However, not all CYP450s share this flexibility in conformation like CYP2B4. CYP2A6 has been crystallised in the presence of seven different ligands of various sizes and shapes (Gay, Roberts and Halpert, 2010) and despite the difference in compounds when bound to an active site, CYP2A6 remains in a single closed conformation with little rearrangement of side chains (Gay, Roberts and Halpert, 2010).

In addition to the pharmacological effects of drugs, cytochrome P450 is also responsible for toxicological effects of drugs. For example, the primary enzyme responsible for paracetamol metabolism is CYP2E1, a Cytochrome P450, which is one of the enzymes involved in the metabolism of xenobiotics in humans. Metabolism via this enzyme forms a toxic alkylating metabolite, known as N-acetyl-p-benzoquinione imine (NAPQI). At therapeutic doses, NAPQI is detoxified via a conjugation reaction with glutathione. However, at doses exceeding therapeutic levels, the glutathione conjugation pathway becomes saturated and NAPQI conjugates with thiol groups on proteins and nucleic acids causing hepatotoxicity; a typical feature of paracetamol poisoning (Cameron et al., 2014).

The most promiscuous enzyme in the family is CYP3A4, which is involved in the metabolism of more than half of all drugs (Table 1.3) (Westerink et al., 2008). This is attributed to its very large and flexible active site, which can undergo conformational changes upon binding of substrates (Rydberg & Olsen, 2012).

It is worth mentioning, that whilst CYP enzymes are mentioned as examples, these enzymes won’t be part of the analysis, as the predictions for CASP and CAFA are randomly assigned and typically relate to proteins where the structures are unknown but are soon to be released and/or proteins where the functional annotation is yet to be determined.

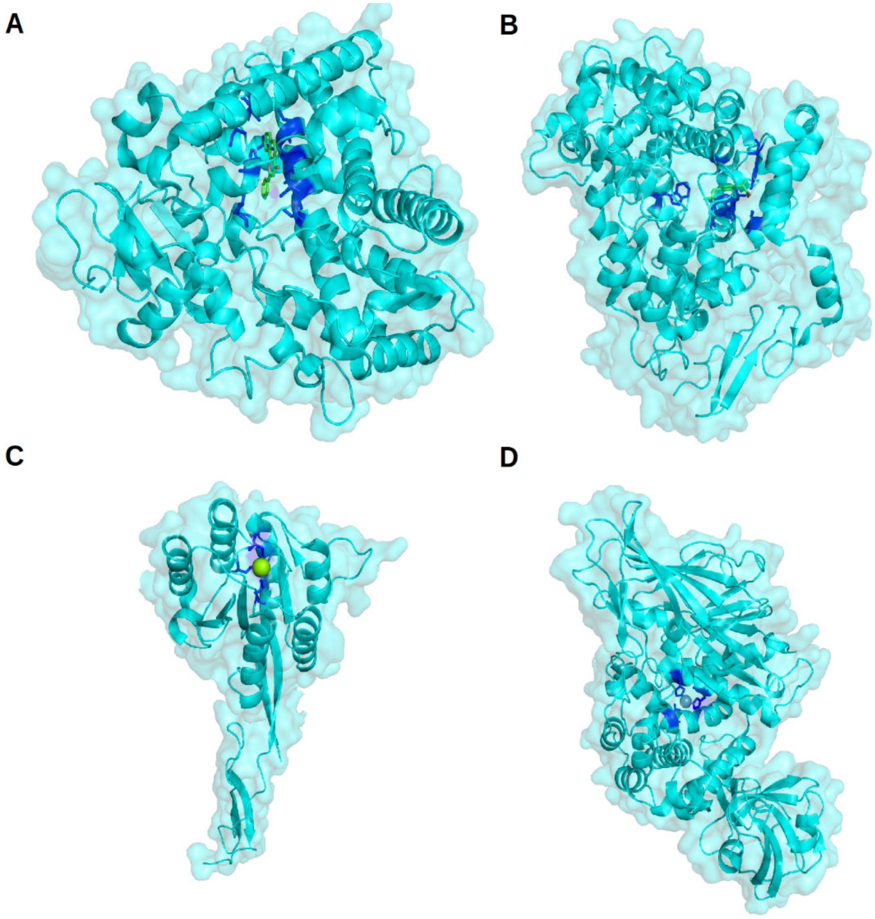


Figure 1.5. Examples of protein-ligand interactions
Proteins are shown in cartoon form, with the surface highlighted and coloured in cyan, binding site residues are shown as sticks and coloured blue, and ligands shown as sticks or spheres and coloured by element; **(A)** The Human cytochrome P450 1A1 protein (PDI ID 4i8v) bound to the drug N-Benzylformamide; **(B)** Cyclooxygenase-2 (PDB ID 4ph9) from *Mus musculus* bound to the drug Ibuprofen; **(C)** The *Plasmodium vivax* TRAP protein (PDB ID 4hqo, T0686) bound to magnesium and; **(D)** The aminopeptidase N family protein Q5QTY1 (PDB ID 4fgm, T0726) from *Idiomarina loihiensis* bound to zinc.

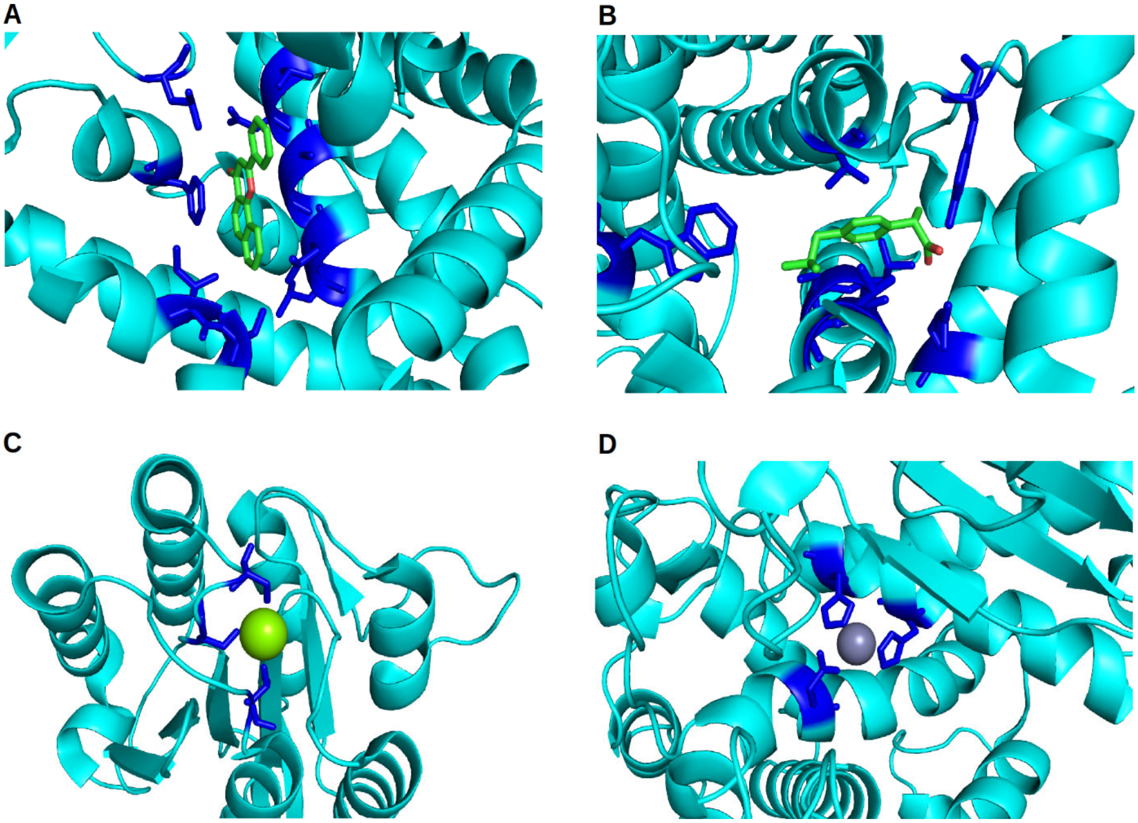


Figure 1.6. Examples of protein-ligand interactions, focusing on the ligand-binding site
Proteins are shown in cartoon form, with the surface highlighted and coloured in cyan, binding site residues are shown as sticks and coloured blue, and ligands shown as sticks or spheres and coloured by element; **(A)** The Human cytochrome P450 1A1 protein (PDI ID 4i8v) bound to the drug N-Benzylformamide; **(B)** Cyclooxygenase-2 (PDB ID 4ph9) from *Mus musculus* bound to the drug Ibuprofen; **(C)** The *Plasmodium vivax* TRAP protein (PDB ID 4hqo, T0686) bound to magnesium and; **(D)** The aminopeptidase N family protein Q5QTY1 (PDB ID 4fgm, T0726) from *Idiomarina loihiensis* bound to zinc.

Table 1.3. Promiscuity of CYP450 enzymes
Examples of ligands acting on the CYP450 enzymes. Ligands listed in red are inhibitors and ligands in green are inducers of the CYP450 enzyme system. Table adapted from Rydberg & Olsen, 2012

CYP3A4	CY2C9/8	CYP2D6	CYP2C19
Aprepitant	Fluconazole	Isoniazid	Fluconazole
Clarithromycin	Ibuprofen	Ketoconazole	Ketoconazole
Erythromycin	Indomethacin	Methadone	Isoniazid
Isoniazid	Ketoconazole	Nicardipine	Omeprazole
Itraconazole	Sulfamethoxazole		Carbamazepine
Ketoconazole	Trimethoprim		Phenytoin
Metronidazole	Carbamazepine		Rifampin
Valproic Acid	Phenobarbital		
Carbamazepine	Phenytoin		
Dexamethasone	Rifampin		
Phenobarbital			
Phenytoin			
Rifampin			

As well as being promiscuous, proteins can also “moonlight”, meaning the same protein can perform different functions when placed in a different organism or a different location within the same organism. For instance, the same protein can have a different function when it is outside of the cell, expressed in a different cell type, moves into an organelle, interacts with other proteins to form a multi-protein complex or interacts with another protein to form a heterodimer (Figure 1.7) (Haidar & Jeha, 2011). “Moonlighting” is only applicable to a subset of multifunctional proteins in which two or more different functions are performed by one polypeptide chain (Jeffery, 2005). Proteins with the same function in multiple locations are not referred to as “moonlighting” proteins (Jeffery, 2005).

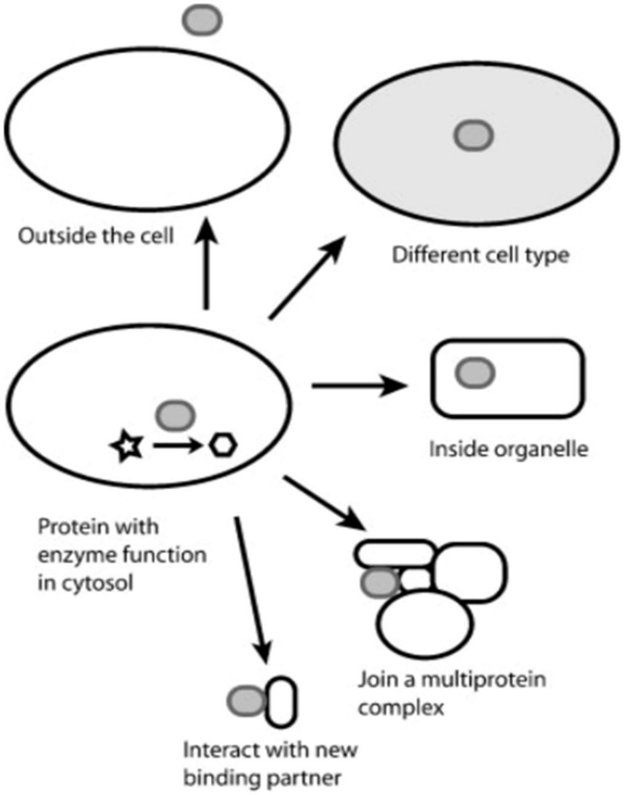

Outside the cell
Different cell type
Inside organelle
Protein with enzyme function in cytosol
Join a multiprotein complex
Interact with new binding partner

Figure 1.7. Examples of protein moonlighting
A protein can have one function in the cytosol of one cell but have a different function outside the cell, when expressed in a different cell type, moves into an organelle, interacts with other proteins to form a multi-protein complex or interacts with another protein to form a heterodimer. Taken from Jeffery, 2005

An example of a “moonlighting” protein is albaflavenone synthase (CYP170A1), a haem-dependent monooxygenase that catalyses the final two steps in the biosynthesis of antibiotic bacterium, steptomyces coelicolou (Nelson *et al.*, 1996) and binds to haem iron in two orientations (Zhao et al., 2009). As a monooxygenase, albaflavenone catalyses the conversion of the terpenoid epi-isozizaene to an epimeric mixture of ablaflavenols, which are then oxidised to the sesquiterpene antibiotic albaflavenone with haem being the cofactor (Zhao et al., 2009). Additionally, albaflavenone synthase has a second, completely distinct catalytic activity corresponding to the synthesis of farnesene isomers (acyclic sesquiterpene farnesene) from farnesyl diphosphate i.e. an intrinsic terpene synthase (Zhao et al., 2009). This activity was independent of protein redox partners (flavodoxin and flavodoxin reductase) and NAPDH, thereby suggesting an intrinsic terpene synthase activity which is distinct from the monooxygenase activity which produces albaflavenone (Zhao et al., 2009). Assessment of the primary sequence and X-ray structure as well as enzymatic data revealed the α-helical domain; indicating the presence of a novel terpene synthase, which is moonlighting in the structure of albaflavenone and is illustrated in Figure 1.8 (Zhao et al., 2009).

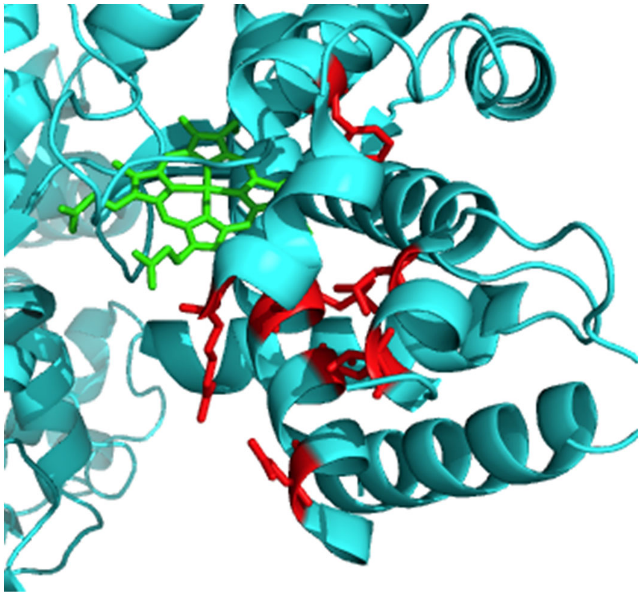


Figure 1.8. Structure of albaflavenone monooxygenase highlighting the haem ligand-binding site and the moonlighting terpene synthase active site
Albaflavenone monooxygenase (PDB ID 3dbg) is shown in cartoon form and coloured cyan, with the haem ligand-binding site residues shown as sticks and coloured green, and the moonlighting active site for terpene synthase shown as sticks and coloured red. Image adapted from Zhao et al., 2009

In addition, there are critical proteins involved in transportation, such as the sodium/potassium pump, which is responsible for transporting three intracellular sodium ions in exchange for two extracellular potassium ions (Zhao et al., 2009). The effect of this exchange has important implications for the transport of amino acids, sugars, bile acids, neurotransmitters and ions (Glitsch, 2001). As well as sodium pumps there are voltage-gated sodium channels, which are responsible for action potentials in cardiac myocytes, skeletal muscle and neurons (Glitsch, 2001). Key ligands, such as lidocaine, block the initiation and propagation of action potentials by this ion channel and is used as local anaesthetics (Kwong & Carr, 2015).

Receptor proteins are also of primary interest in biomedical research due to their roles in intercellular signalling, for example, ligand-gated ion channels, G-protein-coupled receptors (GPCRs), kinase- linked receptors and nuclear receptors (Rang et al., 2015). By far, the largest family of receptors is G-protein-coupled receptors. These receptors bind a diverse range of substrates such as hormones and slow transmitters (e.g. muscarinic acetylcholine receptors) (Rang et al., 2015). Incorrect signal activation of GPCRs particularly in the Wnt signalling pathway, is implicated in a number of cancers, including breast cancer (Rang et al., 2015) as GPCRs act as primary receptors for Wnt signals (Ng *et al.*, 2019). The Wnt signalling pathway is an evolutionarily conserved signal transduction pathway that regulates a wide range of cellular functions and controls multiple aspects of development, including cell proliferation and apoptosis (Clevers, Loh and Nusse, 2014). Inappropriate activation of the Wnt pathway is a major factor in human oncogenesis (Polakis, 2000).

Each of the proteins mentioned above vary in terms of structure, function and target specificity and their subsequent cellular effects. However, there is one common denominator; each of the proteins needs to bind a substrate to enable their functionality. A substrate can also be referred to as a ligand. From a pharmacology perspective; a ligand is a small molecule, frequently although not always, a drug (Koval *et al.*, 2014).

It is important to note that drug binding does not always lead to receptor activation. Binding can also result in either an inhibitory effect or no effect. Hence, binding and activation are two distinct steps in receptor response (Rang et al., 2015). In general, there are three main types of interactions between a receptor and a ligand; agonistic, antagonistic and inverse agonistic (Rang et al., 2015).

An agonist, quite simply occupies the binding site on a receptor and activates the receptor; leading to a response or effect. Agonists can be divided further, into full agonists (producing an optimal response) or partial agonist (suboptimal response) (Rang et al., 2015). Antagonists are usually competitive in nature, whereby they compete for the receptor-binding site with an agonist, but upon binding no response occurs. Thus competition for the binding site ensues, as usually the receptor protein can only bind one ligand at a time (Rang et al., 2015). An inverse agonist binds to a receptor, in the same way as an agonist; however, it induces a conformational change within the receptor that decreases the affinity of the receptor for a cofactor (Rang et al., 2015). Inverse agonists often show preference for binding to the resting state of a receptor (Kwong & Carr, 2015).

1.4 The physiology of protein-ligand interactions

Protein-ligand interactions are essential for biochemical functionality and are implicated in all biochemical roles in all kingdoms of life. Protein-ligand interactions are key to the pharmacological effects of drugs or the response to endogenous ligands, such as neurotransmitters. Without a direct interaction between a protein and its ligand, there can be no observed effect. This interaction is based on ligand specificity; the interacting protein will only recognise ligands of a precise type based on ligand affinity and specificity (Rang et al., 2015). It could be assumed that closely related ligands would either be ignored or produce a weak interaction. However, the molecular similarity principle, and subsequently named bioisosteres, show chemical substituents with similar physical or chemical properties, will produce a broadly similar biological property to another chemical compound. Key examples of this are morphine, diamorphine and codeine that bind to μ- ∂- and/or κ- opioid receptors. All these analogues are a phenanthrene derivative with two planar rings and two aliphatic ring structures, which occupy a plane at approximate right angles to the rest of the molecule. This structural similarity means all the analogues behave in the same way (Rang et al., 2015). In chemoinformatics, the process of searching for compounds that are structurally diverse and share biological activity is called scaffold hopping and is often exploited in patent-breaking (Rang et al., 2015).

More often, investigating how ligand agonists exert their effect is undertaken through the use of drugs, which can be targeted against the receptor protein, with the effect of this response termed antagonistic or inversely agonistic. A prime example of this would be naloxone, used in the treatment of opiate overdose. Naloxone acts as a μ, δ, κ-opioid protein receptor competitive antagonist (Willett, 2014), with the greatest potency of the μ receptor, naturally occurring ligands are ß-Endorphin, enkephalins and dynorphin A, respectively (Pasternak and Pan, 2013). Morphine is a pure agonist, particularly at the μ receptor and the opposing effects of naloxone to morphine is used in emergency situations to reverse the effects of opiates (Feng *et al.*, 2012).

A more recent example, which is still under pharmacological development, is ghrelin. Ghrelin is an endogenous peptide hormone, which plays an important role in the regulation of appetite, food intake and alcohol dependence (Koval *et al.*, 2014) by binding to a specific signalling G protein-coupled receptor (GPCR). Understanding the signalling pathways of ghrelin, as a result of investigating its ligand-binding properties and thus its potential effects, has led to several pharmaceutical companies investigating potential drug targets. For example, specific antagonist or inverse agonist properties to be used as anti-obesity medication (Cameron et al., 2014).

The ability of ligands to bind to proteins and disrupt normal cellular pathways, has been utilised in the treatment of cancer. For example, the vinca alkaloid drug, vincristine, is used in the treatment of leukaemias, malignant lymphomas, multiple myeloma, solid tumours, paediatric solid tumours and idiopathic thrombocytopenic purpura. Vincristine exerts its pharmacological effects by binding to the protein tubulin. This inhibits the assembly of microtubule structures and prevents mitosis in the metaphase (Routledge *et al.*, 1998). Whilst this inhibition targets rapidly dividing cells, such as cancer cells, it also affects other rapidly dividing healthy cells.

The examples above have focused on the ligand-binding ability, but what happens with defective proteins that cannot bind ligands? X-linked lymphoproliferative (XLP) syndrome or Duncan disease, is a rare immunodeficiency disorder, which only affects males (Rowinsky &

Donehower, 1991). Sufferers exhibit abnormalities in the functions of T and natural killer (NK) cells, leading to premature death. The condition is caused by a mutation in the serum amyloid P component (SAP) protein (Li *et al.*, 2003). Li et al., 2003 showed that SAP mutations found in XLP patients are defective in binding to the endogenous ligand (signalling lymphocyte activating molecule, or SLAM), which is vital for T cell activation. As a result of the failure to activate T cells, patients develop infectious mononucleosis or B cell lymphoma (Li *et al.*, 2003).

Illustrated in Figures 1.5 and 1.6 are four examples of protein-ligand interactions for diverse types of ligands, which are important in health and disease. This includes Cytochrome P450 bound to the drug N-Benzylformamide, which targets asthma (Figures 1.5A and 1.6A); Cyclooxygenase-2 from *Mus musculus* is responsible for the metabolism of ibuprofen (Figures 1.5B and 1.6B); the *Plasmodium vivax* thrombospondin related adhesion protein (TRAP) protein bound to magnesium, is involved in phosphate ester hydrolysis (Figures 1.5C and 1.6C); and the aminopeptidase N family protein Q5QTY1 from *Idiomarina loihiensis* bound to zinc (a co-factor), (Figures 1.5D and 1.6D), can be used as a biomarker to detect kidney damage. These four diverse examples along with the examples discussed in this section highlight the central role protein-ligand interactions can play in health and disease.

Proteins can interact with a broad range of molecules which are broadly referred to as ligands, such as small ions (e.g. Zn^{2+}), small molecules (e.g. adenosine triphosphate) and macromolecules (e.g. proteins) to perform their respective functions and the role of these ligands may vary markedly from being a substrate, inhibitor or activator. Whilst some interactions are unspecific and transient (e.g. water molecules and other solutes in the cell), others are very specific and essential for the function of the protein. This is of paramount importance in the characterisation of a new protein as information about ligands often provides crucial hints about its function (Li *et al.*, 2003). However, experimental determination of the structure of protein ligands with medium to low binding affinities are often lost during the purification procedure (Gallo Cassarino et al., 2014).

1.4.1 Protein-ligand binding models

Three different models exist to explain protein-ligand interaction exist; “lock-and-key” (Figure 1.9A), “induced fit” (Figure 1.9B) and “conformational selection” (Figure 1.9C) (Du *et al.*, 2016). With respect to the “lock-and-key” model, it is assumed that both the protein and the ligand are rigid and their respective binding interfaces are perfectly matched. The obvious limitation of this model is only a correctly sized ligand (the key) can fit within the binding pocket of the protein (lock) and does not explain when a protein and ligand bind when initial shapes do not match (Du *et al.*, 2016). The “induced fit” model overcomes this limitation, and assumes that the binding site of a protein is flexible and the interacting ligand induces a conformational change within the binding site of the protein. A limitation of the induced fit model is this only takes into account the conformational changes after ligand binding (Du *et al.*, 2016). The conformational model, aimed to overcome the limitations of both lock-and-key and induced fit by postulating that the native state of a protein does not exist as a rigid conformation, but instead as a vast ensemble of conformational states that coexist in equilibrium. The ligand can bind selectively to the most suitable conformational state. Thus, the unbound protein can have the same conformation as that of the ligand-bound state (Du *et al.*, 2016).

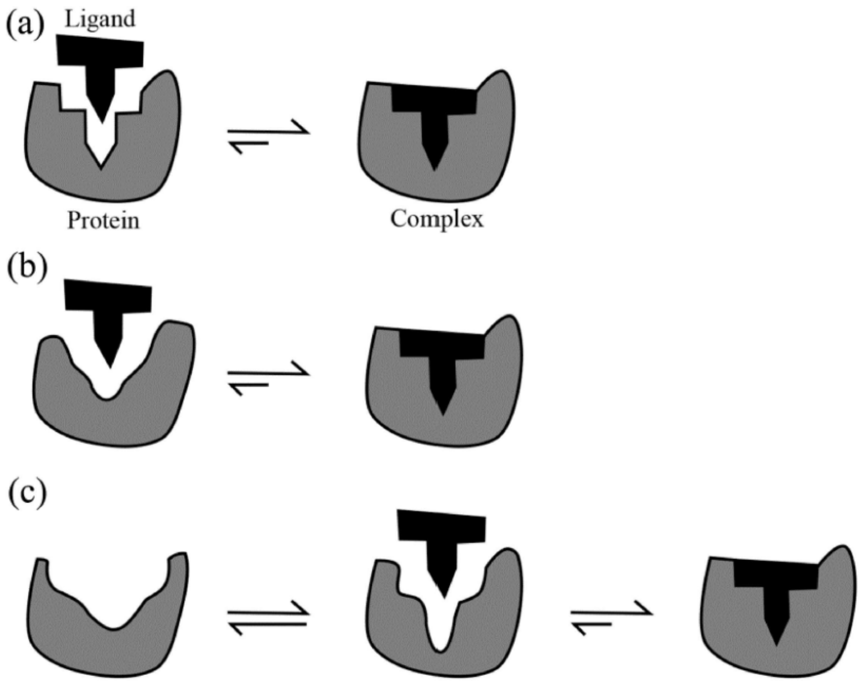


Figure 1.9. Schematic illustrations of the three protein-ligand binding models
(**a**) Lock-and-key (**b**) Induced fit and (**c**) Conformational selection. Figure taken from Du et al., 2016

1.4.2 Binding-site definition

An important consideration in defining ligand-binding, is the definition of a binding-site, particularly a definition which goes further than simply a region on a protein that binds a molecule with affinity and specificity. The tenth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP10) provided a more objective definition - all protein residues in the target structure having at least one (non-hydrogen) atom within a certain distance (dij) to a biologically relevant ligand atom (Gallo Cassarino et al., 2014):

Equation 1.1. Binding-site definition for the CASP10 FN category(*10th Community Wide Experiment on the Critical Assessment of techniques for protein structure prediction*, 2012)
https://predictioncenter.org/casp10/index.cgi?page=format#FN
The equation below shows a binding-site is defined, where, **dij** is the distance between a residue atom ***i*** and a ligand atom, ***j***, ***ri*** and ***rj*** are the Van der Waals radii of the involved atoms, while ***c*** is a tolerance distance of 0.5 Å (Gallo Cassarino et al., 2014).

$$d_{i.j} \leq r_i + r_j + c$$

1.5 Protein-protein interactions

Thus far, the focus has been protein-ligand interactions and whilst the biochemistry of these interactions are crucial, especially in drug development, and will form the focus of this thesis, protein-protein interactions are also significant both physiologically and biochemically. In general, protein interactions cover a full range of interactions, from rigid to dynamic, weak to strong, obligate and non-obligate (Nooren & Thornton, 2003) Figure 1.10 shows examples of these interactions. In obligate protein-protein interactions, the protomers are not found as stable structures on their own *in vivo*. Non-obligate interactions are those whose protomers exist independently (Bera & Ray, 2009). Additionally, protein complexes can also be categorised as permanent and transient according to lifetime *in vivo* (Bera & Ray, 2009). Permanent interactions are stable and exist in complexed form, whereas a transient interaction associates and dissociates *in vivo* (Nooren & Thornton, 2003). Enzyme-inhibitor and antigen-antibody are composed of proteins that are required to bind tightly and permanently and are examples of naturally occurring protein complexes (La *et al.*, 2013). Individual subunits or monomers of these complexes are individually unstable and hence non-functional. In comparison, the heterotrimeric G protein dissociates into Gα and Gßγ subunits upon guanosine triphosphate (GTP) binding, in comparison when bound to guanosine diphosphate (GDP) a stable trimer is formed, thus showing strong transient associations require a molecular trigger to shift the oligomeric equilibrium (Nooren & Thornton, 2003). In general, proteins involved in signaling pathways have a mechanism for dissociation after binding and thereby assist in regulating protein activity at specific times and are transient (La *et al.*, 2013).

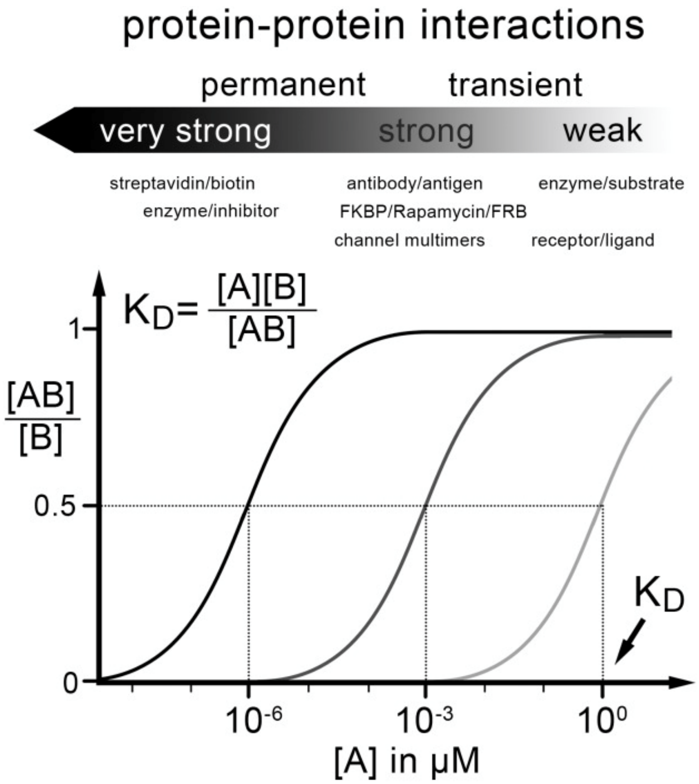


Figure 1.10. Kinetics and affinities of protein-protein interactions
The dissociation constant (K_D) is inversely correlated to the binding affinity. Protein-protein interactions can exist as permanent or transient interactions. Figure taken from (Xing *et al.*, 2016)

Both direct and indirect protein-protein interactions are essential for performing and regulating cellular activities e.g. signal transduction, cell-cycle, morphological differentiation, cell motility, transcription and translation (Bertoni *et al.*, 2017) and can occur between identical or non-identical chains i.e. homo- or hetero-oligomers (Nooren & Thornton, 2003).. Understanding the type of interactions has significant implications for understanding the nature and function of protein-protein interactions (La *et al.*, 2013). The structure and affinity of a protein-protein interaction is related to its biological function, physiological environment and control mechanism and may have evolved to optimise functional efficacy (Nooren & Thornton, 2003).

As with protein-ligand interactions (see Section 1.4), most proteins are very specific in their choice of binding partner with different surface properties. However, some can be multispecific with multiple (competing) binding partners on coinciding or overlapping interface

(Nooren & Thornton, 2003). The prediction of small ligands is considered easier than discovering protein-protein interaction modulators as proteins which bind small molecules generally contain a well-defined ligand-binding site that small molecules interact with (Santos *et al.*, 2017). Designing a small molecule to bind to a protein-protein interface has the following problems:

1. Protein-protein interactions occur on the interface of a specific domain where either two identical proteins or different proteins are in contact (Lu *et al.*, 2020) and the interface area of the interaction usually reaches 1500-3000Å (Jones & Thornton, 1996), whereas for a receptor-ligand contact are is small at 300-1000Å and is hydrophobic (Pagadala, Syed and Tuszynski, 2017).
2. Protein-protein interface tends to be flat and contains few grooves or pockets making it difficult for small molecules to bind (Buchwald, 2010b; Lipinski et al., 1997).
3. Amino acid residues involved in protein-protein interactions are either continuous or discontinuous resulting in high-affinity protein binding. Thus, making it harder for smaller molecules to inhibit such an interaction (Ivanov, Khuri and Fu, 2013).
4. Protein-protein interactions lack endogenous small molecule ligands for reference to act as a starting point (Ivanov, Khuri and Fu, 2013).
5. Traditional small molecule drugs have a molecule weight of 200-500 Da and in comparison drugs acting on protein-protein interactions have higher molecular weights of >400 Da (Buchwald, 2010a). Thus, making it difficult to apply Lipinski’s rule of five (<5 hydrogen bond donors, <10 hydrogen bond acceptors, <500 Da molecular mass and a octanol-water partition coefficient (log *P*) that does not exceed 5) (Lipinski *et al.*, 1997)

Despite the above challenges, identification of hot spots which is 600Å and usually located near the protein-protein interface are able to assist in understanding protein-protein interactions and can identify ligands. Hot spots (Shangary & Wang, 2009) are identified through a point mutation experiment where amino acid residues in protein-protein interactions are mutated to alanine and the change of the binding-free energy is measured to determine the residues that contributes significantly to the binding-free energy (Lu *et al.*, 2020). Experimentally, dual polarisation interferometry is a capable of capturing the conformational changes of proteins and thereby assist in understanding the behaviour of proteins in terms of structure and function (Cross *et al.*, 2003).

The function of proteins is multi-faceted; their role is not just limited to receptors, and the interesting of protein-protein interactions *in vivo*, has led to proteins being harnessed as biologic drugs in the treatment of cancer and autoimmune disease. The use of proteins as drugs is leading to more personalised treatment of diseases and in some cases has transformed treatment.

A prime example, of the latter is trastuzumab (Herceptin®), a monoclonal antibody which is used to treat human epidermal growth factor receptor 2 (HER2) positive breast cancer and is a therapeutic IgG. Human epidermal growth factor receptor 2 is a member of the erb epidermal growth factor receptor tyrosine kinase family and is found to be overexpressed in 20-30% of human breast cancers (Harries & Smith, 2002). When HER2 is overexpressed multiple HER2 heterodimers are formed and cell signalling is enhanced which activates multiple signalling pathways to stimulate cellular migration and cell proliferation, resulting in malignant growth (Harries & Smith, 2002; Wolf-Yadlin et al., 2006). Trastuzumab binds to subdomain IV of the HER2 N-terminal extracellular domain and induces apoptosis in breast cancer cells via antibody-dependent cellular cytotoxicity (ADCC) (Rubin & Yarden, 2001).

Trastuzumab, like all humanised monoclonal antibodies and some chimeric monoclonal antibodies are based on the immunoglobulin G family of antibodies, in particular IgG1 subclass. There are five main classes of immunoglobulins these are IgM, IgG, IgA, IgD and IgE isotypes with the IgG being split into four subclasses, IgG1, IgG2, IgG3 and IgG4, each with its own biologic properties and IgA can similarly be split into IgA1 and IgA2 (Schroeder & Cavacini, 2010).

The IgG1 is preferred for humanised monoclonal antibodies because, this type of immunoglobulin has the longest half-life of all immunoglobulin isotypes, exhibits more pronounced effector functions compared with other subclasses of this type and classes of immunoglobulins and it is the most extensively studied class of immunoglobulins (Schroeder & Cavacini, 2010). As a result of studying the effects of immunoglobins, several antibody biotherapeutics have been developed aided by homology modelling (Schwede *et al.*, 2009). In 2007, of the 21 antibodies on the market, it was estimated that 11 were as a result of computational design of humanised constructs via homology modelling with Herceptin being one of the examples (Schwede *et al.*, 2009).

The utilisation of immunoglobulins to produce therapeutic monoclonal antibodies shows the importance of understanding *in vivo* interactions and the impact it can have on drug development. This is not just limited to protein-protein interactions but can also be relevant to protein-ligand interactions.

1.6 Investigating protein-ligand interactions *in silico*

Prediction of ligand-binding sites from protein structure has many applications, mainly being elucidation of protein function (Dorokhov *et al.*, 2016). This can be further expanded to protein-ligand docking to new compound screening and drug design (Kriválc & Hoksza, 2018).

As mentioned previously, in Section 1.2.2 *in silico* methods are used to address the problem with the knowledge gap on structures, bioinformatic approaches that utilise information from existing protein-ligand complexes are becoming increasingly important, because ligands that bind to a protein are pivotal to understanding protein function. In these approaches, an assumption is made that similar binding sites are likely to bind similar ligands (Wass et al., 2010). This rests on the premise that a known ligand of one protein can be transposed to a similar binding site in another protein, that was previously known to bind the ligand (Konc & Janežič, 2014). In order for these methods of predictions to be successful, the 3D structure of the protein; more specifically the protein-binding site needs to be determined. There are two general approaches for doing this; standard sequence alignments (sequence-based) or sequence to structure alignments (structure-based) (Konc & Janežič, 2014).

Investigating ligand interactions is quite varied and mainly stems from the way the protein structure is predicted and there exists roughly two ways to do this; structure-based and sequence-based (Dukka, 2013). Selection of homologous sequences is a critical step in both sequence-based and structure-based approaches for the prediction of a protein functional site. It has been shown that certain degree of sequence divergence is required in multiple sequence alignment (MSA) for the identification of functional sites (Konc & Janežič, 2014) as homologous proteins do not always share the same function but can be derived from a common ancestral protein (Nemoto & Toh, 2012). The most effective methods in the prediction of functional regions of a protein are detection of conserved residue clusters (e.g. ConSurf (Glaser *et al.*, 2003)) in the tertiary structure. In these methods the homologous amino acid sequence of prediction target are collected and a MSA of the sequence is constructed (Nemoto & Toh, 2012). Then, the conserved residues are identified among all the sites in the MSA and these are assigned to corresponding residues on the tertiary structure. The cluster of conserved residues on the structure are predicted as the functional regions (Nemoto & Toh, 2012).

Sequence comparison is used to infer homology and collect evidence about membership in a given family. The key requirement is to properly choose similarity measures and related cut-off values in order to avoid false positives and false negatives. If two sequences diverge, it becomes impossible to find annotated homologs. This is termed ‘global sequence alignment’ and relies on the evolutionarily related segments of two proteins, which could consist of binding sites and domains (Krivák & Hoksza, 2018). The first global sequence alignment method was developed by Needleman and Wunsch in 1970 (Altschul *et al.*, 1990) and consists of three phases; initialisation (assign values for the first column), fill (aka induction and the entire matrix is filled with scores) and trace-back (recover alignment from the matrix). Since the development of the Needleman and Wunsch algorithm, sequence conservation has been used to predict ligand-binding sites (Needleman & Wunsch, 1970; Berezin et al., 2004). The main strength of sequence-based approaches for prediction of binding sites is that methods that utilise this approach have the ability to determine ligand-binding motif in proteins that may not have the same overall fold (Fischer et al., 2008). Homology-based methods require related proteins with significant identity to the query protein to be available in the PDB because the conservation of biochemical function drops rapidly for proteins sharing <35-40% sequence identity (Dukka, 2013). Therefore, a limitation of this approach is that methods do not work for remote homologs (<30% pairwise identity). For sequence-based methods, the homologous sequence of the target sequence is required, and a multiple sequence alignment (MSA) is constructed. Then, using the specific approach, conserved residues are identified among all the sites in the MSA.

The situation is not as clear when it comes to “moonlighting” proteins and it has been postulated (Dukka, 2013) that mass spectrometry protein-expression profiles are likely to become a key method to identify more “moonlighting” proteins. It is known that sequence information alone cannot provide a distinction about the multiple functions of proteins (Jeffery, 2005). The Critical Assessment of Function Annotation 3 (CAFA3) experiment provided a set of 40 “moonlighting” proteins, for which one or more functions are yet undocumented. In previous years, it has been shown that the best methods to predict protein function rely on homologous proteins (Piovesan et al., 2015). An approach to try and address the problem of predicting protein function without sequence similarity is GAS (Guilty by Association on STRING). The principle of GAS is that if a protein physically interacts (association) with other proteins it should share a similar function (quality) (Piovesan et al., 2015).

Structure-based methods require a knowledge of the 3D dimensional structures of related proteins, which can be used as templates (Piovesan et al., 2015). Currently, there are two principle experimental methods for solving structures; x-ray crystallography and nuclear magnetic resonance (NMR) with the former being the preferred process for obtaining high resolution data (Danishuddin & Khan, 2015). Based on the analysis of existing protein-ligand binding sites/complexes it is quite apparent that homologous protein with similar global topology will bind similar ligands and there will be conserved residues (Danishuddin & Khan, 2015). As a result, there are methods utilising both geometric match and energy scores, in addition to evolutionary information, in order to identify binding sites (Dukka, 2013). In general, these methods are broadly classified into geometry based approaches and energetic based approaches. Geometry-based approaches identify binding residues by searching for pockets or cavities in a protein structure whereas, energetic-based approaches identify binding residues by using various interaction energies (Dukka, 2013). Whilst sequence-based and structure-based methods have different approaches, in reality most successful methods are based on a combination of approaches (Dukka, 2013).

Based on the large number of solved protein structures in databases like PDB, it is possible to develop methods based on structure alignment of proteins. Methods in this area can be classified, broadly as global structure alignment and local structure alignment based methods (Krivák & Hoksza, 2018). Global structure alignment based approaches are based on the observation that there is a tendency of certain protein folds to bind substrates at a similar location. This observation suggests distantly homologous proteins can have common binding sites and, if that is the case, then it should be possible to identify ligand-binding sites for structures requiring prediction (Dukka, 2013). In comparison, local alignment approaches are suited to detect locally conserved patterns of functional groups, which often appear in ligand-binding sites and have relevant involvement in ligand binding (Dukka, 2013). COFACTOR is a method which covers both global and local structure alignment based approaches (Dukka, 2013). This method utilises the amino acid sequence and then generates a 3D structure model for the protein in question using the I-TASSER method (Roy & Zhang, 2012). Subsequently, information based on the global structure similarity to the query protein is obtained using the TM-align structure alignment program, to identify template proteins with bound ligands in PDB (Dukka, 2013).

In contrast, sequence-based methods exploit sequence conservation or the tendency of functionally or structurally important sites to accept fewer mutations relative to the rest of the protein (Dukka, 2013). ConSurf is an example of a method which provides visualisation of sequence conservation values on the surface of a protein structure (Capra *et al.*, 2009).

The natural step, after determination of protein structure, is the prediction of ligand-binding sites and an important consideration when investigating protein-ligand interactions is whether the ligand is biologically relevant. The most direct way to investigate the biological relevance of a ligand is by manual verification (Capra *et al.*, 2009). Verification can primarily consist of reading literature, however given the growth of proteins this can be time consuming. Additionally, novel proteins may not have adequate amounts of literature available to deduce the biologically relevant ligands. As a result, there has been a need to develop automatic procedures to select biologically relevant ligands based on proteins available in PDB (Yang et al., 2012), such as; FireDB (Lopez et al., 2007), LigASite (LIGand Attachment SITE) (Dessailly *et al.*, 2007), Binding MOAD (Mother of All Databases) (Benson *et al.*, 2007), PDBbind (Wang *et al.*, 2004), BindingDB (Liu *et al.*, 2007) and BioLiP (Yang et al., 2012), which are each described in further detail and a summary is shown in Table 1.4. For the Critical Assessment of protein Structure Prediction (CASP) competitions, biologically relevant ligands were defined using information from the literature, Swiss-Prot ligand annotations (Yang et al., 2012), sequence conservation of functionally important residues and information from homologous structures (Magrane & UniProt Consortium, 2011).

Table 1.4. Availability of methods to predict biologically relevant ligands
Table 1.4 below is a summary of the resources available to predict biologically-relevant ligands

Method	Year of first publication	Summary
PDBbind (Wang *et al.*, 2004)	2004	Ligand-binding affinity database for protein-ligand complexes with known 3D structures
FireDB (Lopez et al., 2007)	2007	Selects ligands based on a mapping between inorganic ligands and GO annotations
Binding MOAD (Benson *et al.*, 2007)	2007	Ligand-binding affinity database that selects ligands based on a combination of automated procedure and manual validation
BindingDB (Liu *et al.*, 2007)	2007	Experimentally determined binding affinities of protein-ligand complexes
LigASite (Dessailly *et al.*, 2007)	2008	Consists exclusively of biologically relevant binding sites for each protein with and least one apo- and one holo- structures
BioLiP (Yang et al., 2012)	2012	Semi-manual curated database of biologically relevant protein-ligand interactions

FireDB is a database for functional information on proteins with known structures and is orientated towards small molecule ligands. Therefore, interactions with proteins, DNA and RNA are not considered and large ligands where the number of ligand atoms is 2/3 or greater than the number of protein atoms are also rejected (Gallo Cassarino et al., 2014). Ligands are selected based on a mapping between inorganic ligands and Gene Ontology (GO) annotations. Nevertheless, FireDB does have limitations; as inorganic ligands which are biologically relevant can be missed off if there is no GO annotation or no mapping for the ligand (Lopez et al., 2007).

LigASite consists exclusively of biologically relevant binding sites in proteins for which at least one *apo*- and one *holo*- structure are available. Apo is the structure of a protein with unbound ligand(s) and *holo* is the structure of the protein bound to its ligand(s) (Yang et al., 2012). There is a clear advantage of having both the *apo*- and *holo*- structures available and that is the recognition that the structure may change upon binding of its ligand. Additionally, LigASite is used for benchmarking but does have strict requirements (Seeliger and de Groot, 2010). Ligands are selected if they have >10 heavy atoms on the basis that biologically irrelevant molecules in PDB files are generally very small (Yang et al., 2012) and have >70 inter-atomic contacts with the protein atoms. As a result this may miss metal ion biological ligands (Dessailly *et al.*, 2007).

Binding MOAD selects ligands based on a combination of automated procedure and manual validation. Each structure is hand curated by reading the crystallography paper, which presents the structure in literature and is used to validate ligands and acquire binding affinities (Benson *et al.*, 2007). Binding MOAD contains all appropriate protein-ligand complexes such as; protein-ligand, protein-cofactor and protein-ligand cofactor. The database is also able to present complexes when no binding data is available (Benson *et al.*, 2007). However, as with LigASite, Binding MOAD excludes metal ions and additionally small DNA/RNA molecules (Benson *et al.*, 2007). In comparison, PDBind has less strict requirements than Binding MOAD, such as the inclusion of DNA/RNA molecules and peptides although if there is no binding data reported in the literature then the complex is excluded from the databases (Yang et al., 2012).

BindingDB collects binding data directly from the literature, focusing mainly on proteins that are drug- targets or candidate drug-targets and the structure needs to be present in the PDB. This restriction allows BindingDB to complement, rather than overlap other binding databases (Liu *et al.*, 2007). As with PDBind, if the information is not available on the ligand in literature, then it is excluded from the database.

Most of the existing databases miss biologically relevant ligand-protein interactions, which are important for protein function annotations. Therefore, there was a need for a comprehensive database of biologically relevant ligand-protein interactions collected from the PDB. BioLiP contains both computational and manual examinations to have a precise assessment of ligand entries into the database (Yang et al., 2012). Each entry in the BioLiP database contains a comprehensive list of annotations on ligand- binding residues, ligand-binding affinity, catalytic site residues, Enzyme Commission (EC) numbers, GO terms and cross-links to other popular databases. In order to annotate the function of uncharacterised proteins; a new algorithm called COACH was used to predict ligand-binding sites from either protein sequence or 3D structure. A ligand is deemed biologically relevant using the BioLiP database if it interacts with the protein and plays a biological role; consisting of inhibitor, activator and substrate analog (Yang et al., 2012). Due to the use of several unique aspects, for example a four-step hierarchical procedure to automatically verify the biological relevance of a ligand, comprehensive function annotation and a new reliable algorithm COACH to predict ligand-binding sites; BioLiP will be used in order to assist with the prediction of biologically relevant ligands and ultimately protein function by FunFOLD3.

1.6.1 BioLiP: database for biologically relevant ligand-protein interactions

The biological function of a protein may have several different meanings; a protein can function as a catalyst in chemical reactions, as a transporter for materials across a cell, receiving and sending chemical signals, responding to stimuli and providing structural support (Roche et al., 2012). Most of these functions are determined by interactions with other proteins or small molecules. Therefore, interfaces/interactions between proteins and/or small molecules are critical to understanding function (Liu *et al.*, 2018). The role of a ligand can be in the initiation of a process following binding of the ligand molecule to a protein molecule or conversely block the initiation of activity by occupying the binding-pocket (Alberts et al., 2002). Typically, ligands can be thought of as a signalling molecule and will bind to a specific site on a protein or other molecule (Du *et al.*, 2016). As a result, a ligand will be an extracellular signalling molecule, so a small hydrophobic molecule that can enter a cell and bind to proteins within a cell or a water-soluble, polar molecule that binds to proteins outside the cell (Alberts et al., 2002).

When predicting ligands, a key question is whether ligands are biologically relevant. Ligands associated with PDB entries are not always biologically relevant and can be 'ligands' left over from crystallisation conditions or additives for solving protein structures (Liu *et al.*, 2018). In order to reduce the prediction of non-biologically relevant ligands, FunFOLD3 utilises BioLiP – a semi manually curated database for biologically relevant protein interactions (Yang et al., 2012). Establishing the interactions between a protein and its ligand aids in the understanding the function of proteins. As FunFOLD3 uses templates from PDB it is important to determine the biological relevance, if any, of predicted ligands.

Each entry in BioLiP contains a comprehensive list of annotations on; ligand-binding residues in the database, ligand-binding affinity, catalytic site residues, EC numbers, GO terms and cross-links to other popular databases. To assist with annotation of the function of uncharacterised proteins, another algorithm COACH is used to predict ligand-binding sites from either protein sequence or 3D structure. COACH is a consensus-based approach for ligand-binding site prediction that combines the results from five methods; COFACTOR, FINDISTE, ConCAVITY, TMSITE and SSITE (Yang et al., 2012). SSITE is used to identify the ligand-binding information based on the sequence profile-to-profile search of the target against BioLiP library where the hits of the highest E-value is returned. All five methods are used to generate the binding prediction and the consensus hits from multiple searches are selected (Yang et al., 2012).

The BioLiP database consists of three steps (Yang et al., 2012):

Step 1: For each entry available in the PDB, the 3D structure is downloaded and the modified residues are translated to standard residues based on the record ‘MODRES’ in the PDB structure file

Step 2: Ligands, defined as small molecules are extracted from the PDB file. Three types of ligand molecules are collected in the BioLiP database: the molecules from the HETATM record (excluding water and modified resides), small DNA/RNA and peptides with <30 residues. Metal ions are considered as potential biologically relevant ligands paradoxically, whilst some are first listed as possible artifacts, others can be deemed as biologically relevant ligands (Yang et al., 2012).

Step 3: Each ligand molecule is submitted to a composite automated and manual procedure to decide biological relevance. If the ligand is seemed biologically relevant it is deposited into the BioLiP database. Additional information on ligand-binding affinity, catalytic site resides, EC numbers, GO terms and crosslinks to PDB, UniPort, PDBsum, PDBe and PubMed databases are also collected and deposited into BioLiP (Yang et al., 2012).

The definition of a biologically relevant ligand is if the ligand in question interacts with the protein and plays certain a biological role, such as co-factor, inhibitor, activator or substrate.

The validation of a biologically relevant ligand consists of both manual and automated procedure to eliminate possible false positives, such as crystallisation additives (Yang et al., 2012).

Figure 1.11 outlines the four step automated filtering process for BioLiP. The first step, is if the candidate ligand is in the artifact list and appears >15 times in the same structure file, then it is likely to be a crystallisation additive and is considered as biologically irrelevant (Yang et al., 2012). During the second step, the contacts between the receptor and ligand atoms are computed. For a receptor residue, if the closet atomic distance between the residue and the ligand is within certain distance cutoff, then the residue is defined as a ligand-binding residue. The cut off is set at 0.5 plus the sum of the Van der Waal's radius of the two atoms. If the number of binding site residues is less than two or all the binding site residues are consecutive, it is deemed biologically irrelevant because most biologically relevant ligands are usually tethered by multiple residues, which are further apart in the sequence space (Yang et al., 2012). Step three, if the ligand is not present in the artifact list, then it is considered as biologically relevant and kept for manual verifications (Yang et al., 2012). Step four, review of literature e.g. PubMed abstracts to filter out biologically irrelevant ligands. Automatically regarding ligands as artifacts if they fit this criteria will miss some ligands (false negatives) that are biologically relevant (Yang et al., 2012). Step five, involves manually verification across the literature and other databases (Yang et al., 2012).

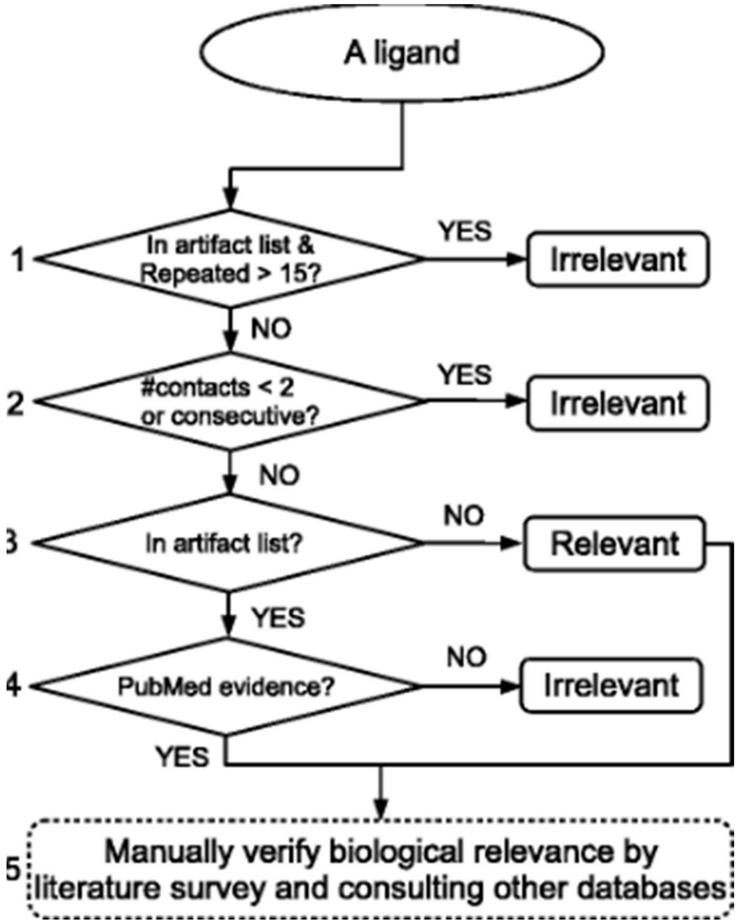


Figure 1.11. Flowchart for the biological relevance assessment of ligand molecules. Figure taken from Yang et al., 2012

The most convenient and well-known computational method for function prediction is based on the detection of significant sequence similarity to gene products of known gene function (Yang et al., 2012) and has been shown that computational prediction methods also play a key role in the prediction of cancer-gene function, as traditional experimental approaches are laborious and expensive (Hu *et al.*, 2007). Basic local alignment search tool (BLAST), is a rapid sequence alignment algorithm for homology searching of sequence libraries and works on the assumption that proteins with a similar sequence probably have similar biological properties. However, there is an important restriction with this simplistic approach – only functions tied directly to sequence, such as enzymatic activity, can be predicted accurately (Hu *et al.*, 2007). Position Specific Iterative-BLAST (PSI-BLAST) further improved the speed and sensitivity of the BLAST algorithm (Hu *et al.*, 2007).

It could be assumed the use of in silico methods removes the need for in vivo methods for investigating protein interactions. However, in silico methods can be used before in vivo methods to provide initial confirmation of interactions. Guarienti et al., 2015 used computational analysis methods to predict if recombinant human erythropoietin could interact with zebra fish erythropoietin receptors in vivo. The computational analysis enabled the investigation into the functional similarity between human and zebra fish erythropoietin receptors. This showed recombinant human erythropoietin could recognise and bind to zebra fish erythropoietin receptors in the same way it binds to human erythropoietin receptors. When recombinant human erythropoietin was used in vivo, results showed the zebra fish could be utilised as an animal model to study safety and efficacy of biologics. Thus, supporting the use of computational methods in protein interactions.

1.6.2 Experimental models of protein-ligand binding affinity

Various experimental techniques can be used to investigate protein-ligand binding, with X-ray crystallography, nuclear magnetic resonance (NMR), small-angle X-ray scattering and cryo-electron microscopy also used in the determination of 3D protein structure and mentioned previously in Section 1.2.1. Crystal structures of protein-ligand complexes provide a detailed view of their spatial arrangement and interactions (Schlichting, 2005). Protein complexes with reactive short-lived ligands e.g. chemical or binding reactions are determined using X-ray diffraction techniques such as Laue method (Du *et al.*, 2016).

In the Laue method, a stationary single crystal is bathed in a beam of “white” radiation and using general wave optical principles a 3D lattice concept are used to deduce three equations which must be simultaneously satisfied to explain that X-rays are scattered selectively in certain well-defined directions (Smallman & Ngan, 2014). A transmission photograph or a back reflection photograph is taken and the Laue path is able to indicate the symmetry of the crystal (Smallman & Ngan, 2014). Advantages of Laue include a 'niche of excellence' in the study of cyclic, ultra-fast, light-triggered reactions (Bourgeois & Royant, 2005) and speed of data collection that may be achieved while maintaining an adequate signal-to-noise ratio (Ren *et al.*, 1999). A disadvantage is the Laue method is best suited for visualisation of intermediate states that cannot be cleanly trapped by cryocooling (Ren *et al.*, 1999).

All provide atomic-resolution or near-atomic-resolution structures of the unbound proteins and the protein-ligand complexes, which can be used to study the changes in structure and and/or dynamics between the free and bound forms as well as relevant binding events (Du *et al.*, 2016).

1.7 FunFOLD webserver

FunFOLD is a template-based method for protein-ligand binding prediction (Roche et al., 2011, 2013) and uses an automatic approach for cluster identification and residue selection (Roche et al., 2015). The main requirement for FunFOLD is a 3D model, amino acid sequence and a list of templates as inputs (Roche et al., 2011) FunFOLD3 will provide:

1. A list of ligand-binding site residues in the target sequence that are most likely to bind a ligand
2. A list of putative binding ligand(s)
3. 3D models of the likely protein-ligand interactions
4. List of likely GO terms and EC identifiers

A flow diagram of the FunFOLD2 prediction server pipeline (pre-dates FunFOLD3) is illustrated in Figure 1.12 below with a simplified flowchart in Figure 1.12A and a detailed overview in Figure 1.12B:

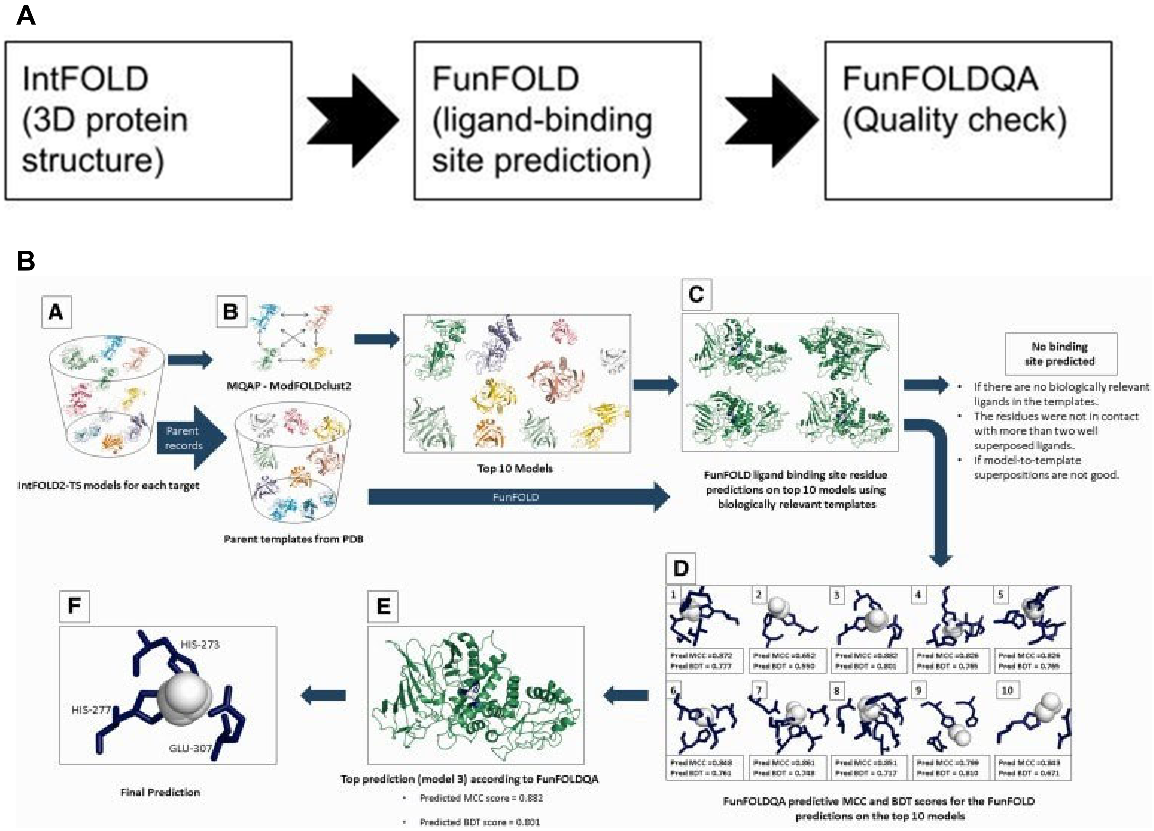

A
IntFOLD (3D protein structure)
FunFOLD (ligand-binding site prediction)
FunFOLDQA (Quality check)
B
A
IntFOLD2-TS models for each target
Parent records
B
MQAP - ModFOLDclust2
Parent templates from PDB
Top 10 Models
FunFOLD
C
FunFOLD ligand binding site residue predictions on top 10 models using biologically relevant templates
No binding site predicted
If there are no biologically relevant ligands in the templates.
The residues were not in contact with more than two well superposed ligands.
If model-to-template superpositions are not good.
D
1
Pred MCC =0.872
Pred BDT = 0.777
2
Pred MCC =0.652
Pred BDT = 0.550
3
Pred MCC =0.882
Pred BDT = 0.801
4
Pred MCC =0.826
Pred BDT = 0.785
5
Pred MCC =0.826
Pred BDT = 0.765
6
Pred MCC =0.848
Pred BDT = 0.761
7
Pred MCC =0.861
Pred BDT = 0.748
8
Pred MCC =0.851
Pred BDT = 0.717
9
Pred MCC =0.799
Pred BDT = 0.810
10
Pred MCC =0.843
Pred BDT = 0.671
FunFOLDQA predictive MCC and BDT scores for the FunFOLD predictions on the top 10 models
E
Top prediction (model 3) according to FunFOLDQA
Predicted MCC score = 0.882
Predicted BDT score = 0.801
F
HIS-273
HIS-277
GLU-307
Final Prediction

Figure 1.12. Flow diagram of the FunFOLD2 prediction server pipeline.
(**A**) A number of alternative models are built for the target sequence using the IntFOLD2-TS protocol. (**B**) The FunFOLD2 pipeline then uses ModFOLDclust2 to determine the top models for each target. (**C**) The FunFOLD algorithm is subsequently used to predict ligand-binding site residues for the top models. (**D**) The quality is assessed for the resultant FunFOLD predictions, using our ligand-binding site quality assessment tool, FunFOLDQA (**E**) The predicted MCC and BDT scores [according to FunFOLDQA] are provided, along with the propensity of which ligand type the binding site is most likely to contain, along with ligand functional propensity. (**F**) Final prediction. Figure taken from Roche, et al., 2011

Briefly, the FunFOLD method for predicting ligand-binding site residues is based on the concept that, target proteins (e.g. CASP targets), may contain similar binding sites as those identified in templates from the PDB, which have the same fold (Roche et al., 2013). The FunFOLD server predicts protein-ligand binding sites from a single sequence via 3D structures built using the IntFOLD server (Roche et al., 2013). FunFOLD uses the predicted 3D model of the target protein under analysis and using structural superpositions of this model and related templates with bound ligands in order to identify putative contacting residues (Roche, Tetchner and McGuffin, 2011). This is based upon the concept that ligand containing templates from the PDB with the same folds as the target protein may contain similar binding sites. For further information on the prediction of ligand-binding sites by FunFOLD refer to Section 1.7.2.

1.7.1. FunFOLDQA

Quality assessment gained attention to become an integral part of tertiary structure prediction (McGuffin *et al.*, 2019) and it was later proposed that similar metrics should become an integral part of the ligand binding site residue predictions. Several QA tools exist and the Cheng group have numerous QA tools such as MUTLICOM, APOLLO, QMEAN,QMEANolust, ProQ, Kalman & Ben-Tal and DISCERN (McGuffin & Roche, 2011).

FunFOLDQA feature scores are derived from data generated by running the FunFOLD method. It is worth mentioning that similar data are also produced by the majority of the top structure based binding site residue prediction methods (Roche et al., 2012). Following identification of ligands from FunFOLD, ligands are then assigned to clusters using an agglomerative hierarchical clustering algorithm that identifies each continuous mass of contacting ligands, thereby indicating a putative binding pocket (Roche et al., 2012). Ligands are considered part of the cluster if any of their atoms are in contact with the continuous mass (Roche et al., 2012). Once each continuous mass of contacting ligands was identified, the cluster with the largest number of ligands was selected as the location of the most likely binding pocket (Roche et al., 2012). Determination of which residues are most likely to be the predicted ligand-binding site relies on a residue voting procedure. For a residue to be included in a prediction it must have at least one contact with two or more ligands and at least 25% of the ligands in the cluster (Roche et al., 2012).

1.7.2 FunFOLD3 for the prediction of ligand-binding sites

The FunFOLD algorithm utilises the TMalign method to superpose all identified templates containing biologically relevant ligands with the predicted 3D structure from IntFOLD2-TS (Roche et al., 2015). TM-align is an algorithm for structural alignment between two protein structures (Zhang & Skolnick, 2005). TM-align works by firstly generating optimised residue-to-residue alignment based on structural similarity using a heurtistic dynamic programming iteration (dividing the full sequence into a series of smaller sequences and uses the solutions to the smaller problems to find an optimal solution to the full sequence) (Needleman & Wunsch, 1970; Zhang & Skolnick, 2005). TM-align will provide an optimal superposition of the two structures (for the purpose of this thesis, it will be the observed and predicted structures), built on the detected alignment (Zhang and Skolnick, 2005). Additionally, a TM-score is used to scale the structural similarity. TM-score has a scale of 0-1, with 1 indicating a perfect match between the two structures. Scores <0.2 correspond to randomly chosen unrelated proteins and scores >0.5 assume two proteins generally have the same fold (Xu & Zhang, 2010).

This method is a similar concept to methods developed by the Lee group (Roche et al., 2013) and Sternberg group (Oh et al., 2009). However, the FunFOLD algorithm uses a novel automated method for ligand clustering and identification of binding residues (Wass et al., 2010). This method consists of protein-ligand binding site and quality assessment protocols for the prediction of protein function (the "FN" category in CASP, see link in equation 1) from sequence via structure (Roche et al., 2011).

The input to the FunFOLD3 server is a 3D model of the protein under analysis and a list of template PDB IDs (Roche et al, 2013). Once the 3D model has been inputted into the server, the TMalign method is used to superimpose the template structures of the 3D protein model. Template-model superpositions with a TM-score ≥0.4 are retained (Roche et al., 2011). This is because, TM-scores from 0.4 to 0.6 have previously been shown to mark the transition from unrelated to significantly related folds (Zhang & Skolnick, 2005). Then superpositions are combined and reoriented using a PyMOL script to determine putative ligands (Xu & Zhang, 2010). The next step is ligands are assigned to clusters using agglomerative hierarchical clustering. To determine the ligand binding site residues in the selected binding pocket, a novel residue-voting algorithm is used. Residues are determined to be in contact with a ligand cluster, if the residue is in contact with the ligand cluster (Roche et al., 2013). Ligands are considered to be part of a cluster if the Van der Waals radius is ≤0.5 Å. The most probable ligand-binding site is the site with the largest ligand cluster (Konc & Janežič, 2014).

In 2005, TMalign was developed as an algorithm to identify the best structural alignment between protein pairs combining the TM-score rotation matrix and dynamic programming (Zhang & Skolnick, 2005). TM-score overcomes the problems associated with the root-mean-square deviation (RMSD). Root-mean-square deviation compares the protein structures/models with a specified equivalence between pairs of residues. This is the most commonly used metric in the category, which compares protein structure/models with a specified equivalence between pairs of residues (Matthews, 1975). With RMSD the problem arises from weighting the distances between all residues equally. As a result of a small number of local structural deviations could result in a high RMSD, even when the global topologies of the compared structures are similar (Matthews, 1975). TM-score overcomes this problem by exploiting a variation of Levitt-Gerstein weight factor, which weighs the residue pairs at smaller distances relatively stronger than these at larger distances (Zhang & Skolnick, 2005). The second type of structure comparison compares a pair of structures where the alignment between equivalent residues is not a priori given. TMalign extends the approaches of Levitt & Gerstein and Kihara & Skelnick with the TM-score rotation matrix speeding up the process of identifying the best structure alignments (Zhang & Skolnick, 2005). The TMalign method will also be used to compare the protein models from CASP and the predicted protein models from FunFOLD3 (see Chapter 3).

The latest version of the FunFOLD webserver is FunFOLD3, which incorporates the

FunFOLDQA algorithm (McGuffin & Roche, 2010). This algorithm evaluates the quality of FunFOLD predictions by producing a set of quality assessment scores. These output scores include five sequence- and structure-based features and output from predicted Binding-site Distance Test (BDT) (Roche et al., 2013) and Matthews Correlation Coefficient (MCC) scores, (McGuffin & Roche, 2010) which are used for the assessment of ligand-binding site residue predictions compared with crystal structures. The FunFOLDQA method combines four binding site-dependent protein scores and one structural dependent feature score, using a neural network, trained on either the MCC or BDT metrics to produce local ligand-binding site predictions. The five feature scores are called:(1) BDTalign, (2) Identity, (3) Rescaled BLOSUM62, (4) Equivalent Residue Ligand Distance and (5) Model Quality. BDTalign establishes the distance between residues that are equivalent between the model binding site and the template-binding site. Identity score compares binding site residues between the model- and template-binding site, which are equivalent in 3D space according to their amino acid sequence. Rescaled BLOSUM62 score is similar to the Identity score, but it scores equivalent residues between model and template binding site, using the BLOSUM62 (Eddy, 2004) scoring matrix. BLOSUM (**BLO**cks **SU**bstitution **M**atrix) is a substitution matrix of 2,000 blocks of aligned sequence segments and characterises more than 500 groups of related proteins (Henikoff & Henikoff, 1992). The sequences in each block were sorted into closely related clusters and the frequencies of substitutions between these clusters within a family used to calculate the probability of a meaning substitution. A scoring matrix is required to evaluate the two amino acid residue-pairs in an alignment and is scored according to a match or mismatch, if they occur at the same position. Matches are given a positive score, e.g. +1 and mismatches are given a negative score e.g. -1. Amino acids are grouped according to the chemistry of the side chain (Pertsemlidis & Fondon, 2001). The cut-off values associated with BLOSUM denote the percentage of sequence identity that defines the cluster (e.g. BLOSUM45, BLOSUM62 and BLOSUM80) (Pertsemlidis & Fondon, 2001). Thus, BLOSUM62 would have within each block, the amino acid sequences would be at least 62% identical when the two proteins were aligned (Pertsemlidis & Fondon, 2001) and will interchange with each other and contains the general evolutionary information among the protein families (He *et al.*, 2006). Lower cut-off values allow more diverse sequences into the group and are therefore appropriate for examining more distant relationships (Pertsemlidis & Fondon, 2001). BLOSUM62 is consistent with strong evolutionary pressure to conserve protein function and hence is utilised by FunFOLD3 (O'Connor, 2021). The Equivalent Residue Ligand Distance score, scores the equivalent residues between the model and the template in relation to their distance from the bound ligand. The Model Quality score is the global quality score for the starting model, calculated using ModFOLDclust2 (McGuffin & Roche, 2010; Roche et al., 2011).

Finally, the FunFOLD3 method outputs a putative ligand binding site, putative ligand binding site residues, putative ligands that may bind to the target protein, along with predicted EC numbers and GO (Gene Ontology Consortium, 2015) terms for each target protein (McGuffin & Roche, 2011; Roche et al., 2011). Several other methods for the prediction of ligand-binding sites exist and are shown in Table 1.5. Methods have traditionally been categorised based on their main algorithmic strategy into geometric, energetic, conservation-based, template-based and machine learning/knowledge based (Krivák & Hoksza, 2018). Each method has additional functionality e.g. suggesting possible binding ligands (e.g. FunFOLD3 and GalaxySite), others predict druggability of predicted pockets (Fpocket, DrugSite).

Table 1.5. Availability of existing tools for ligand binding site prediction from protein structure introduced since 2009

Geometric methods, the ligand-binding site is presumed to be located within the largest pocket on the protein surface (Tsujikawa *et al.*, 2016). Energetic methods are based on the concept that a ligand binds the site where the interaction energy with the protein is minimal (Tsujikawa *et al.*, 2016). Template-based methods search for the most similar proteins in a database(s) that have been labelled with ligand-binding sites using structure alignment algorithm and then to transfer the known ligand-binding site from the most similar proteins onto the query protein (Zhao et al., 2020). Consensus methods utilise a multi-strategy approach which combines different methods to potentially perform better than a single-strategy methods (Xie & Hwang, 2012). Protein-ligand docking, involves molecular modeling to predict ligand-protein binding conformations (Zhang et al., 2020). Conservation methods make assumptions that residues located in protein-ligand binding site are usually more important more highly conserved than those located in other parts during evolution (Dai *et al.*, 2011).
Machine learning encompasses both traditional and deep learning prediction methods. Traditional methods focus on machine learning algorithms to carry out both ligand-binding site predictions but also for the binding affinity research (Zhao et al., 2020). Deep learning simulates the learning mechanism of the human brain but still uses algorithms to determine ligand-binding sites (Zhao et al., 2020). Table adapted from Kriválk & Hoksza, 2018

Method	Year of first publication	Type
SiteMap (Halgren, 2009)	2009	Geometric
Fpocket (Le Guilloux et al., 2009)	2009	Geometric
SiteHound (Ghersi & Sanchez, 2009)	2009	Energetic
ConCavity (Capra *et al.*, 2009)	2009	Conservation
3DLigandSite (Wass et al., 2010)	2010	Template
POCASA (Yu *et al.*, 2010)	2010	Geometric
DoGSite (Volkamer *et al.*, 2010)	2010	Consensus
FunFOLD (Roche et al., 2011)	2011	Template
MetaPocket (Zhang *et al.*, 2011)	2011	Consensus
MSPocket (Zhu & Pisabarro, 2011)	2011	Geometric
FTSite (Ngan et al., 2012)	2012	Energetic
LISE (Xie & Hwang, 2012)	2012	Knowledge/conservation
COFACTOR (Roy et al., 2012)	2012	Template
COACH (Yang et al, 2013)	2013	Template
G-LoSA (Lee & Im, 2013)	2013	Template
eFindSite (Brylinski & Feinstein, 2013)	2013	Template
GalaxySite (Heo *et al.*, 2014)	2014	Template/Docking
LIBRA (Hung et al., 2015)	2015	Template
P2RANK (Krivák & Hoksza, 2018)	2015	Machine learning
bSiteFinder (Gao *et al.*, 2016)	2016	Template
ISMBLabLIG (Jian *et al.*, 2016)	2016	Machine learning
DeepSite (Jiménez *et al.*, 2017)	2017	Machine learning

1.7.3 IntFOLD server

As mentioned previously, FunFOLD3 is part of the IntFOLD server, which comprises of five novel methods: IntFOLD-TS, for tertiary structure prediction, ModFOLD, for model quality assessment, DISOclust, for disorder prediction, FunFOLD3, for function prediction by ligand and ligand-binding site prediction and DomFOLD, for prediction of a number of domains and their possible boundaries within a protein sequence (Roche et al., 2011; Roche et al., 2013). The IntFOLD server was designed by the McGuffin group and has been operational since January 2010 and the guiding principles behind the server development were (i) to provide a simple unified resource that makes prediction software accessible to all and (ii) to produce integrated output for predictions that can be easily interpreted (Roche *et al.*, 2011).

Figure 1.13 on the next page demonstrates how the methods within the original IntFOLD server are interdependent, with the output from initial tertiary structure prediction algorithm becoming the input for subsequent methods.

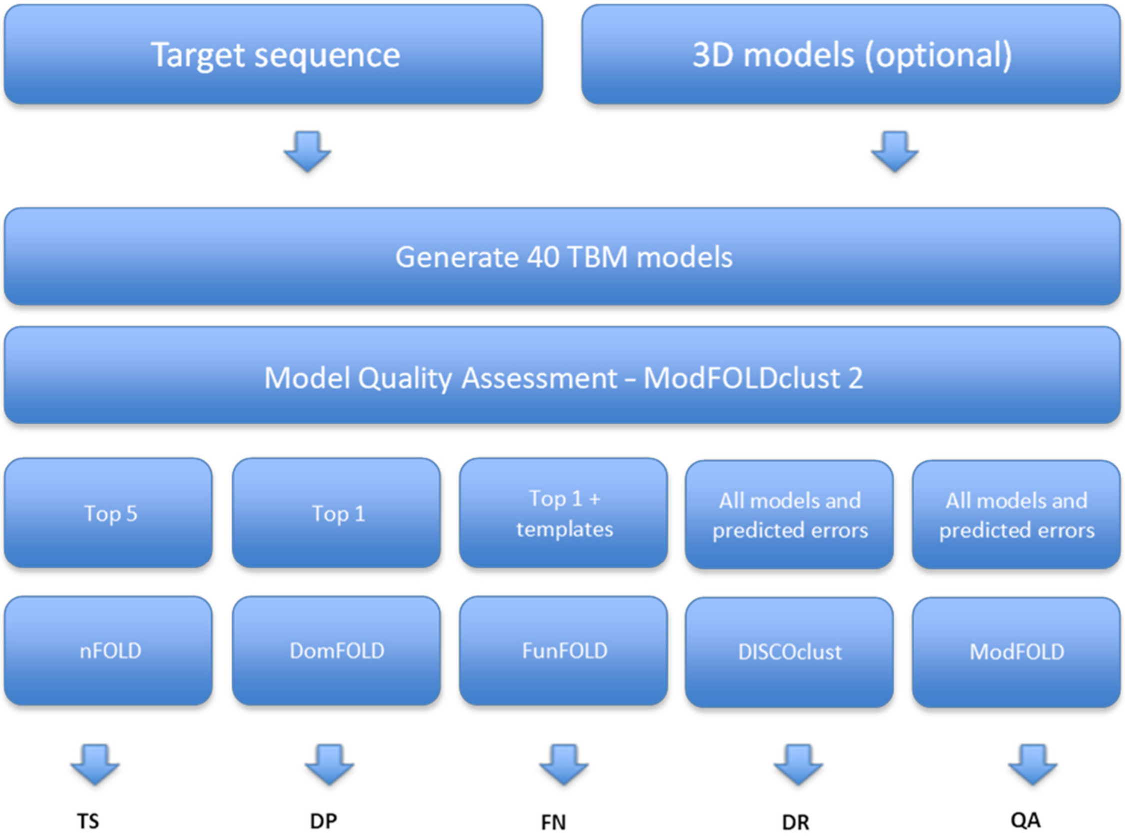


Figure 1.13. Diagram of the software stack implemented for the IntFOLD server
The figure highlights the interdependency of all the different IntFOLD algorithms and highlighting the importance of ModFOLDclust2 as the key algorithm in the pipeline. The models ranked by ModFOLDclust2 are used to produce a resulting output for 3D structure prediction (TS), domain prediction (DP), binding site residue prediction function prediction (FN), disorder prediction (DR) and model quality assessment (QA). Figure adapted from McGuffin & Roche, 2011

The top ranked IntFOLD model and related templates with bound ligands have model-to-template superpositions performed. This aids the identification of putative contacting residues and identifies templates used for model generation that contain biologically relevant ligands, in order to produce ligand binding site residue predictions (Roche *et al.*, 2011). A prototype version was developed during the CASP9 prediction season and incremental improvements to the server have been made since, which has enhanced performance and reliability (Roche *et al.*, 2011).

1.8 Assessing the performance of the FunFOLD3 webserver

There are several community wide prediction experiments such as; Critical Assessment of techniques for protein Structure Prediction (CASP) (Liu *et al.*, 2018), the Continuous Automated Model EvaluatiOn (CAMEO) project (López et al., 2009) and the Critical Assessment of Function Annotation (CAFA) (Radivojac *et al.*, 2013). CAMEO, used to have a ligand-binding section which evaluated FunFOLD predictions called CAMEO-LB. However, this was discontinued in April 2016, and a beta version of CAMEO, CAMEO-3D will include ligand-binding predictions once again. Each of these projects have been essential for the independent benchmarking of the performance of the FunFOLD3 webserver. The focus of this thesis will be about CASP and CAFA competitions and further details are provided in Section 1.8.1 and Section 1.8.2, respectively.

1.8.1 Critical Assessment of protein structure prediction

The Critical Assessment of techniques for protein Structure Prediction was launched in 1994 and is a community-wide experiment for tertiary protein structure prediction taking place biennially (Moult *et al.*, 1995). The only way to objectively assess the usefulness of prediction methods, is to ensure predictions are made without any knowledge of the answers (Moult *et al.*, 1995). The procedure for CASP consists of three parts: (1) the collection of targets for prediction from the experimental community, (2) the collection of predictions from the modelling research groups and (3) the assessment (and discussion) of the results (Moult *et al.*, 1995).

CASP appreciated the difficulty of predictions depended on the extent of the relationship of the target protein to already known structures and predictions were therefore divided into three types (Moult *et al.*, 1995):

1. Comparative modelling
2. Threading or fold identification
3. *Ab initio* predictions

CASP provides structural biologists an opportunity to objectively measure their structure prediction methods independently, as the protein targets are double-blinded and will have experimental structures released imminently or are solved but not released publicly by either X-ray crystallography or NMR spectroscopy and will be available on Protein Data Bank (Moult *et al.*, 1995). This blinding enables benchmarking to be utilised with the purpose of benchmarking to compare predicted structures to experimental structures so ultimately the data can be as robust as it can be and also be trusted.

CASP aimed to show that objective testing of structure prediction methods is both practical and necessary (Moult *et al.*, 1995). Since inception of CASP1 where 27 groups took part to the recent CAP14 in which over 200 groups participated shows the increasing importance of understand structure and function prediction to the community.

As CASP provided an objective measure of not only structure but also protein function by prediction of ligands and ligand-binding residues, it was deemed as an invaluable method to objectively measure FunFOLD3 and identify specific strengths and weaknesses and ultimately highlight where the opportunities are for further development of this method.

1.8.2 Critical Assessment of Function Annotation

The CAFA challenge is a worldwide effort aimed at analysing and evaluating protein function prediction methods (Radivojac *et al.*, 2013). This has begun to provide an objective overview of the state-of-the-art in the field of automatic protein function prediction (AFP). The experiment consists of two tracks (i) the eukaryotic track and (ii) the prokaryotic track. In each track, a set of targets is provided by the organisers.

The CAFA experiment is also responsible for defining new criteria for evaluation, which are (Radivojac *et al.*, 2013):

1.5.1.1 Validation data set used for the blind set
1.5.1.2 Definition of function space through GO terms
1.5.1.3 Scoring metrics for comparing different methods

In comparison to CASP, CAFA's main objective is to gather all AFP researchers to fairly assess and compare the latest computational methods using a centralised and independent assessment (Piovesan et al., 2015).

Protein function can be described in multiple ways, in CAFA the focus is on classification schemes provided by the GO consortium (Kahanda *et al.*, 2015). The CAFA experiment involves a set of proteins lacking experimentally validated functional annotation being released to participants. Proteins are then annotated by the participants with the annotations submitted to assessors (Radivojac *et al.*, 2013). The evaluation of protein function prediction is assessed using the maximum F-measure (Fmax) and considers predictions across the full spectrum from high to low sensitivity. A perfect predictor would be characterised with Fmax = 1 (Radivojac *et al.*, 2013). The approach is not without its limitations however, mainly the penalisation of specific predictions (Radivojac *et al.*, 2013).

There have been changes to CAFA since its inception; CAFA1 had participants make computational predictions using their own AFP method on protein targets missing previous annotations. The computational predictions were compared against annotations to assess the accuracy of each AFP method. The second CAFA challenge; CAFA2 had the exact same concept, except the protein targets consist of both annotated and unannotated proteins. The addition of annotated proteins makes CAFA2 a more realistic representation of function prediction problems, as it better models the accumulation of annotations over time (Radivojac *et al.*, 2013). Whereas, CAFA3 introduced predictions of macromolecular binding sites in a protein (DNA and RNA) and metal binding sites. Unlike CASP, CAFA allows researchers to assess three different methods, however only one model, typically the best algorithm can be officially evaluated.

1.9 Outline of thesis and rationale of study

Chapter 3 of the thesis will analyse FunFOLD3 in the CASP11, CASP12 and CASP13 double-blinded experiment which aims to objectively establish the performance of FunFOLD3 in function prediction with Chapter 4 focusing on IntFOLD4 in CASP12 and Chapter 5 reports the performance of FunFOLDQ in CAFA3. Chapter 6 will explore the use of docking to refine the ligand-binding site using AutoDock Vina. Assessing the performance of FunFOLD3 and FunFOLDQ in two different competitions aids in providing a more complete picture of the server's performance. When FunFOLD3 is utilised in CASP competitions, it mainly informs the user that the protein has a biologically relevant ligand and the 3D structure of the protein, however, while GO terms are predicted as part of the FunFOLD3 output, they do not form part of the analysis for the CASP assessors. In CAFA, the prediction of GO terms enables the user to uniquely and precisely define the features of genes and gene products in a species independent manner (Kahanda *et al.*, 2015). Therefore, when CASP and CAFA results are analysed in combination for the same protein, ligand-binding sites can be identified but also the function of the protein, allowing for the elucidation of any impact on disease. Chapter 7 will be applying FunFOLD3 in the 2020-2021 COVID-19 global pandemic and gaining valuable insights into potential ligand-binding residues.

Translocator Protein kDa (TSPO) is an example of a protein in literature where knowledge solely about a ligand can assist in further information into the role of a protein and also highlights a potential role for FunFOLD3, in that FunFOLD3 can be utilised to identify ligands and ligand-binding site residues and understanding the role of these ligands can assist in providing insight into the function of a protein. TSPO, is a ubiquitous mitochondrial protein containing separate drug and cholesterol binding domains and was previously known as the peripheral-type benzodiazepine receptor because it was identified as a binding site for the benzodiazepine, diazepam. (Pinoli et al., 2015). On PDB, a number of ligands can be identified for TSPO and these range from protoporphyrin IX, formic acid, 1-Oleoyl-R-glycerol to tetraethylene glycol (Papadopoulos et al., 2017). TSPO along with the binding of its ligands is illustrated in Figure 1.14. Due to the diversity of ligands, which bind to TSPO, it would be reasonable to assume there is a diverse role of TSPO, especially if the ligands are biologically relevant. For example, the function of TSPO in haeme biosynthesis is quite obvious due to protoporphyrin IX ligand. Seven GO annotations are associated with TSPO, illustrating the diverse range of functions and is illustrated in Table 1.6.

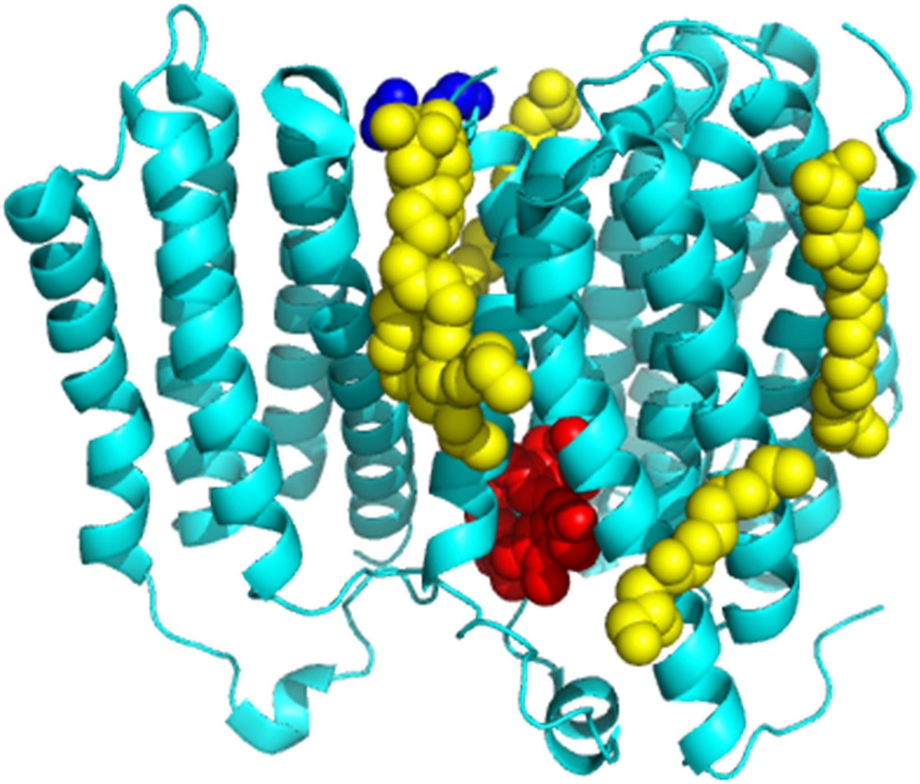


Figure 1.14. Structure of TSPO bound to ligands
Translocator protein 18kDa 2.4 Å (PDB ID 5duo) is shown in cartoon form with the surface highlighted and coloured cyan, with the proroporphyrin IX ligand shown as sphere and coloured red, the formic acid ligand shown as sphere and coloured blue and 1-Oleoyl-R-glycerol shown as sphere and coloured yellow

Table 1.6. GO identifiers and GO term name associated with TSPO
Seven annotations have been identified for TSPO illustrating the diversity of functions

GO identifier	GO term name
GO: 0006783	Haeme biosynthetic process
GO: 0006820	Anion transport
GO: 0006915	Apoptotic process
GO: 0008202	Steroid metabolic process
GO: 0008283	Cell proliferation
GO: 0015485	Cholesterol binding
GO: 0032374	Regulation of cholesterol transport

As can be seen from Table 1.6, the GO terms provide further information into the function of TSPO and demonstrates diversity across a range of biological processes and molecular function. The ancestor charts for GO: 0006783 and GO: 0008283 as shown in Figure 1.15A show the function of TSPO in haeme biosynthetic process is well understood as shown by the number of “branches”, however the exact role of TSPO in cell proliferation (Figure 1.15B) remains to be fully understood and this is supported by the limited data in literature (Berman, Henrick and Nakamura, 2003).

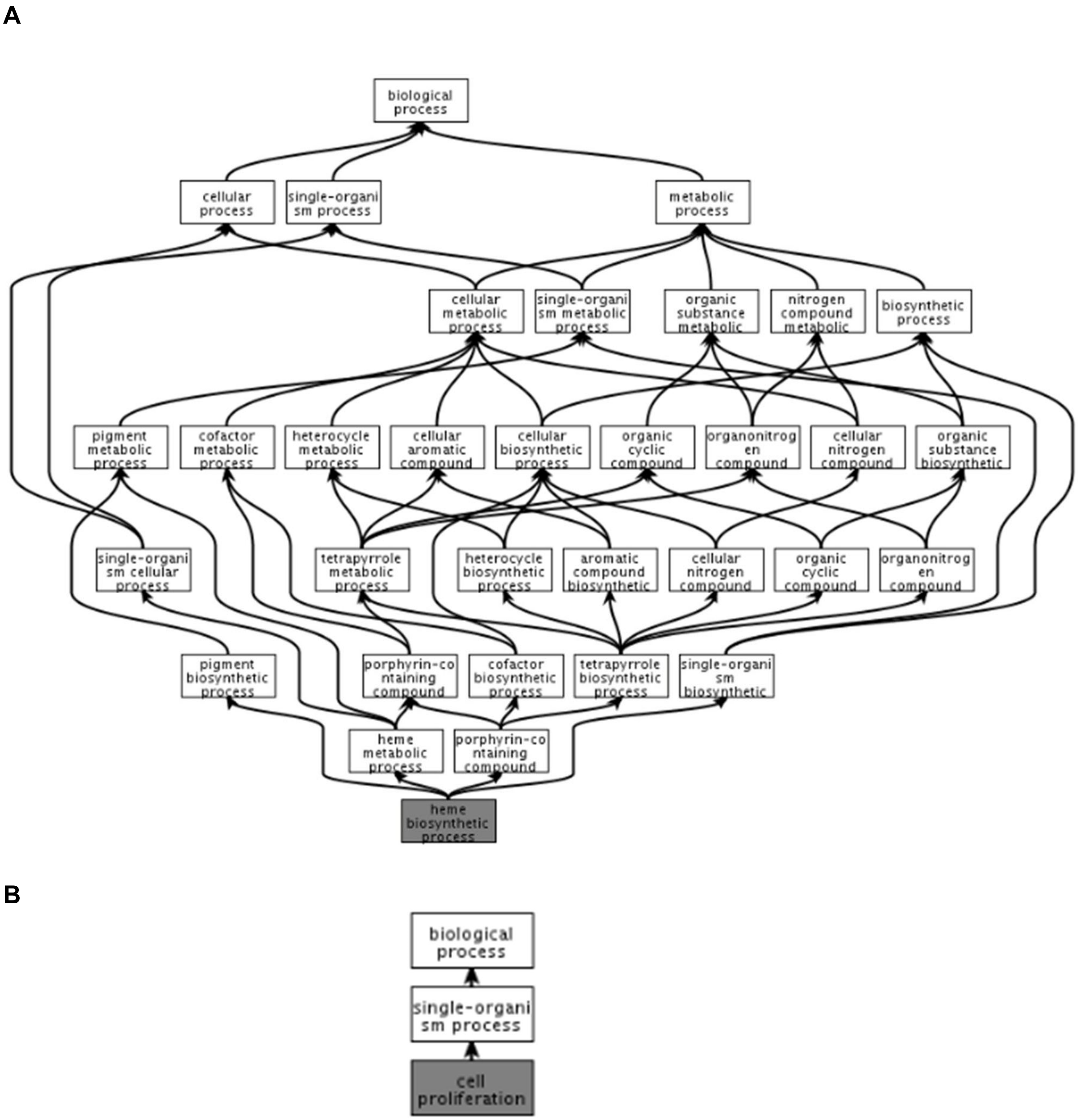

A
biological process
cellular process
single-organi sm process
metabolic process
cellular metabolic process
single-organi sm metabolic process
organic substance metabolic
nitrogen compound metabolic
biosynthetic process
pigment metabolic process
cofactor metabolic process
heterocycle metabolic process
cellular aromatic compound
cellular biosynthetic process
organic cyclic compound
organonitrog en compound
cellular nitrogen compound
organic substance biosynthetic
single-organi sm cellular process
tetrapyrrole metabolic process
heterocycle biosynthetic process
aromatic compound biosynthetic
cellular nitrogen compound
organic cyclic compound
organonitrog en compound
pigment biosynthetic process
porphyrin-co ntaining compound
cofactor biosynthetic process
tetrapyrrole biosynthetic process
single-organi sm biosynthetic
heme metabolic process
porphyrin-co ntaining compound
heme biosynthetic process
B
biological process
single-organi sm process
cell proliferation

Figure 1.15. Hierarchical mapping of Gene Ontology (GO) term haeme biosynthetic process (A) and cell proliferation (B) for TSPO

(A) Mapping illustrates that haeme biosynthetic process belongs to the biological process ontology and is a chemical reactions and pathway resulting in the formation of haem, any compound of iron complexed in a porphyrin (tetrapyrrole) ring, rom less complex precursors. **(B)** Mapping for cell proliferation, showing limited understanding of the role of cell proliferation for TSPO. Figure taken from Papadopoulos et al., 2017

Two Examples of other naturally occurring ligands which are important for protein function are the Toll-like receptors (TLR), which are a family of eleven protein recognition receptors that recognise and respond to conserved components of microbes and play a critical role in both innate and adaptive immunity (Yu, Wang and Chen, 2010) and nuclear receptors (NRs) which are hormone-sensing transcription factors that translate dietary or endocrine signals into changes in gene expression. Endogenous ligands of TLR include proteins and peptides (e.g. fibrogen and surfactant protein A), polysaccharides and proteoglycan (e.g. biglycan and heparan sulphate), nucleic acids (e.g. DNA and RNA) and phospholipids (e.g. OxPAPC) all of which are extracellular matrix degradation products and binding of these ligands to the receptor regulates the inflammation process by activation of the immune cells (Yu, Wang and Chen, 2010). This in turn leads to the production of cytokines and chemokines and inflammatory responses. The function of TLR, is dependent on the protein's ability to recognise endogenous stimulators and essential in the function of regulating non-infectious inflammation (Yu, Wang and Chen, 2010). Nuclear receptors are a superfamily which controls processes such as development, inflammation, toxicology, reproduction and metabolism (Mangelsdorf *et al.*, 1995). Endogenous ligands, when bound to NRs elicits a conformational change and this conformational change alters the cellular location of the NRs and/or their interaction with cofactors (Mangelsdorf *et al.*, 1995). This in turn translates into changes in gene expression and explains why NRs are called ligand-activated and the endogenous ligands are bile acids, phospholipids, steroid hormones, thyroid hormones, retinoids and vitamin D (Mangelsdorf *et al.*, 1995).

As previously eluded, identifying ligand-binding residues can aid the overall understanding of the role and function of a protein by using them to subsequently predict the types of ligands which they bind and for enzymes, the types of reactions that are catalysed (Dutta et al., 2017). Accurate modelling of protein-ligand interactions is an important step to understanding many biologically process. Additionally, the knowledge of residues involved in protein-ligand interactions is not just limited to understanding the function of proteins but can have applications in drug discovery (Fischer et al., 2008).

1.10 Problem statement and aims of thesis

FunFOLD3 is an integral part of the IntFOLD server and uses the top predicted tertiary structure, scored by ModFOLD, to predict the function of proteins by identifying likely associated ligands and ligand-binding site residues. One of the goals of structure prediction is to provide insights into biological functions. However, it is difficult to quantify and benchmark the utility of protein structure prediction for functional inference (Skolnick & Brylinski, 2009). Thus, the aim of this thesis is to address the need to integrate functional data into structure prediction pipelines in order to infer the functions of individual proteins, in order to answer the question and address the problem of how do we know the structure has biological relevance? In addition, the thesis has explored why should structural biologists care about *in silico* protein structure prediction and why develop these methods further, in light of experimental methods to predict protein structure and ultimately function being costly. However, it is imperative that structure-function predictions are properly evaluated. The thesis will do this by independent accuracy benchmarks and will help to improve developments in the field of protein function prediction as a whole. Once results of these competitions are released then successive improvements and development can be made to FunFOLD3 to enhance predictions of protein function prediction.

1.10.1 Objectives

The research objectives of this thesis can be summarised as follows:

- Objectively measure the performance and accuracy of FunFOLD3 for the prediction of ligands and ligand-binding site residues in the double-blinded structure and function experiment of CASP11, CASP12 and CASP13 competitions and will be explored in Chapters 3 and Chapter 4. FunFOLD3 is the method which this thesis is developing.

- Objectively measure the performance and accuracy of FunFOLDQ for the prediction of GO terms in the double-blinded CAFA3 competition and is explored in Chapter 5. Ultimately, the focus on two different methods to predict protein function will determine if there is a gold standard when it comes to protein function prediction e.g. ligands and ligand-binding site residue prediction or GO term prediction.
- Following on from the analysis in Chapters 3 and 4, determine if docking utilising AutoDock Vina can improve the ligand-binding site predictions by FunFOLD3 and results are presented in Chapter 6. The utilisation of TBM and docking, has been explored previously in literature. However, the novel aspect of the research will be inclusion of four different grid box calculations around the ligand space in order to determine an optimum cut-off when integrating docking into FunFOLD3 function prediction.
- CASP Commons will provide the framework in order to go into the unknown and provided a unique opportunity for FunFOLD3 to be used for the prediction of function and/or ligands for the SARS-CoV-2 virus and this could potentially help determine the role of proteins in the novel SARS-CoV-2 virus and therefore gain valuable insight into the role of FunFOLD3 in the prediction of novel proteins. Results from CASP Commons are presented in Chapter 7

Chapter 2: Methodology

As mentioned previously in Chapter 1, FunFOLD3 is a template-based method for protein-ligand binding site prediction (Roche and McGuffin, 2016). FunFOLD3 is the methodology which will be utilised in Chapter 3 for the analysis of protein targets from CASP11, CASP12 and CASP13 and also in Chapter 6 to aid in the functional elucidation of SARS-CoV-2. The FunFOLDQ element of FunFOLD3, which will be the prediction of GO terms, is used in Chapter 5.

Instructions for installing and running the FunFOLD3 method have been described previously (Roche and McGuffin, 2016). A downloadable version of the FunFOLD3 method is available as an executable JAR file, which can be run locally. The dependencies and system requirements are described below and the executable and example input and output data can be downloaded from http://www.reading.ac.uk/bioinf/downloads/ (Roche and McGuffin, 2016).

The system requirements are as follows:

1. A linux-based operating system such as Ubuntu
2. A recent version of Java (www.java.com/getjava/)
3. A recent version of PyMOL (www.pymol.org)
4. The TM-align program (Zhang and Skolnick, 2005) (http://zhanglab.ccmb.med.umich.edu/TM-align/).
5. wget and ImageMagick installed system wide.
6. The CIF chemical components database file (Feng *et al.*, 2004) should be downloaded from here: ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif
7. The BioLip databases(Yang, Roy and Zhang, 2012) containing ligand and receptor PDB files are also required. The databases need to be downloaded in two sections: firstly all annotations prior to 6/3/2013 can be downloaded from here for the receptor database:

http://zhanglab.ccmb.med.umich.edu/BioLiP/download/receptor_2013-03-6.tar.bz2 and from here for the ligand database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/ligand_2013-03-6.tar.bz2. The text file of the BioLip annotations can be downloaded from here: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2. To update the databases to include annotations after 2013-03-6 it is recommended to download and use this perl script which will update the databases: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/download_all_sets.pl. The BioLip text file: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2 and all the weekly update text files should be concatenated to form a large text file containing all of the annotations. Additionally, a shell script is available as downloadBioLipdata.sh, which can be downloaded from here: http://www.reading.ac.uk/bioinf/downloads/, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

8. Set system environment to English, as utilising other languages may cause problems with the FunFOLD calculations.

9. To run the program you can simply edit the shell script (FunFOLD3.sh)

10. For example, if the path of your model was "/home/dani/bin/FunFOLD3/MUProt_TS3", your list of templates was "/home/dani/bin/FunFOLD3/T0470_PARENTNew.dat" (all templates should be listed on a single line separated by a space), your FASTA sequence file was "/home/dani/bin/FunFOLD3/T0470.fasta", your output directory was "/home/dani/bin/FunFOLD3/" and your target was called T0470:

```
$JAVA_HOME/java -jar FunFOLD3.jar /home/dani/bin/FunFOLD3/MUProt_TS3 T0470
/home/dani/bin/FunFOLD3/ /home/dani/bin/FunFOLD3/T0470_PARENTNew.dat
```

```
/home/dani/bin/FunFOLD3/T0470.fasta $BIOLIP_TXT $BIOLIP_LIGAND $BIOLIP_RECEPTOR
$CIF
```

Or, using the shell script provided:

```
./FunFOLD3.sh /home/dani/bin/FunFOLD3/MUProt_TS3 T0470 /home/dani/bin/FunFOLD3/
/home/dani/bin/FunFOLD3/T0470_PARENTNew.dat /home/dani/bin/
FunFOLD3/T0470.fasta
```

11. The user requires a model generated for their target protein, this can be achieved using a homology modeling method either in-house or via a web server such as IntFOLD (see Chapter 4). Additionally, the user needs a list of structurally similar templates. Again this list of templates can be generated from the list of templates used to generate the target protein model (e.g. IntFOLD). The program utilises the templates that have the same fold and contain biologically relevant ligands in the prediction process. Furthermore, it is important to download and install the BioLip databases (Yang, Roy and Zhang, 2012) and CIF chemical components library file (Feng *et al.*, 2004). Additionally, it is important that the full paths for all input files are used, the output directory should also end with a "/" and must contain the input model, template list, and FASTA sequence file. A shell script is available called downloadBioLipdata.sh, which can be used to download and update the BioLip and CIF libraries. The shell script and the required perl script can be found on the downloads page, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

13. A number of output files are produced in the output directory (e.g. "/home/dani/bin/FunFOLD3/") and a log of the prediction process is output to screen as standard output. A description of the output files are as follows:

(a) The final ligand binding site prediction file "T0470_FN.txt" is supplied, conforming to CASP FN format. This file contains a list of predicted binding site residues, ligands, along with associated EC and GO terms.

(b) A PDB file “T0470_lig.pdb”, which contains superpositions of all templates, having the same fold and containing biologically relevant ligands, onto the 3D protein model.

(c) A reduced version of the PDB file “T0470_lig2.pdb”, which contains only the target model with all possible ligands.

(d) Another reduced version of the PDB file “T0470_lig3.pdb”, which contains only the target model with the predicted centroid ligand.

(e) A graphical representation of the protein–ligand interaction prediction “T0470_binding_site.png” is automatically generated using PyMOL.

(f) Finally, the PyMOL script “pymol.script” that was used to generate the image file is also output.

8. An example of output produced by FunFOLD3 for target T0470 can be found in the compressed directory: “T0470_Results.tar.gz” along with an example of the required input: “T0470_Input.tar.gz”. These example directories can be found on the downloads page: http://www.reading.ac.uk/bioinf/downloads/

Chapter 3: Analysis of CASP11, CASP12 and CASP13 protein targets by FunFOLD3

3.1 Introduction

The Critical Assessment of techniques for protein Structure Prediction uses blind testing of modelling methods to assess the state-of-the-art capabilities in the field (Radivojac *et al.*, 2013). Contributors are provided with amino acid sequences of unknown structures and are asked to deposit structure models (Moult *et al.*, 2016). The deposited models are then compared with newly determined experimental structures (Moult *et al.*, 2016). In the first CASP experiment in 1994, the primary concern was establishing what then-current methods could or could not deliver and (Moult *et al.*, 2016) three categories were used to define predictions and assess modelling performance, with a fourth being added in CASP2. Currently six categories are used (Moult *et al.*, 1997):

1. Models based on homologous templates (template based modelling (TBM), the most useful form of modelling)
2. Models produced without detectable homologous templates (free modelling; FM)
3. Refinement
4. Predicting the accuracy of a model
5. Predicting three dimensional contacts within structures (an area which has dramatically improved since CASP11)
6. Exploiting predicted contacts and sparse experimental structure data to build improved models (Moult *et al.*, 2016)

The gold standard for evaluating models is comparison of their coordinates with those of the corresponding experimental structure. The determination of an experimental structure is whether a specific biological question is answered. Consequently, the key question becomes not if the model is accurate as an experimental structure, but whether the structure is accurate enough to answer a biological question (Moult *et al.*, 2016).

3.1.1 History of function prediction at CASP

At the time of the inception of CASP, a majority of the progress in structure prediction over the years was knowledge based (Moult *et al.*, 2016). Meaning, the more successful methods made direct use of experimentally determined structures. CASP2 had four categories to reflect how extensively reliant structure prediction was based on other structures (Moult *et al.*, 1997):

1. Comparative or homology modelling is a prime example of utilising knowledge-based prediction, when the sequence of a target structure is clearly related to that of one or more structures it is right to presume that the structure will also be similar? (Moult *et al.*, 1997)
2. Threading or fold identification – structures deposited into the PDB can have fold(s) that have been seen before despite not demonstrating obvious sequence homology between related structures. At the stage of CASP2, this method was suggested to be of growing importance. Two main questions needed to be asked (1) how successful are the different methods at identifying fold relationships (2) when successful, what is the quality of the models produced. At CASP2 inception, techniques included advanced sequence comparison methods e.g. Hidden Markov models (Moult *et al.*, 1997)
3. *Ab initio* prediction methods that do not directly rely on knowledge on complete similar structures, encompass a wide range of techniques. At the time, the best method for doing this was to use secondary structure prediction tools, then attempt to assemble three-dimensional folds from predicted secondary structure (Moult *et al.*, 1997)
4. Docking – when the structure of two molecules is compared, is it possible to produce a detailed model of the complex between them? This is of paramount importance in drug design (Moult *et al.*, 1997)

CASP3 saw a growth in predictions with 4,000 received, four times as much as those received in CASP2, demonstrating the rapid increase in protein prediction in a short space of time (Moult *et al.*, 1997). CASP3 showed an impression of further improvement in comparative modelling areas and improvement in both comparative modelling and fold recognition categories (Moult, Hubbard and Fidelis, 1999). Several predictors produced reasonably accurate models of proteins up to 60 residues; at that stage it was encouraging. By CASP11, there was successful prediction of proteins up to 256 residues with the generation of accurate three-dimensional models for targets without templates (Moult, Hubbard and Fidelis, 1999). This was a result of much more accurate prediction of contacts between protein residues. Until CASP11, predictions in this area were disappointing with 80% false positives (Moult *et al.*, 2016).

3.1.2 Progress

CASP1 established how effective the then current methods were at predicting protein structures. CASP2 and future CASP experiments focus on the measurement of progress. There was apparent progress between CASP1 and CASP2 in particular with comparative modelling and side chain accuracy. Areas for improvement were incorrect alignments and ineffective refinement methods. However, it was hard to assess progress in threading.

In CASP3 the assessors made changes to allow structures, which did not fit into rigid frames of prediction categories to be scored favourably (Moult, Hubbard and Fidelis, 1999). Targets should not be divided firstly into comparison modelling, fold recognition and *ab initio* prediction, but all relevant categories would be carried out for all predictors for all targets, and then targets would be assigned into these three categories a posteriori. In CASP3 fold recognition methods did not achieve a high level of accuracy, due to either differences in methods or capturing different details (Zemla *et al.*, 1999). For example:

(1) Correct protein fold but poorly aligned target sequence;

(2) Large part of correct structure in a different protein fold but wrongly predicted the structure of conserved and functionally important regions in the rest of the target sequence; or

(3) Correctly predict the functionally important but otherwise small part of the target sequence and failed the rest of the sequence

CASP3 aimed to ascertain homogeneity between different predictions of fold recognition and comparative modelling. CASP3 also showed progress in *ab initio* prediction on small targets with greater success with α structures (Zemla *et al.*, 1999). Another consideration for CASP3, was the introduction of the Critical Assessment of Fully Automatic Structure Prediction methods (CAFASP) and was used in parallel to CASP3, but is independent to CASP, whilst utilising CASP target distribution and prediction collection infrastructure (Zemla *et al.*, 1999). The goal of CAFASP, was to assess the state of the art in the fully automatic methods of structure prediction, whereas CASP allows any combination of computational and human methods (Moult *et al.*, 2001).

In CASP4 there was the inclusion of large-scale benchmarking of prediction methods; EVA and LiveBench. LiveBench focused on the area of fold recognition and EVA focused on secondary structure predictions. CASP4 identified the needs for recognising correct architecture, even in cases where the topology is incorrect. As with previous years, there was an increase in participation with 163 groups taking part (including CAFASP) (Moult *et al.*, 2001). CASP4 showed an element of stability in numerical evaluation of predictions, specifically in fold recognition. The global distance test (GDT) introduced in CASP3 was found to be useful (Moult *et al.*, 2001). This ensured stability in contact prediction evaluation, which has previously been a controversial area. However, the problems of numerical evaluation had not been completely addressed. The need for additional numerical criteria would assist with automatic evaluation, especially for new fold predictions, where only some fragments of the structure were modelled successfully. For the first time, contact predictions were approaching a useful level of accuracy; however new fold models were still not accurate enough to be useful for assigning function. In fold recognition, the detection of correct folds and the quality of alignments were of particular interest (Sippl *et al.*, 2001). The superimposition of predicted models and the target domains could assist in evaluation, by indicating the number of equivalent residues between target and prediction domains and determining whether a correct fold had been recognised. Furthermore, the fraction of correctly aligned residues was used to determine the quality of the alignment (Sippl *et al.*, 2001).

As with previous CASP experiments, CASP5 saw an increase in the number of participants with 216 groups taking part. CASP4 introduced GDT and CASP5 built upon that implementing GDT_TS to establish a universal numerical evaluation for model structures, which had remained an area for improvement in all previous CASP experiments (Sippl *et al.*, 2001). Whilst GDT_TS acts as a consensus to evaluating model structures, further development is still required as GDT_TS can be modified by visual inspection (Moult *et al.*, 2003). The New Fold (formally *ab initio*) category saw continuous progress from CASP1 through to CASP4, but in CASP5 there was little evidence of further improvement (Moult *et al.*, 2003). In respect to fold recognition, CASP5 showed progress due to the emergence of metaservers which were developed as a direct result of the LiveBench experiment.

Results from the best metaservers were competitive with the best human servers and interestingly; some of the best human performances were obtained by starting from metaserver output (Moult *et al.*, 2003). A further advancement in CASP5 was the prediction of loop regions in unknown proteins where a structure for a related sequence was available.

The loops were modelled by the New Fold methods with the remaining regions being closely guided by the template (Moult *et al.*, 2003).

The sixth CASP competition, which also marked a decade of structural prediction by CASP, saw a revision in how extensively models could be based on knowledge of other structures, this consisted of three categories (i) comparative or homology modelling (ii) fold recognition– targets were assigned to this category if the target structure was found to be similar to one or more already in the PDB and did not meet the criteria for comparative modelling and (iii) New Fold Methods (Moult *et al.*, 2003). Progress over a decade of CASP experiments depended on the category prediction and the least amount of progress came in comparative modelling from high sequence identity templates (Moult *et al.*, 2005). There was steady but modest progress in difficult comparative modelling and homologous fold recognition, with respect to the extent of sequence dependent superposition between model and target, and in alignment accuracy (Moult *et al.*, 2005). Scores in this area had roughly doubled from CASP1 to CASP6. The most dramatic advances were made in the New Fold or Template-Free Modelling category (Moult *et al.*, 2005).

The seventh CASP competition saw the merging of comparative modelling and fold recognition and the introduction of a new category; high-accuracy modelling. The category consisted of template-based models, where problems of alignment and template coverage were expected to be small enough that the accuracy of resulting models should be competitive with experimental structures (Moult *et al.*, 2005). A finer measure of main chain accuracy was also implemented, GDT_HA, which had thresholds of 0.5, 1, 2 and 4Å, as opposed to 1,2, 4 and 8Å utilised in GDT_TS (Moult *et al.*, 2007). Assessment of this category looked at detailed features such as, side chain accuracy and accuracy of regions most relevant to function (Moult *et al.*, 2007). More importantly, CASP7 saw a larger emphasis on function prediction, which was included in CASP6, however lack of experimental data meant initial evaluation was complicated (Moult *et al.*, 2007) The reason for including function prediction for another CASP experiment, illustrates an increased emphasis beyond relatively simple structure accuracy to a more practical and applied area, which can have an impact (Moult *et al.*, 2007). Of the 63,717 models deposited; 1,930 were function predictions. Function prediction was difficult to determine, as there was no agreed definition prediction of function. Therefore, it was decided that EC and GO categories were to be used as definitions of function within the prediction category. Both were scored for evaluation purposes and ranged from 0 and 1, the definition is shown below (López *et al.*, 2007). Gene Ontology scores are calculated by each annotated term being compared directly with the most similar predicted term in the target predictions (López *et al.*, 2007). The pairing between the annotated term and the most similar predicted term is referred to as the computable pair (López *et al.*, 2007). Common ancestor depth was calculated for all computable pairs in a target (López *et al.*, 2007). The prediction score for each target prediction was obtained by summing the common ancestor depths of all computable pairs (López *et al.*, 2007). The final score is normalised by dividing the maximum possible score for the given target i.e. the sum of the annotated term depths (López *et al.*, 2007).

Equation 3.1. Calculation of GO Score
The equation below illustrates how a GO Score is derived. A score between 0 and 1 is obtained

$$\text{GO Score} = \frac{\text{sum of common ancestor depths of computable pairs}}{\text{sum of the annotated terms depth}}$$

Equation 3.2. Calculation of EC Score
The equation below illustrates how an EC Score is derived. A score between 0 and 1 will be obtained

$$\text{EC Score} = \frac{\text{sum of computable pair scores}}{\text{maximum possible score}}$$

Despite the inclusion of GO and EC scores, there was still a need for a specific binding site measure that could be used in the assessment and so the focus on the ligand-binding residues was established. Binding site residues were defined as all relevant residues in contact with biologically relevant ligands. Two atoms were considered to be in contact if they were within a distance of 0.5Å, plus the sum of Van der Waal's distance (López *et al.*, 2007).

One of the main focuses for CASP8 was the evaluation of template-based models and several groups were identified whom performed well in the subset of human and server targets (283 IBT_LT, 489 DBAKER, 71 Zhang, 426 Zhang-Server, 57 TASSER, 434 fams-ace2, 196 ZicoFullSTP, 46 SAM-T08-human, 299 Zico, 453 MULTICOM, 371 GeneSilico, 138 ZicoFullSTPFullData, 379 McGuffin, 282 3DShot1) (López *et al.*, 2007). At the time of publication for CASP8, 426 Zhang-Server was the only group officially registered as a server and performed the best in comparison to other groups (Cozzetto *et al.*, 2009). Following on from the introduction of function prediction category (FN) in CASP6, it became quite obvious that using EC numbers and GO terms wouldn't fall within scope and would not remain suitable in assessing predictions (Cozzetto *et al.*, 2009). The main problem was the availability of new GO terms being associated with targets, with newer GO terms being identified with targets after the end of the CASP competition (López, Ezkurdia and Tress, 2009). Instead, binding site predictions were measured using MCC, which had advantages in that it takes into consideration the imbalance between the binding site residues (positives) and non-binding residues (negatives) (López, Ezkurdia and Tress, 2009). Furthermore, the MCC score provided a statistical score for the comparison of predicted ligand binding sites to observed ligand binding site residues. Residues were assigned to one of the following; true positives, false positives, true negatives and false negatives. This provides a score of between -1 and 1, with 1 being a perfect prediction, whereas 0 was a random prediction. The main disadvantage of the MCC score was that it is a purely a statistical measure and does not consider the overall tertiary structure of the protein (López, Ezkurdia and Tress, 2009).

Matthews Correlation Co-efficient was first used in CASP8 to measure binding site prediction success (López, Ezkurdia and Tress, 2009) and is a dichotomous form of the Pearson correlation co-efficient (Powers, 2020). Due to the relationship between MCC and Pearson correlation co-efficient MCC values can be thought of in the same way (Powers, 2020). Therefore, a score of one is perfectly positive, 0.8 is strongly positive and 0.5 is moderately positive. Conversely, -1 is perfect negative, -0.8 is strongly negative, -0.5 is moderately negative and -0.2 is weakly negative (Ratnasari *et al.*, 2016). The calculation for MCC is given below (Matthews, 1975):

Equation 3.3. Matthews correlation coefficient
The equation below illustrates the calculation of the MCC, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives

$$MCC=\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

In the ninth CASP experiment, ligand-binding site prediction was explored further with more methods taking part than in the previous CASP experiment; 33 methods as opposed to 23 in the previous year (López, Ezkurdia and Tress, 2009). The prediction of ligand binding in CASP differs from typical ligand binding studies, such as docking or virtual screening. In these studies, the chemical identity of the ligand is given and the correct geometric orientation of the molecule in the receptor protein has to be determined. In the CASP experiment, the chemical identity of the ligand is unknown at the time of prediction and only the interacting residues are predicted. Thus the evaluation of ligand binding site predictions consisted of three steps; (i) identification of biologically relevant ligands in the target structure (ii) definition of binding site residues (iii) assessment of the prediction performance (Schmidt *et al.*, 2011). A major factor in function prediction is determining what a “biologically relevant” ligand is. While 73% of the target structures in CASP9 had various ligands present, most were not considered biologically relevant due to originating from solvent, crystallisation, precipitant or buffers (Schmidt *et al.*, 2011). Determining whether a ligand had biological relevance was based on type and location of the ligand, literature information and UniProt annotations (Schmidt *et al.*, 2011). As with CASP8, in CASP9 the Zhang group (FN096

Zhang and FN339 I-TASSER-FUNCTION) performed the best as demonstrated in terms of MCC (Schmidt *et al.*, 2011).

As mentioned previously, the main disadvantage of using the MCC score is that there was no consideration of the tertiary structure of the protein. In order to address this, the McGuffin group developed a new scoring metric, the BDT score (Roche, Tetchner and McGuffin, 2010). The BDT score takes into consideration the distance in 3D space that a predicted binding site residue is from the observed binding site residue. The BDT has a score ranging from 0 to 1 (1 being a prefect prediction and 0 a random prediction). The higher the score, means the closer the predicted site is from the observed site. The calculation for BDT is given below:

Equation 3.4. Binding distance test score
Where *Sij* is the *S*-score between a predicted residue *i* and an observed residue *j*, *Sij* is the Euclidean distance between the C-alpha coordinates of residues i and j and d0 is a distance threshold (values between 1 and 3 Å are recommended. The maximum Sij score, max(Sij), is then determined for each predicted residue. The final BDT score is the sum of the maximum Sij scores normalised by the greater value of the number of predicted residues (Np) and the number of observed residues (No):

$$\text{BDT}=\frac{\sum_{i=1}^{N_p}\max(S_{ij})}{\max(N_p, N_o)}$$

The BDT score has been used in CASP experiments since CASP9 (Schmidt *et al.*, 2011) and the utilisation of BDT in CASP9 by the McGuffin group was mentioned in the official function assessment publication(Schmidt *et al.*, 2011). The BDT score was applied to predictions and a few deviations in group ranking from the MCC-based prediction assessment were observed for the top groups (Figure 3.1), supporting BDT as a viable option for including in measures of binding site prediction performance.

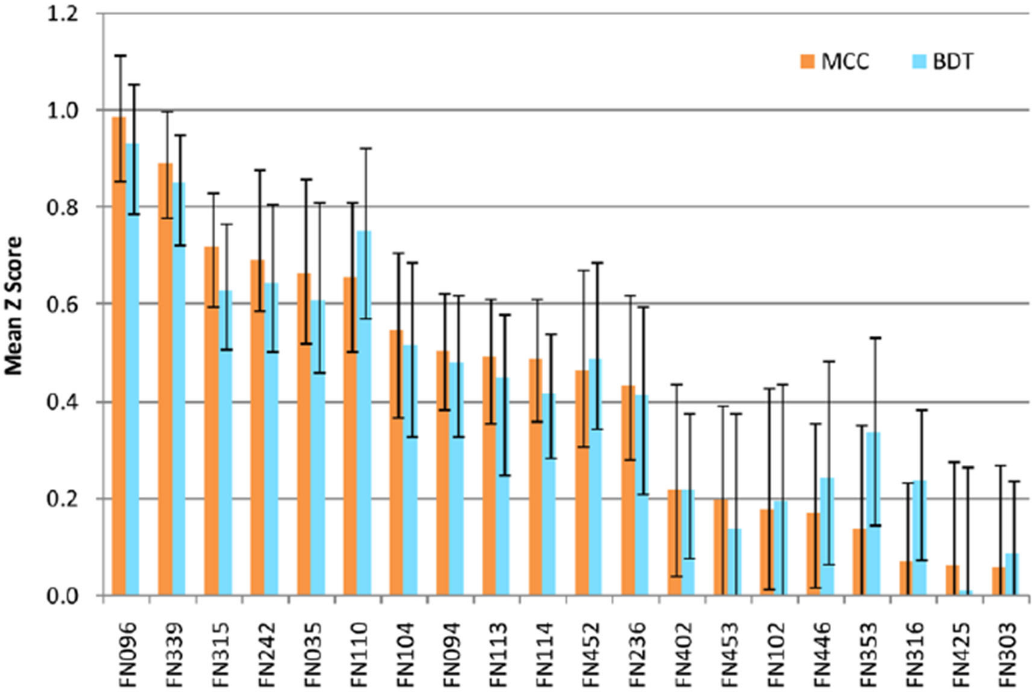


Figure 3.1. Comparison between the overall prediction performances evaluated using the Mathews Correlation Coefficient (MCC, in orange) and the Binding site Distance Test (BDT, in cyan).
Overall prediction performance is shown in mean Z Scores over all targets. Z scores are used to show the energy separation between the native fold of a protein from the ensemble of misfolded structures. Figure taken from Schmidt et al., 2011

As with previous CASP competitions, CASP10 saw a year on year increase in the number of participants taking part, with 17 groups participating in the function prediction category and 13 targets were included as part of the prediction category (Gallo Cassarino, Bordoli and Schwede, 2014). There were difficulties in analysing accuracy for different ligand types or overall structure difficulty; therefore the FN category was dropped from CASP and subsequently assessed as part of the CAMEO project using different metrics (Gallo Cassarino, Bordoli and Schwede, 2014). The prediction format in CASP, does not include a confidence score, so residues are classified in a binary way; thus either binding or not binding to any ligand. This is not to be confused with MCC score, as mentioned previously and focuses solely on the evaluation of the quality of the binding site predictions (Gallo Cassarino, Bordoli and Schwede, 2014). Since the use of BDT in CASP9, BDT was also used to assess the accuracy of predictions in CASP10 (Gallo Cassarino, Bordoli and

Schwede, 2014). More servers participated in CASP10, compared to CASP9 with six groups instead of two and performance was indistinguishable from human predictors (Gallo Cassarino, Bordoli and Schwede, 2014). In order to better understand the usefulness of the methods in practice, the performance of the methods was compared with ligands identified using DELTA-BLAST (Gallo Cassarino, Bordoli and Schwede, 2014). The average MCC was 0.339 for DELTA-BLAST, compared to the average predictor group score of 0.62 for the CASP competition and only two methods performed worse than the baseline. This demonstrated that the methods assessed in CASP10 gave advantages on the ligand binding site prediction compared to a naïve homology search approach (Gallo Cassarino, Bordoli and Schwede, 2014).

The Critical Assessment of techniques for protein Structure Prediction 11, like all other CASP experiments, aimed to obtain an in-depth and objective assessment of current abilities and inabilities in the area of protein structure prediction. However, CASP11 had two major new initiatives (Gallo Cassarino, Bordoli and Schwede, 2014):

1. Assessment of models in terms of how well they address relevant biological questions
2. Collaboration with Critical Assessment of Protein Interactions (CAPRI) to assess modelling of oligomeric relationships and of interdomain relationships

Additionally, CASP11 assessment addressed the following questions (*The Protein Prediction Center*, 2014)

1. How good are methods in identifying the most reliable predicted contacts (using RL analysis)
2. How accurate are the methods in predicting contact with the highest reliability (RL)
3. How accurate are all submitted contact predictions, including those predicted with lower reliability (FL)

A pilot scheme assessment category was introduced in CASP11; biological relevance and this aimed to assess models on the basis on how well they provide answers to biological questions. The onus being on the target providers to answer questions on why they are determining certain structures, and the ability of models to provide answers to those questions (Monastyrskyy *et al.*, 2016). Retrospective analysis of CASP11, CASP12 and CASP13 will be discussed later in this Chapter.

3.1.3 Aim

The aim of this Chapter is to objectively assess the performance of FunFOLD3 across three different CASP experiments; CASP11, CASP12 and CASP13. Critical Assessment of protein Structure Prediction provides a unique opportunity to compare known with unknown, as CASP is a double-blinded experiments, with observed structures either soon to be determined or not yet publicly available, this will act as a benchmark and will provide specific details of what needs to be improved by FunFOLD3 in the prediction of ligands and ligand-binding site residues. Results from this Chapter will guide the refinement and development of FunFOLD3.

3.2 Methods and Materials

The methodology has been described previously in Chapter 2.

FunFOLD3 is the server implementation of the refinement methods of FunFOLD and FunFOLD2. The FunFOLD standalone method takes as its input a 3D model of the protein, which is being analysed, and a list of PDB IDs, which can be obtained from the templates and used to build a 3D protein model by IntFOLD. The prototype version of FunFOLD was first developed during CASP9 and FunFOLD3 will be used in the retrospective analysis of CASP11 and CASP12.

The FunFOLD3 method was assessed using information about templates and models for each target obtained from the CASP11, CASP12 and CASP13 server predictions. The 39 targets/domains with associated PDB IDs were analysed for biologically relevant ligands in CASP11, 61 targets/domains analysed for CASP12 and 183 targets/domains were analysed for CASP13. All associated 3D models were downloaded from the CASP website (http://predictioncenter.org/download_area/).

Upon prediction of ligands(s) by FunFOLD3 using available data on the target from PDB the accuracy of the ligand(s) prediction was determined. In CASP12, the predicted GO terms were compared against available data about a target's function from UniProtKB. As this generally related to proteins with minimal annotations and no available data on PDB, further assessment will be performed using protein-ligand docking experiments (models the interaction between a ligand and the protein in order to determine the best orientation of a ligand in the ligand-binding space). Additionally, each predicted ligand-binding site residue was compared against the actual ligand-binding site residues. Incorrect predictions were deemed as either under- or overprediction.

Following the prediction of ligand-binding site residues and biologically relevant ligands, the next stage is to provide an objective measure of the ligand-binding site residues. The predicted and observed ligand-binding residues are compared against each other to output MCC and BDT scores. The MCC and BDT scores are used to rank the FunFOLD3 predictions and the MCC score can be used to compare against all of the other function prediction groups participating in CASP11, CASP12 and CASP13. Furthermore, MCC and BDT scores can assess performance across previous CASP experiments. Explanations of BDT and MCC were provided previously in this Chapter and Chapter 1.

Amino acid sequences for the prediction of tertiary structure are provided by the CASP organisers and are double-blinded, so neither the predictors nor assessors are aware of the structure at the time of prediction. Amino acid sequences for the CASP11, CASP12 and CASP13 targets, which were deemed to have biologically relevant ligands, and subsequently included in the analysis in this Chapter are presented in Table 3.1. The input for FunFOLD3 was the top-ranking 3D model from ReFOLD, a list of templates (PDB IDs) and the amino acid sequence for the target protein in question. All these inputs were processed by the FunFOLD3 algorithm. The output consisted of several files including, but not limited to, the predicted ligand-binding site residues and functional prediction with associated ligands. In order to objectively analyse the predictions all associated observed 3D models were downloaded from the CASP website.

Table 3.1. Amino acid sequences for CASP11, CASP12 and CASP13
The table below shows the amino acid sequences for which a functional prediction was made by FunFOLD3

CASP 11 target ID	Amino acid sequence
T0783	SMHPQAVAAVLPAGGCGERMGVPTPKQFCPILERPLISYTLQALERVCWIKDIVVAVTGENMEVMKSIIQKYQHKRISLVEAGVTRHRSIFNGLKALAEDQINSKLSKPEVVIIHDAVRPFVEEGVLLKVVTAAKEHGAAGAIRPLVSTVVSPSADGCLDYSLERARHRASEMPQAFLFDVIYEAYQQCSDYDLEFGTECLQLALKYCCTKAKLVEGSPDLWKVTYKRDLYAAESIIKERISQEICVVMDTEEDNKHVGHLLEEVLKSELNHVKVTSEALGHAGRHLQQIILDQCYNFVCVNVTTSDFQETQKLLSMLEESSLCILYPVVVVSVHFLDFKLVPPSQKMENLMQIREFAKEVKERNILLYGLLISYPQDDQKLQESLRQGAIIIASLIKERNSGLIGQLLIA
T0786	MDTQQLYFLNDIGKQKPESIRNRSAACPFCDRENLTDILATEGSIIWLKNKFPTLKDTFQTVLIETDNCEDHIATYTEEHMRSLIRFSIKHWLNLQKNEEFTSVILYKNHGPFSGGSLHHAHMQIIGMKYVNYLDNVEQDNFQGVIVQKNEHIELNISDRPIIGFTEFNIIIEDIGCIDELANYIQQTVRYILTDFHKGCSSYNLFFYYLNEKIICKVVPRFVVSPLYVGYKIPQVSTKIEDVKIQLAAYFTKQNDAIIHKKIE
T0798	YDYLFKVVLIGDSGVGKSNLLSRFTRNEFNLESKSTIGVEFATRSIQVDGKTIKAQIWDTAGQERYRAITSAYYRGAVGALLVYDIAKHLTYENVERWLKELRDHADSNIVIMLVGNKSDLRHLRAVPTDEARAFAEKNNLSFIETSALDSTNVEEAFKNILTEIYRIVSQKQIADCAAHDESPGNNVVDISVPPTTD
T0807	MVKKTVRFGEQAAVPAIGLGTWYMGEHAAQRQQEVAALRAGIDHGLTVIDTAEMYADGGAEEVVGQAIRGLRDRVVLVSKVYPWHAGKAAMHRACENSLRRLQTDYLDMYLLHWRGDIPLQETVEAMEKLVAEGKIRRWGVSNLDIEDMQALWRTADGEHCATNQVLYHLASRGIEYDLLPWCQQHSLPVMAYCPLAQAGRLRDGLFQHSDIINMANARGITVAQLLLAWVIRHPGVLAIPKAASIEHVVQNAAALDIVLSGEELAQLDRLYPPPQRKNRLDMV
T0813	MAQQFQTIALIGIGLIGSSIARDIREKQLAGTIVVTTRSEATLKRAGELGLGDRYTLSAAEAVEGADLVVVSVPVGASGAVAAEIAAHLKPGAIVTDVGSTKGSVIAQMAPHLPKDVHFVPGHPIAGTEHSGPDAGFAGLFRGRWCILTPPAGTDEEAVARLRLFWETLGSMVDEMDPKHHDKVLAIVSHLPHIIAYNIVGTADDLETVTESEVIKYSASGFRDFTRLAASDPTMWRDVCLHNKDAILEMLARFSEDLASLQRAIRWGDGDKLFDLFTRTRAIRRSIVQAGQDTAMPDFGRHAMDQK
T0819	MSAFSRFTPLIQSLPASVPFVGPEALERQHGRKIAARIGANESGFGPAPSVLLAIRQAAGDTWKYADPENHDLKQALARHLGTSPANIAIGEGIDGLLGQIVRLVVEAGAPVVTSLGGYPTFNYHVAGHGGRLVTVPYADDREDLEGLLAAVGRENAPLVYLANPDNPMGSWWPAERVVAFAQALPETTLLVLDEAYCETAPRDALPPIESLIDKPNVIRARTFSKAYGLAGARIGYTLSTPGTAQAFDKIRNHFGMSRIGVAAAIAALADQDYLKEVTLKIANSRQRIGRIAADSGLAPLPSATNFVAVDCGKDASYARAIVDRLMSDHGIFIRMPGVAPLNRCIRISTAPDAEMDLLAAALPEVIRSLAAT
T0845	MNKLYTTLLIACLAAGFTACNDDDCEDLHLGNLAHYPNVLKGTFPTESQVLELGETLEITPELLNPEGATYSWLVNGKEYSTEPTFSYKIDNPCRADLSCIIKNKYGKVEMSTSFSSNHNFSKGFFYVADGTFNFYDTEKKTAYQDCYASLNAGKTLGIGNYDSANIIHSNGKFYLLVGTSTSNRDHFYIVDAKTLYYENSAVVGANLSGLTILNEQYGLVTGDGIRRIDLKSLNNVRIKNERLLCFYNSIIYNGKVLSNDTYKDESKVKYYDVNELIAAKEGEAPAVTELDIIQKQKINFVLAKDGNVYTLESADNGCNIVKIKNDFTLEKVFANFQPAKGPYHSSPTIGMVASETENIIYLVSTDGAIYKYILGDSDSLKAPFIAAESGVSITAPLQLNQQSGELYVTYTEERKDESKIVVYSKDGKVLHTVDCGESVPSQILFNN

T0854	MAVSSNGGAIKAVIYDCDGVMFDSFEANLAFYQRIMEMMGRPRLSRDNEEQMRILHTYANREVLAHFFPSPGDWEEAVRCAGAIDYRELVPLMIMEEGFREALDTLKGRVGLGVCTNRSTSMDMVLRLFSLDSYFSIVMTASRVTNPKPHPEPLLKVLEHFGIGPREALFVGDSEVDRLSAEAAGVPFVAYKAPLPAAYRMEHHREIIDLLG
T0849	MNEPIILRYFPVLGRAQALRHALADAELAFRDLRIPLEQWSQHKDSDAGGPYGSLPTLRWHGVEVAETIAIASFLARSLGHYEGRDNGEIARLEAVVSLCYTEVSLQIAQLLWLDLFNPGVDLAAAVPLQFGRLVARLTRLEAHTPEAGWFGGERPVMADYFAAEAIEALRYLLGREHDDALRTRLPHLCALARRMAQRPALAQAWSTRPQTFTAHPDEAAMLERLRALPLAATIGASME

CASP 12 target ID	Amino acid sequence
T0868	MGASSGSNISASNGSSSPTTIVASNPVDLNAFDRLNVVDPAVGKFRPGEAGAAAELENYLGGTLQRAPQGSSVDFVFSSGPNNGKTVDFMLTPDTVAQAAKINQFFDKNLNNFMNTLSDHAAAADFVPLASRFLSEANKTLLVKAIGNLPQKLQAKIILIK
T0872	VTVDDLVEGIAFSITHDSENPNIVYLKSLMPSSYQVCWQHPQGRSQEREVTLQMPFEGKYEVTFGVQTRGGIVYGNPATFTIDSFCADFVN
T0892	SPSINVALKAAFPSPPYLVELLETAASDNTTIYYSLLDRIAKGHFAEATTDKALYEKFLEVLRDDGHMDPEALSAFKLALSLRTATPRVEAHYQYYTATVEPSLSGTQEGCDQWFLIDGEQYCSPTLDTSHGKVKGEDQLRTLPFDRKFGVGSRDVILYADITSKSFAPFHEVAMDLAKKGKASYRVRYRRSP
T0899	MKQIIALCIYTSALMLVTGCSPESPGLMVQQDPIPETPVEIPQYEMPAQQEFKWITEDGGQSQLDFNPQVDILFVTDNSESMKSAQENLVRNLDRFTNGINKNAMIDYQIGVISTWDSSERFSATKKDKYGIGELRHIKDGKSQNYNKRFVTKKEKHLLASTLDLGVAPYAQGGPEDEEFFAPLTAALEKSGRGGVNEGFFREDAQLVVVFLTDADEFKQSRITAEQMARTLLDFKKGKANKLAVYGALVKASDSDQYKDWALRIHPKYNEQCFDMTQKTPKNNGTCTGFGPEKLEELIVRANEDKGAPDAIKSKYIVGIVNKNFGEDLARIGSDITKKTLAKEIFLTQRPRATADGSLQIRVRYGTPEQLNAGRGQVIPNKANGGWTYDPENNSVLLPGDIEYKYQDKARFAVDLIPLTLAQ
T0901	MKKDFQLLVVAAASSLLMACAQNVSFDLPETQDNFGQSITYNNKVDILWIVDNSTSMLKHQQRLSEQVPDLVSKLNTLKMDYHMAVVTTSMGGTSPDGGKFIGSPKYVTSKTPDLVNSLKNRMIVGEAGSNLERGLESMENALSANYLANEGKGFFRNDALLVVIALSDEDDFSKSSSSAGITYYTNLLDGIKEPWVDGSRSWVFNFIGVLSLTSQCKTFNDFASAGLSYMGLADTSGGVKESICSTNLSSAVGNIRSRIYQILTDFKLSKVPLEESITVSINGVSIPRDTTNGWDYLAASNVIRFYGTAVPAADASIKVDFKPKDAN
T0905	MKKTVASKALMMASAVALVAGCSKGTGSYSLLNDAEDYKQQAVFIPKQIDILWVIDNSGSMKTSQDNLAANFQSFISRFQQYNYDFHMAVTTTEAWEKQFNSASEKARIKDGAVLQTNPKIETHSGVFIMDKNTANLGDVFSTNAKQGTLGNGDERAFSSFKEALLEPQNAGFRRSEAFLAVIIVSDEEDFSSSSAAFNESYNNANLHTVQSYVDFLDGYVGSRNYSVSTITVPDDACKTSLSTDGFARKISTRLPELATLTAGVKGSLCSNFGSTLELISDSIIQLSSVFKLNREPQEDTIVITVNGVSVPNDAVNGWTYDASNLTITFHGSSVPAADANITIDFYPKSIKL
T0907	MGHHHHHHSSGVDLGTENLYFQSVNSITAQVIPQSQIVMSGDTYKANIVLSSVDTTQRPDVFVNGKLLSPENMGLFTATAGAPGTYPVKGYIEMMGNDGVKIRRDFESEYFVTEPMASVAPTMMNVLYAGIDNPINIAVPGVAQQNVSATINNGTLTRRGNLWIARPTKVGSEAIISVTAQSGGRTIQMAKTTLRVRALPDPLPYIEYKDVQGNTKRFKGGRLGKREILAAGGIKAALDDDLLEVNYTVVKFQLVFYDSMGNSIPEVSDGASFSERQKRQIQNLGKGKRFYVTEVIARGPDGIERKIPAIEVIVN
T0909	LLPGQSPDEAFARNSVVFLVPGAEYNWKNVVIRKPVWIYGERCHGEDFRPRAIIHIMGDLDNPMDVRIQDLTFIGGDSPDRLVPFSAVLTNQMALWCIDPRITIRGCSFYNFGGAAIYLERSERDTGFRFGRGQVMITDCRFRGCRIGIANGGSVEYGLASQNNFSDCQICFNVVGGNWTRSGNVASNCRCMYLHTQGMWYEGAAGNFNPAHGSFTSNTLNHCDYGGNLWPTEFQLPDRVINLAGFYFDNAAARLPNFSGNSQWYGDMKLINFLPDSTFVINGGALYGGPGDTGVIAVATALAAKVFVIGCQGNAGQQIVNVPAANIIPEVGTRKDDATQ

T0911	MVSGFAMPKIWRKLAMDIPVNAAKPGRRRYLTLVMIFITVVICYVDRANLAVASAHIQEEFGITKAEMGYVFSAFAWLYTLCQIPGGWFLDRVGSRVTYFIAIFGWSVATLFQGFATGLMSLIGLRAITGIFEAPAFPTNNRMVTSWFPEHERASAVGFYTSGQFVGLAFLTPLLIWIQEMLSWHWVFIVTGGIGIIWSLIWFKVYQPPRLTKGISKAELDYIRDGGGLVDGDAPVKKEARQPLTAKDWKLVFHRKLIGVYLGQFAVASTLWFFLTWFPNYLTQEKGITALKAGFMTTVPFLAAFVGVLLSGWVADLLVRKGFSLGFARKTPIICGLLISTCIMGANYTNDPMMIMCLMALAFFGNGFASITWSLVSSLAPMRLIGLTGGVFNFAGGLGGITVPLVVGYLAQGYGFAPALVYISAVALIGALSYILLVGDVKRVG
T0912	MGSSHHHHHHSSGPQQGLRHLLSAGEIWISPQGNDLNDGTRPSPKATLTSALRQAREWRRTDDERVRGGITICMEGGTYALYEPVFIRPEDSGTEDSPTVIRPVADEKVVLSGGIRIGGWKKQGKLWVADVPMFNGRPLDFRQLWVNGKKAVRARDVEDFEKMNRICSVDEKNEILYVPAVAIRRLVDGKGALKAKYAEMVLHQMWCVANLRIRSVELAGDSAAIRFHQPESRIQFEHPWPRPMVTTDGHNSAFYLTNARELLDVAGEWYHDIDARKVYYYPREGEKLQDAGTEVIVPAIETLIQVKGTFDRPVSHIRFEKITFSHTTWMRPSEKGHVPLQAGMYLTDGYRIDPKMERDYLNHPLDNQGWLGRPAAAVSVAAANQIDFERCRFDHLGSTGLDYEEAVQGGVVRGCLFRDIAGNGLVVGSFSPAAHETHLPYDPTDLREVCAHQQISNCYFTEVGNEDWGCLAILAGYVKDINIEHNEICEVPYSGISLGWGWTQTVNCMRNNRVHANLIHHYAKHMYDVAGVYTLGSQPKSYVTENCVHSIYKPGYVHDPNHWFYLYTDEGSSFITVRDNWTEGEKYLQNANGPGNVWENNGPQVDTVIRERAGLEAEYRDLKK
T0916	SRNMKEKLEDMESVLKDLTEEKRKDVLNSLAKCLGKEDIRQDLEQRVSEVLISRELHMEDSDKPLLSSLFNAAGVLVEARAKAILDFLDALLELSEEQQFVAEALEKGTLPLLKDQVKSVMEQNWDELASSPPDMDYDPEARILCALYVVVSILLELAEGPTSVSS
T0919	SRSVTVVGHSSFCTSDVVMSSTELNRLLGTDIYNFARGGASDVEVAMMSQEAITRQYAPVGGSSIPASGSVALTPTEVGIFWNGATGKCIIFGGIDGTFSTTLVNAGTGETQLVFTRDSAGSAVSSVSTTATFAMRPYTRFNTNTIPAGRRRKHSSLHRDDIYIVWGGRNSTDYTRYVSELHTMVANMHTQRRRFVICPEFPYDTETTGTTGATNLAALNNNLKADFPDNYCQISGGVDLLQNFKSKYNPAYAGDVTDIANGITPRSLREDNLHPSETTLQPNGLYIGAKVNADFIAQFIKSKGWG
CASP 13 target ID	**Amino acid sequence**
T0949	MAAKKGMTTVLVSAVICAGVIIGALQWEKAVALPNPSGQVINGVHHYTIDEFNYYYKPDRMTWHVGEKVELTIDNRSQSAPPIAHQFSIGRTLVSRDNGFPKSQAIAVGWKDNFFDGVPITSGGQTGPVPAFSVSLNGGQKYTFSFVVPNKPGKWEYGCFLQTGQHFMNGMHGILDILPAQGS
T0953s2	MAVQGPWVGSSYVAETGQNWASLAANELRVTERPFWISSFIGRSKEEIWEWTGENHSFNKDWLIGELRNRGGTPVVINIRAHQVSYTPGAPLFEFPGDLPNAYITLNIYADIYGRGGTGGVAYLGGNPGGDCIHNWIGNRLRINNQGWICGGGGGGGGFRVGHTEAGGGGGRPLGAGGVSSLNLNGDNATLGAPGRGYQLGNDYAGNGGDVGNPGSASSAEMGGGAAGRAVVGTSPQWINVGNIAGSL
T0954	SPSSQGQHKHKYHFQKTFTVSQAGNCRIMAYCDALSCLVISQPSPQASFLPGFGVKMLSTANMKSSQYIPMHGKQIRGLAFSSYLRGLLLSASLDNTIKLTSLETNTVVQTYNAGRPVWSCCWCLDEANYIYAGLANGSILVYDVRNTSSHVQELVAQKARCPLVSLSYMPRAASAAFPYGGVLAGTLEDASFWEQKMDFSHWPHVLPLEPGGCIDFQTENSSRHCLVTYRPDKNHTTIRSVLMEMSYRLDDTGNPICSCQPVHTFFGGPTCKLLTKNAIFQSPENDGNILVCTGDEAANSALLWDAASGSLLQDLQTDQPVLDICPFEVNRNSYLATLTEKMVHIYKWE
T0955	SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC
T0957s2	SNAMINVNSTAKDIEGLESYLANGYVEANSFNDPEDDALECLSNLLVKDSRGGLSFCKKILNSNNIDGVFIKGSALNFLLLSEQWSYAFEYLTSNADNITLAELEKALFYFYCAKNETDPYPVPEGLFKKLMKRYEELKNDPDAKFYHLHETYDDFSKAYPLNN
T0958	MNKKSKQQEKLYNFIIAKSFQQPVGSTFTYGELRKKYNVVCSTNDQREVGRRFAYWIKYTPGLPFKIVGTKNGSLLYQKIGINPCNNSTPSKGGDC
T0961	MKNFYQDGPQLSNTFRSDEALQKILKSLLPADAQKVALPHLEHLGERAVTDMLTWAQEAESQPPVHVPFDPWGRRIDDIKTSHGWKALEKVAAEEGIVATAYDRRFGAASRVYQMALLYLYSPSSAIFSCPLAMTDGAARALELYADADLKARVLPHLLSRDPKTFWTAGQWMTERTGGSDVSGTSTDAHPFTGTSEFGATHSLHGTKWFTSATTSQMALTLARPDGAAPGSRGLSLFFLELRNDKGELNHIQ

	IHRLKDKLGTKALPTAELSLQGTPARMIGGVGEGVKRIASVLNITRIYNSICAVGHIRRALDLAQDYSGKRQAFGKLLKDHPLHKSTLDSLEADFRKCIAFSFFVANLLGQEEVGEASASEKILLRVLTPIL KLYTAKKSIHISSEVVEMFGGAGYVEDTGIPRLLRDAQVFSIWEGTTNVLSLDMLRAFEKDQAGQILEQFLVLNEAGSEELVRLQKLLTLSGEQKEQHAREIAFLIGNAVARIAMKKYSL
T0965	MGSSHHHHHHSSGLVPRGSHMEGKKILVTGGTGQVARPVAEALAERNEVWCLGRFGTPGVEKELNDRGITTFHWDMDDPGAAAYEGLPDDFTHVLHSAVRRGEDGDVNAAVEVNSVACGRLMT HCRGAEAFLFVSTGALYKRQTLDHAYTEDDPVDGVADWLPAYPVGKIAAEGAVRAFAQVLNLPTTIARLNIAYGPGGYGGVPMLYFKRMLAGEPIPVPKEGQNWCSLLHTDDLVAHVPRLWEAAATP ATLVNWGGDEAVGITDCVRYLEELTGVRARLVPSEVTRETYRFDPTRRREITGPCRVPWREGVRRTLQALHPEHLPSESRHSAV
T0970	KRSGFLTLGYRGSYVARIMVCGRIALAKEVFGDTLNESRDEKYTSRFYLKFTYLEQAFDRLSEAGFHMVACNSSGTAAFYRDDKIWSSYTEYIFFRP
T0972	MVVDNTQKTSNAIFSTTTKVKEKNTSADEFQATLNEVKNKEEKEDKKTNSSKFTNEDIDLGAVREDFRSYAWQKMREDQYKKNEETLLNKLFTTIDAGNATNNTKA
T0973	PQAADIVIADAQATPVNHTFVPIGPDPKDATIYWWEDQSQASPAGYWRLSMQLVRPAPAKAGQNTNQRMIRVRVSTFEPILEVAVTATYSGIAPSPTVSYVPKAFTEFVLPERATLDNRKDIRKMHALA LTTSEAIAMIESLQFVY
T0974s1	MSYDYSSLLGKITEKCGTQYNFAIAMGLSERTVSLKLNDKVTWKDDEILKAVHVLELNPQDIPKYFFNAKVH
T0975	LEDAQESKALVNMPGPSSESLGKDDKPISLQNWKRGLDILSPMERFHLKYLYVTDLATQNWCELQTAYGKELPGFLAPEKAAVLDTGASIHLARELELHDLVTVPVTTKEDAWAIKFLNILLLIPTLQSE GHIREFPVFGEGEGVLLVGVIDELHYTAKGELELAELKTRRRPMLPLEAQKKKDCFQVSLYKYIFDAMVQGKVTPASLIHHTKLCLEKPLGPSVLRHAQQGGFSVKSLGDLMELVFLSLTLSDLPVIDIL KIEYIHQETATVLGTEIVAFKEKEVRAKVQHYMAYWMGHREPQGVDVEEAWKCRTCTYADICEWRKGSGVLSSTLAPQVKKAK
T0980s1	SLKPFTYPFPETRFLHAGPNVYKFKIRYGKSIRGEEIENKEVITQELEDSVRVVLGNLDNLQPFATEHFIVFPYKSKWERVSHLKFKHGEIILIPYPFVFTLYVEMKWFHE
T0980s2	VNNMVTGYISIDAMKKFLGELHDFIPGTSGYLAYHVQNEINMSAIKNKLKRK
T0983	MFGPEHAEVYEAAYRGRGKSWHDEAADVADRIRAARPDAARLLDVGCGTGAHLETFATRFPHVEGLELAPAMLALARHRLPGVRLHAGDMRTFDLGVTFDAVTCLFTAVNFLGTVAEMRAAVAAM SAHLAPGGVLVLEPWWFPERFIDGYVGGDLVREEGRTVARVSRSTRQGRVTRMEERWLVGDAAGIREFSQVGLLTMFTREEYDAAFAAAGCESAYVEGWLTGRGLFVATRTGGHATPTMV
T0985	MAHHHHHHVGTGSNDDDDKSPMKLKQDVISIYQKISLFESGQLNITKLASGAYYLDDELTLITDPVNSGARFPYAVNGMTIWAYASGYISINHSSYYILPPNLEGKEPFLDFFGIEQDGNNTYPVSLLGV SERNDEIENKRYTVFSKNIAYYITVTKNFLYAVTVYISKDFKIYFNTVAHNLTGETKQITLSSFFNMLFKYDSGESIETKWFKKVSYENNMFIYDAPEDIDRHTRIENYGVVKRHLHTKPKNIQNTTSRIDYV GKRYRSVRNALSIRSLKFEKAPLVTNFTDTAINADLINYEVKAYDTIISSYRIETCHDKDTLNKMMASDLTDKEIKKVYEGLSNTQSYDFDNFGISFKGVNDNRVDDKVLNQFLKLVNYQIHFSSLSSNSG TVFLGVRDVMQQLESSLIWDRKNVRSKILEVLSFIDPSGLPPRQYALPPKEGNPRMDLRPFIDQGLWIISTLHTYLAYTEDYDILNEVCGYYERIEPNSAKKSKVENSVLEHLIRVTNYLVSNIDPSTYGL KALYGDWNDALDGLGLIEGSSGYGNGVSVMATLQLYENLERMIEILKLVDPQNEHINTYEVVRHNLSLGINKYAVVIKQDEKRVLHGWGHDRSYFVGSFNDPDGHSRNSLTSNAFYIISDMIKNTPEM KPHLLHAFHNLDSKYGLKTFDPAMQDFHGFGRIINLPPGTAENAATYVHATLFGVLALYMLGEGDFANEQVLKVLPITKKEMSTSPFIMPNSYVHNEELNMDGESMSDWYTGSANTLLKTLIRGLFGLE VKFDHLRLRPSKAFFSKEATLMVSIGNKLTRIVYKNNNNGNRTFKLNGKVIEAKLDTLSGLLYIDINKSILEHQNVIHIQD

T0986s2	MKELFEVIFEGVNTSRLFFLLKEIESKSDRIFDFNFSEDFFSSNVNVFSELLIDSFLGFNGDLYFGVSMEGFSVKDGLKLPVVLLRVLKYEGGVDVGLCFYMNDFNSAGKVMLEFQKYMNGISADFGFENFYGGLEPASDQETRFFTNNRLGPLL
T0992	HGEDKPGPHGGHIQMPGAFHTEITVDKDQSVHVYLLDMNFANPTIKDSSVAVTAKNKKSEIKYTCSVMGNDHYHCIPNGKVPAKTNLIVQATREKAVGNEAVYKLPLPAFKESKKESKKEDHSHHH
T0993s1	MEQSVANLVDMRDVSFTRGNRCIFDNISLTVPRGKITAIMGPSGIGKTTLLRLIGGQIAPDHGEILFDGENIPAMSRSRLYTVRKRMSMLFQSGALFTDMNVFDNVAYPLREHTQLPAPLLHSTVMMKLEAVGLRGAAKLMPSELSGGMARRAALARAIALEPDLIMFDEPFVGQDPITMGVLVKLISELNSALGVTCVVVSHDVPEVLSIADHAWILADKKIVAHGSAQALQANPDPRVRQFLDGIADGPVPFRYPAGDYHADLLPGS
T0994	MLSSFLMLSIISSLLTICVIFLVRMLYIKYTQNIMSHKIWLLVLVSTLIPLIPFYKISNFTFSKDMMNRNVSDTTSSVSHMLDGQQSSVTKDLAINVNQFETSNITYMILLIWVFGSLLCLFYMIKAFRQIDVIKSSSLESSYLNERLKVCQSKMQFYKKHITISYSSNIDNPMVFGLVKSQIVLPTVVVETMNDKEIEYIILHELSHVKSHDLIFNQLYVVFKMIFWFNPALYISKTMMDNDCEKVCDRNVLKILNRHEHIRYGESILKCSILKSQHINNVAAQYLLGFNSNIKERVKYIALYDSMPKPNRNKRIVAYIVCSISLLIQAPLLSAHVQQDKYETNVSYKKLNQLAPYFKGFDGSFVLYNEREQAYSIYNEPESKQRYSPNSTYKIYLALMAFDQNLLSLNHTEQQWDKHQYPFKEWNQDQNLNSSMKYSVNWYYENLNKHLRQDEVKSYLDLIEYGNEEISGNENYWNESSLKISAIEQVNLLKNMKQHNMHFDNKAIEKVENSMTLKQKDTYKYVGKTGTGIVNHKEANGWFVGYVETKDNTYYFATHLKGEDNANGEKAQQISERILKEMELI
T0995	MTSIYPKFRAAAVQAAPIYLNLEASVEKSCELIDEAASNGAKLVAFPEAFLPGYPWFAFIGHPEYTRKFYHELYKNAVEIPSLAIQKISEAAKRNETYVCISCSEKDGGSLYLAQLWFNPNGDLIGKHRKMRASVAERLIWGDGSGSMMPVFQTEIGNLGGLMCWEHQVPLDLMAMNAQNEQVHVASWPGYFDDEISSRYYAIATQTFVLMTSSIYTEEMKEMICLTQEQRDYFETFKSGHTCIYGPDGEPISDMVPAETEGIAYAEIDVERVIDYKYYIDPAGHYSNQSLSMNFNQQPTPVVKHLNHQKNEVFTYEDIQYQHGILEEKV
T0997	AGQDYSSAEVLPDDTEMEQTIPETNTADKTTAEETEPAALEDTTTLMESAAVLKNYDHLDPKRMINSKALAEAVLYFDKNQSRIKNKKYMSLIDFGKRSTQARFFIINMSTGEVTAIHTAHGKGSDANHGYAEKFSNNSGSNASSLGYYLAAETYYGKHGLSLKLDGLSSTNSKARARAVVIHGASYVKESSVIQGRSWGCPAVANHLRDKVIGMLKGGSLIYAFAK
T1001	SISTRIGEYRSAQSKEDLIQKYLNQLPGSLCVFFKFLPSVRSFVATHASGIPGSDIQGVGVQLESNDMKELSSQMAIGLLPPRFTEMLVEAFHFSPPKALPLYAHNALEGVFVYSGQLPAEEVARMNEEFTLLSLCYSHF
T1003	LQDGKSKIVQKAAPEVQEDVKAFKTGNYVFSYDQFFRDKIMEKKQDHTYRVFKTVNRWADAYPFAQHFSEASVASKDVSVWCSNDYLGMSRHPQVLQATQETLQRHGVGAGGTRNISGTSKFHVELEQELAELHQKDSALLFSSCFVANDSTLFTLAKILPGCEIYSDAGNHASMIQGIRNSGAAKFVFRHNDPDHLKKLLEKSNPKIPKIVAFETVHSMDGAICPLEELCDVSHQYGALTFVDEVHAVGLYGSRGAGIGERDGIMHKIDIISGTLGKAFGCVGGYIASTRDLVDMVRSYAAGFIFTTSLPPMVLSGALESVRLLKGEEGQALRRAHQRNVKHMRQLLMDRGLPVIPCPSHIIPIRVGNAALNSKLCDLLLSKHGIYVQAINYPTVPRGEELLRLAPSPHHSPQMMEDFVEKLLLAWTAVGLPLQDVSVAACNFCRRPVHFELMSEWERSYFGNMGPQYVTTYA
T1008	TDELLERLRQLFEELHERGTEIVVEVHINGERDEIRVRNISKEELKKLLERIREKIEREGSSEVEVNVHSGGQTWTFNEK
T1009	TYFAPNSTGLRIQHGFETILIQPFGYDGFRVRAWPFRPPSGNEISFIYDPPIEGYEDTAHGMSYDTATTGTEPRTLRNGNIILRTTGWGGTTAGYRLSFYRVNDDGSETLLTNEYAPLKSLNPRYYYWPGPGAEFSAEFSFSATPDEQIYGTGTQQDHMINKKGSVIDMVNFNSYIPTPVFMSNKGYAFIWNMPAEGRMEFGTLRTRFTAASTTLVDYVIVAAQPGDYDTLQQRISALTGRAPAPPDFSLGYIQSKLRYENQTEVELLAQNFHDRNIPVSMIVIDYQSWAHQGDWALDPRLWPNVAQMSARVKNLTGAEMMASLWPSVADDSVNYAALQANGLLSATRDGPGTTDSWNGSYIRNYDSTNPSARKFLWSMLKKNYYDKGIKNFWIDQADGGALGEAYENNGQSTYIESIPFTLPNVNYAAGTQLSVGKLYPWAHQQAIEEGFRNATDTKEGSACDHVSLSRSGYIGSQRFCSMIWSGDTTSVWDTLAVQVASGLSAAAT

	GWGWWTVDAGGFEVDSTVWWSGNIDTPEYRELYVRWLAWTTFLPFMRTHGSRTCYFQDAYTCANEPWSYGASNTPIIVSYIHLRYQLGAYLKSIFNQFHLTGRSIMRPLYMDFEKTDPKISQLVSS NSNYTTQQYMFGPRLLVSPVTLPNVTEWPVYLPQTGQNNTKPWTYWWTNETYAGGQVVKVPAPLQHIPVFHLGSREELLSGNVF
T1011	DYKDDDDGAPKETRGYGGDAPFCTRLNHSYTGMWAPERSAEARGNLTRPPGSGEDCGSVSVAFPITMLLTGFVGNALAMLLVSRSYRRRESKRKKSFLLCIGWLALTDLVGQLLTTPVVIVVYLSK QRWEHIDPSGRLCTFFGLTMTVFGLSSLFIASAMAVERALAIRAPHWYASHMKTRATRAVLLGVWLAVLAFALLPVLGVGQYTVQWPGTWCFISTGRGGNGTSSSHNWGNLFFASAFAFLGLLALTV TFSCNLATIKALVSRGSNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDATVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTN SLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYGSWGRITTETAIQLMAIMCVLSVCWSPLLIMMLKMIFNQTSVEHCKTHTEKQKECNFFLIAVRLASLNQILDPWVYLLLRKILG RPLEVLFQGPHHHHHHHHHH
T1012	MTEYKPTVRLATRDDVPRAVRTLAAAFADYPATRHTVDPDRHIERVTELQELFLTRVGLDIGKVWVADDGAAVAVWTTPESVEAGAVFAEIGPRMAELSGSRLAAQQQMEGLLAPHRPKEPAWFLA TVGVSPDHQGKGLGSAVVLPGVEAAERAGVPAFLETSAPRNLPFYERLGFTVTADVEVPEGPRTWCMTRKPGA
T1013	DMADEPLNGSHTWLSIPFDLNGSVVSTNTSNQTEPYYDLTSNAVLTFIYFVVCIIGLCGNTLVIYVILRYAKMKTITNIYILNLAIANELFMLGLPFLAMQVALEHWPFGKAICRVVMTVDGINQFTSIFCLTV MSIDRYLAVVHPIKSAKWRRPRTAKMITMAVWGVSLLVILPIMIYAGLRSNQWGRSSCTINWPGESGAWYTGFIIYTFILGFLVPLTIICLCYLFIIIKVKSASTDYWQNWTFGGGIVNAVNGSGGNYSVN WSNTGNFVVGKGWTTGSPFFRTINYNAGVWAPNGNGYLTLYGWTRSPLIEYYVVDSWGTYRPTGTYKGTVKSDGGTYDIYTTTRYNAPSIDGDDTTFTQYWSVRQSKRPTGSNATITFTNHVNAWK SHGMNLGSNWAYQVMATEGYQSSGSSNVTVWSSKRKKSEKKVTRMVSIVVAVFIFCWLPFYIFNVSSVSMAISPTPALKGMFDFVVVLTYANSCANPILYAFLDDNFKKSFQNVLCLVKVSGTDDGE RSDSKQDKSRLNETTETQRT
T1014	SLAPVDIEGLLRQVAELMSPRAHEKGIEIAWAVSSPLPTILADEGRLRQILLNFAGNAVKFTEAGGVLLTASAIDGGRVRFSVADTGPGVAPDARARIFEAFVQTDVTHATQLGGAGLGLAIVSRLSAAM GGAVGVGGELGQGAEFWFEAPFATAAAPLRAAPLEGRNVAIASPNAIVRAATARQIEAAGGRAYAAVDIASALAGAPADAVLLIDAALSGPRGALKPPAGRRSVVLLTPEQRDRIDRLKAAGFSGYLIK PLRAASLVAQVLQAVTA
T1016	MRLWLIRHGETQANIDGLYSGHAPTPLTARGIEQAQNLHTLLHGVSFDLVLCSELERAQHTARLVLSDRQLPVQIIPELNEMFFGDWEMRHHRDLMQEDAENYSAWCNDWQHAIPTNGEGFQAFSQ RVERFIARLSEFQHYQNILVVSHQGVLSLLIARLIGMPAEAMWHFRVDQGCWSAIDINQKFATLRVLNSRAIGVENA
T1017s1	LLLNDKQYNELCEAAEGRNLGAVFSYSEPEEPPPLNFSFEERKKIFLWVLTRLLKEGRIKLAKHGKFLEGSVDEQVERFRQAFPKTEEEMEDGIWFFDESCPGGAVWVLED
T1018	MITSSLPLTDLHRHLDGNIRTQTILELGQKFGVKLPANTLQTLTPYVQIVEAEPSLVAFLSKLDWGVAVLGDLDACRRVAYENVEDALNARIDYAELRFSPYYMAMKHSLPVTGVVEAVVDGVRAGVRD FGIQANLIGIMSRTFGTDACQQELDAILSQKNHIVAVDLAGDELGQPGDRFIQHFKQVRDAGLHVTVHAGEAAGPESMWQAIRDLGATRIGHGVKAIHDPKLMDYLAQHRIGIESCLTSNLQTSTVDSL ATHPLKRFLEHGILACINTDDPAVEGIELPYEYEVAAPQAGLSQEQIRQAQLNGLELAFLSDSEKKALLAKAALRG
T1023s3	MAGGEAGVTLGQPHLSRQDLTTLDVTKLTPLSHEVISRQATINIGTIGHVAHGKSTVVKAISGVHTVRFKNELERNITIKLGYANAKIYKLDDPSCPRPECYRSCGSSTPDEFPTDIPGTKGNFKLVRHVS FVDCPGHDILMATMLNGAAVMDAALLLIAGNESCPQPQTSEHLAAIEIMKLKHILILQNKIDLVKESQAKEQYEQILAFVQGTVAEGAPIIPISAQLKYNIEVVCEYIVKKIPVPPRDFTSEPRLIVIRSFDVNK PGCEVDDLKGGVAGGSILKGVLKVGQEIEVRPGIVSKDSEGKLMCKPIFSKIVSLFAEHNDLQYAAPGGLIGVGTKIDPTLCRADRMVGQVLGAVGALPEIFTELEISYFLLRRLLGVRTEGDKKAAKVQ KLSKNEVLMVNIGSLSTGGRVSAVKADLGKIVLTNPVCTEVGEKIALSRRVEKHWRLIGWGQIRRGVTIKPTVDDD

3.3 Results and Discussion

3.3.1 Summary of results from CASP11, CASP12 and CASP13

The main findings of the Chapter are outlined below:

- Good structural homology as shown by TM-align scores did not always equate to high MCC and BDT scores, thereby suggesting that overall tertiary structure is not the only or key determination in ligand and ligand-binding site predictions (see Figure 3.2, 3.3 and 3.4).
- The above led to an additional question around modelling of binding sites, in particular with respect to the modelling of flexible loops (see Figure 3.23), where the correct ligand is in the incorrect location, what can be done to improve this? Can protein-ligand docking be the answer?
- In some examples (see Figure 3.29) FunFOLD3 over-predicted the ligand-binding site due to inclusion of solvents used in the crystallisation process being included in the ligand-binding residues. Once again, this raised the question of whether protein-ligand docking can improve these predictions by focusing on the orientation of the ligand in the ligand-binding space and excluding the solvents.
- Results from CASP has shown that FunFOLD3 can predict a variety of ligands from small metal ions (e.g. calcium and magnesium), to enzymatic cofactors (e.g. nicotinamide adenine dinucleotide) across a variety of proteins (e.g. enzymes to viral proteins). This includes well annotated proteins to poorly annotated proteins, thus showing the diversity of FunFOLD3 in ligand and ligand-binding site prediction.
- The lowest MCC and BDT score was -0.05 and 0.035, respectively. Albeit not for the same CASP11 target (see Table 3.3). The highest MCC and BDT score was 1 (CASP13 target T0974s:temperate bacteriophage). Demonstrating the diversity of scores and also the potential for FunFOLD3 to predict MCC and BDT scores at the higher end of the scale.

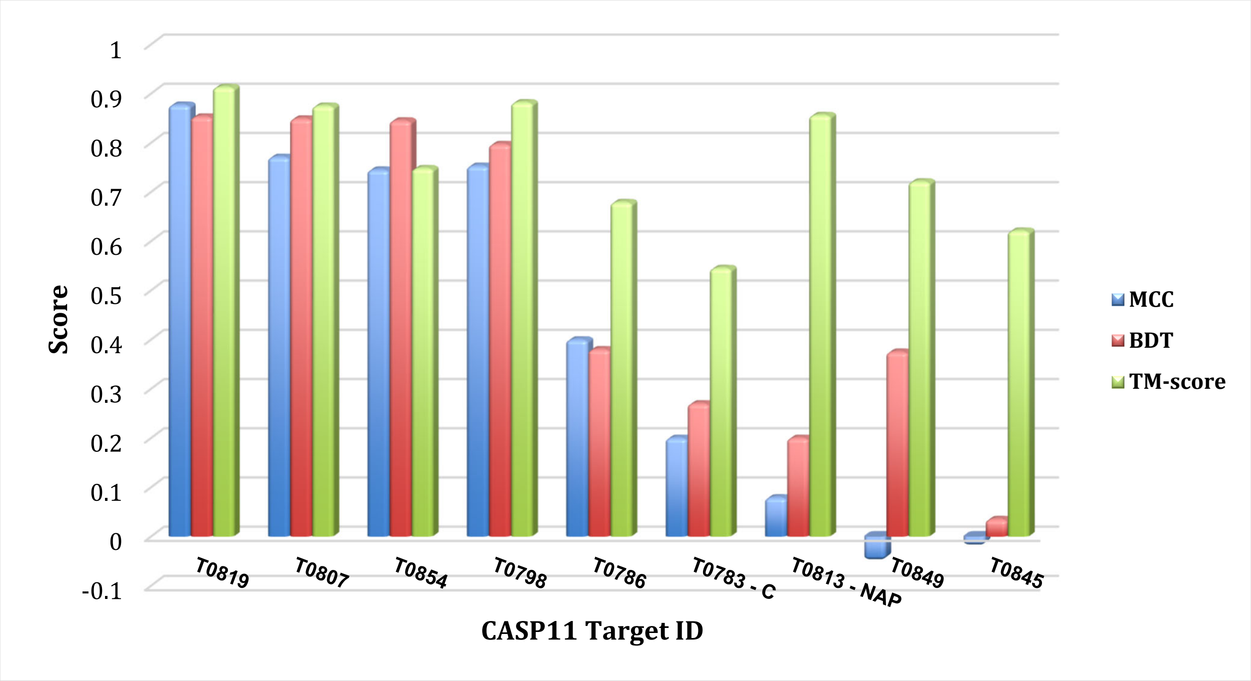


Figure 3.2. Comparison between BDT, MCC and TM-score for CASP11 targets
As can be seen from the figure, the higher TM-scores are associated with higher BDT and MCC scores. Except T0813, which has a higher TM-scores than T0854 but this has not correlated with a higher BDT or MCC score

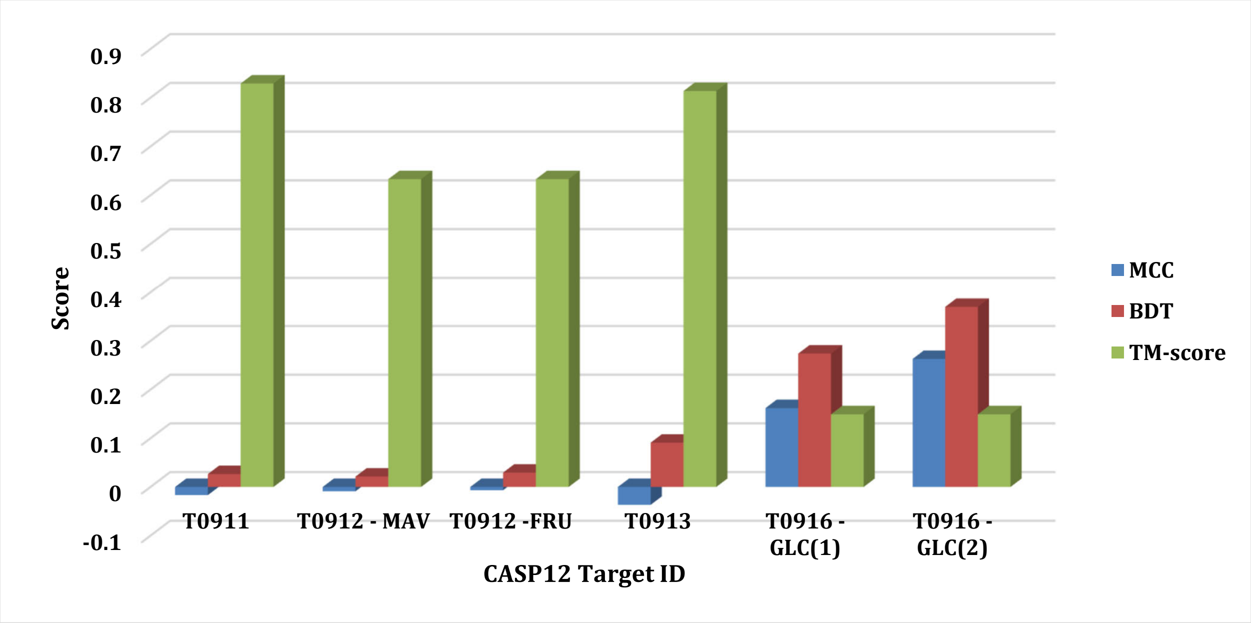


Figure 3.3. Comparison between BDT, MCC and TM-score for CASP12 targets
As can be seen from the figure, higher TM-scores were not necessarily associated with higher MCC or BDT scores, indeed the converse is true for T0916 NAD ligand when compared to the observed GLC ligand at the two different locations. T0909 has not been shown as there was no consensus between the predicted and observed ligands

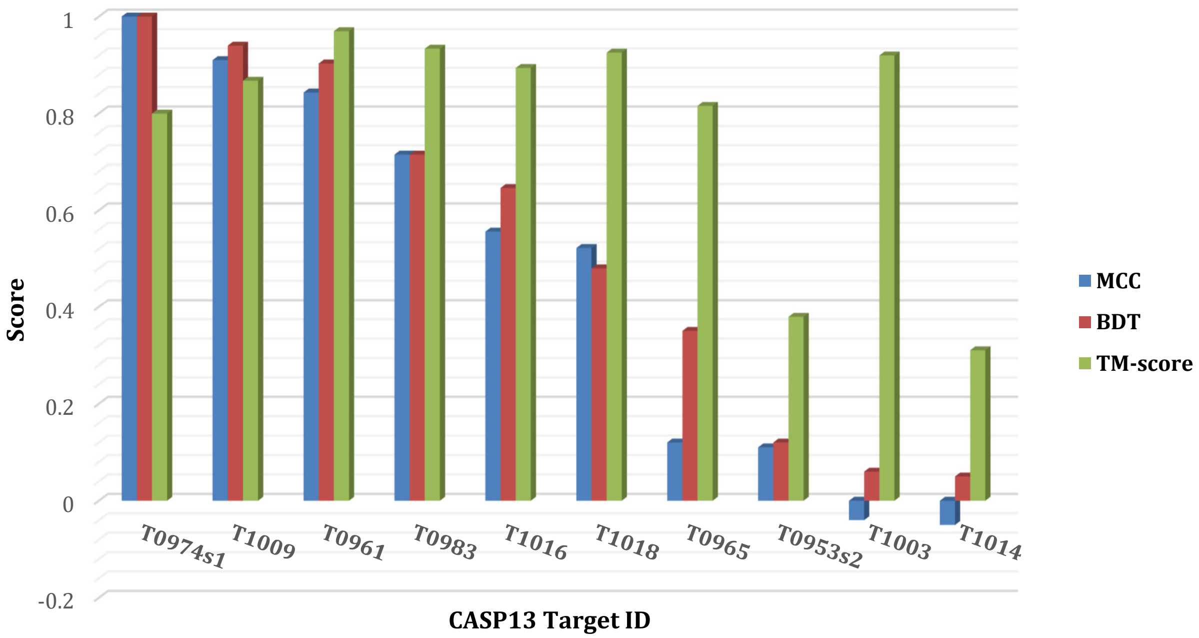


Figure 3.4. Comparison between BDT, MCC and TM-score for CASP13 targets
As can be seen from the figure, higher TM-scores were not necessarily associated with higher MCC or BDT scores. However, there was a general trend with higher TM-scores associated with better MCC and BDT scores.

3.3.2 Analysis of CASP11 Functional Prediction

Targets obtained from CASP11 with associated PDB IDs were analysed. In the first step targets were analysed using the BioLiP database to ascertain if they contained biologically relevant ligands. Next, the targets deemed to contain biologically relevant ligands were further investigated to identify the ligand site residues, this was done using The Van der Waal radius of the contacting atom of a residue and the containing ligand atom plus 0.5 Å.

A total of 39 CASP11 targets/domains were associated with PDB IDs (experimental structures that have been released into PDB). Analysis using the FunFOLD3 server yielded a total of nine proteins containing biologically relevant ligands and binding site residues. These were: T0783 (PDB ID 4cvh), T0786 (PDB ID 4qvu), T0798 (PDB ID 4ojk), T0807 (PDB ID 4wgh), T0813 (PDB ID 4wji), T0814 (PDB ID 4r7f), T0819 (PDB ID 4wbt), T0845 (PDB ID 4r5o) and T0849 (PDB ID 4w66), where the CASP11 target ID is given outside parenthesis and PDB IDs within parenthesis. Protein-ligand interactions were predicted for all the nine FN targets, with a mean MCC score of 0.391 and a mean BDT score of 0.431. Each of the predictions will be discussed in detail.

FunFOLD3 predicted the binding site residues for nine CASP11 targets. The corresponding observed binding site residues are also provided, along with *under* and *over*- predictions. Correct residues are highlighted in red as illustrated in Table 3.2. The top-three scoring predictions (Figure 3.5, 3.9 and 3.12) and a low scoring prediction (Figure 3.14) will be depicted in the Chapter with remaining predictions shown in Appendix 2. An overview of the MCC and BDT scores for all the targets is shown in Table 3.3.

Table 3.2. Predicted and observed ligand-binding site residues for CASP11 targets
Correct ligand binding site residues are depicted in red and bold and presented in ascending CASP11 target ID

CASP 11 target ID	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0783 **(PDB ID 4cvh)**	CTP ligand: 12,13,14,15,16,17,18,19,26,27,83,84, **85,86**,89,116,117,118,223 C ligand: 12,13,14,15,19,27,83,84,**85,86,**89,116,117,223 C5P ligand: 12,13,14,15,19,26,27,83,84,89,117, 118,223 CU ligand: 226,229	MG ligand: 85,86,87,194,195,197 CL ligand: 88,92	87,88,92,194, 195,197	12,13,14,15,16,17,18,19,26,27,82, 83,84,89,116,117,118, 223,226,229
T0786 **(PDB ID 4qvu)**	AMP ligand: 28,29,33,35,50,51,52,109,115,116, 117,118,122,124 FE ligand: 138,208,212 ZN ligand(01): 27,29,30,67,69,120 ZN ligand(02): **152**,154	ZN ligand: 152,177,180	177,180*	154*
T0798 **(PDB ID 4ojk)**	13,**14,15,16,17,18,19,29,30,31,**33,35, 36, 61,62, **117,118,120,121,147, 148,149**	12,14,15,16,17,18,19,29,30,31,32,34,117, 118,120,121,147,148,149	12,32,34	13,33,35,36,61,62
T0807 **(PDB ID 4wgh)**	**20,21,22**,23,**50**,54,**55**,113,**143,165, 193,194,195,196**,197,**198,199**,200, **201**,207,**224,240,241,242,244,248, 251**	20,21,22,50,55,80,142,143,165,193,194, 195,196,198,199,201,224,240,241,242,243 ,244,245,248,251,252	80,142,243, 245,252	23,54,113,197,200, 207

T0813 **(PDB ID 4wji)**	NAI ligand: 11,15,16,38,72,73,74,75,77,98,100,123,124,127,128,131,132,235 NAD ligand: 11,12,**13**,14,15,16,37,38,72,73,74,75,98,100,123,127,128,131,132,235 NAP ligand: 11,12,14,15,16,36,37,38,39,**42**,55,72,73,74,75,98,100,127,128,131,132,235	MG ligand:13,42,46,133	13,46,133	NAP ligand: 11,12,14,15,16,37,38,39,72,73,74,75,98,100,123,127,128,131,132,235 NAD ligand: 11,12,14,15,16,37,38,72,73,74,75,98,100,123,127,128,131,132,235 NAI ligand: 11,15,16,38,72,73,74,75,77,98,100,123,124,127,128,131,132,235
T0819 **(PDB ID 4wbt)**	**93,94,95,119,167,194,197,223,225,226,234**, 347	93,94,95,119,161,167,194,196,197,223,225,226,234	196, 161	347
T0845 **(PDB ID 4r5o)**	130,163,165,177,209,210,263,396,442,443	58,228,237,273,275, 377,379,433	58,228,237,273,275,377,379,433	130,163,165,177, 209,210,263,396,442,443
T0854 **(PDB ID 4rn3)**	**16,18**,19,**173**	16,18,173,177	177	19
T0849 (**PDB ID 4w66)**	9,10,14,15,54,55,56,67,68,108,113,226, 230	168,171,179,182,183,190,194,197	168,171,179, 182,183,190,194,197	9,10,14,15,54,55,56,67,68,108,113,226, 230

* the under- and over-predictions are compared against ZN ligand(02) only

Table 3.3. MCC and BDT scores for CASP11 targets
A list of CASP11 targets with associated MCC and BDT scores. The results are listed from ascending to descending order by MCC and BDT score. For CASP11 target T0783 MCC and BDT scores are compared to the Magnesium observed ligand

CASP11 target	MCC Score	BDT Score
T0819	0.877	0.853
T0807	0.771	0.849
T0854	0.7451	0.845
T0798	0.753	0.797
T0786	ZN(1):-0.014 ZN(2):0.40	ZN(1):0.0139 ZN(2):0.38
T0783	CTP: 0.17 C:0.20	CTP:0.21 C:0.27
T0813	NAD:0.086 NAI:-0.029 NAP:0.079	NAD:0.19 NAI:0.11 NAP:0.2
T0849	-0.05	0.0375
T0845	-0.02	0.035

Figure 3.2 in Section 3.3.1 relates the MCC and BDT scores to the TM-score. As can be seen from the figure, the targets with the highest MCC and BDT score were also associated with the highest TM-scores.

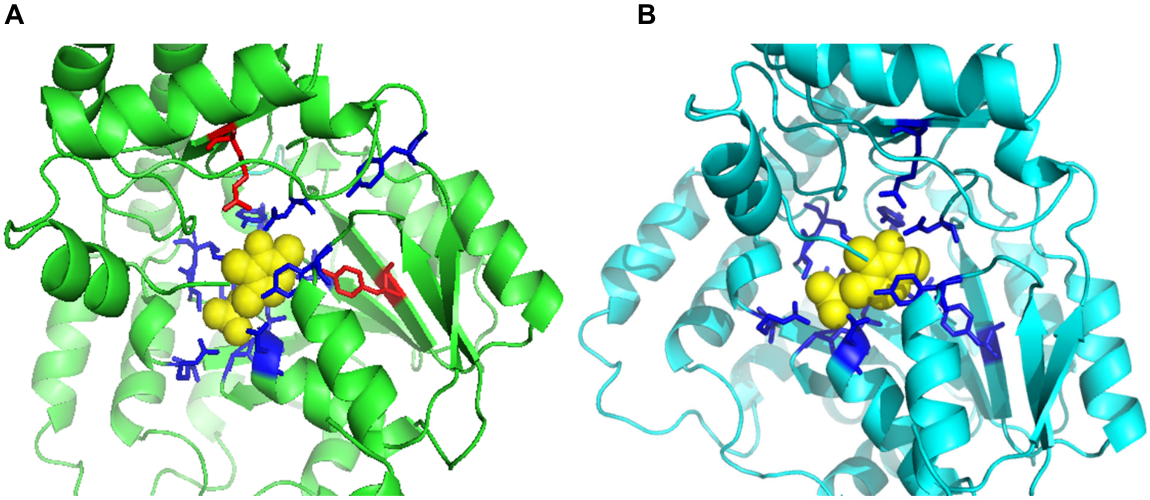


Figure 3.5. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0819 (PDB ID 4wbt). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the pyridoxal-5'-phosphate (PLP) ligand is shown as a sphere and coloured yellow. BDT score of 0.853 and MCC score of 0.877. **(B)** The observed ligand binding site for T0819 (PDB ID 4wbt), with binding site residues shown as sticks and coloured in blue and the ligand PLP coloured yellow

The CASP11 target with the highest BDT (0.853) and MCC (0.877) score is histidinol-phosphate aminotransferase (HisC) from *Sinorhizobium meliloti* (CASP ID T0819 and PDB ID 4wbt). There were two underpredictions (TYR 161 and ALA 196) and one overprediction (ARG 347). An underprediction is a ligand binding site residue that was missed in the prediction whereas, an overprediction is a ligand binding site residue that was predicted to be in contact with a ligand but was not found to be in contact in the observed structure.

Histidinol-phosphate aminotransferase is a pyridoxal 5'-phosphate-dependent (PLP) enzyme, that catalyses the reversible transamination reaction between histidinol phosphate (His-P) and 2-oxoglutarate (O-Glu). Figure 3.5B illustrates the observed binding site for PLP, which is located at the bottom of deep cavities formed at interfaces between the two domains of each subunit (Sedgwick, 2015) as better illustrated in Figure 3.6 which is the structure from PDB and shows all the chains (A, B and C) from the protein and the binding of PLP to each of these chains.

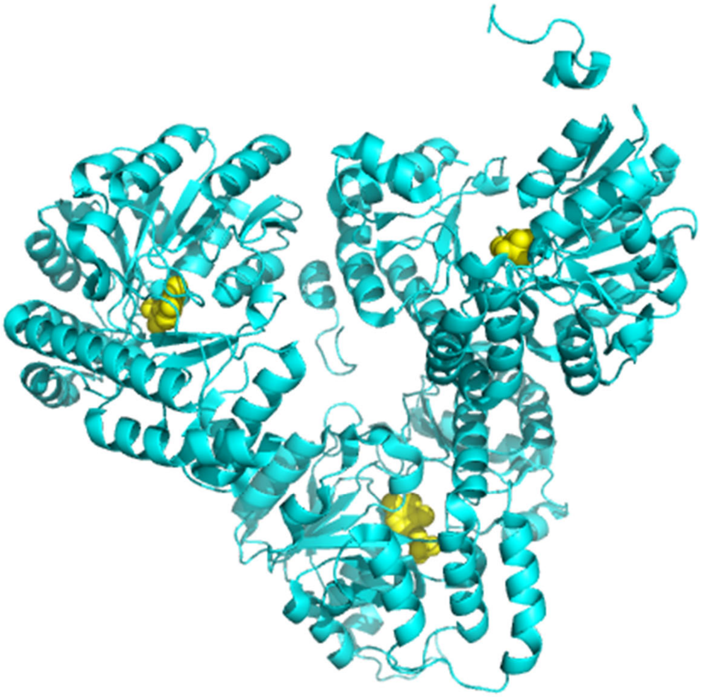


Figure 3.6. Histidinol-phosphate aminotransferase from Sinorhizobium meliloti (PDB ID 4wbt) bound to PLP. Proteins are shown in cartoon form and coloured in cyan and ligands shown as spheres and coloured yellow

When PLP binds to HisC (from *Corynebacterium glutamicum*) it is lined by conserved residues such as GLY 97, SER 98, ALA 199, TYR 200, THR 225, LYS 228 and ARG 236 (Marienhagen *et al.*, 2008). The active site is made up of residues located on the central β-strands of the β-sheet and loops; this is supported by Figure 3.6. The incorrect predictions were also located on these parts of the protein; ALA 196 is located on a loop and both TYR 161 and ARG 347 are on β-sheets. Figure 3.7 shows the view behind the PLP ligand to illustrate the underprediction for ALA 196.

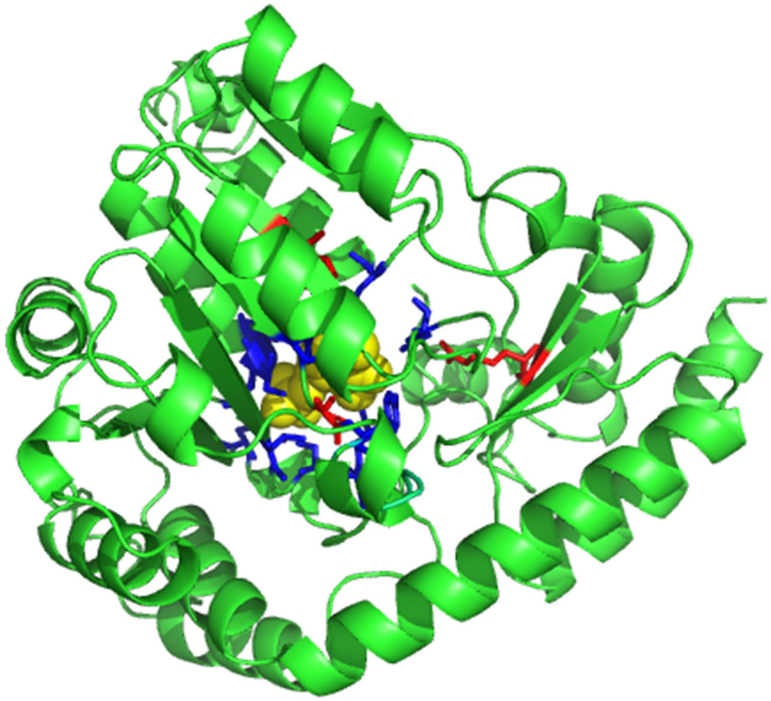


Figure 3.7. Reversed view of CASP11 target T0819 (PDB ID 4wbt)
Illustrating the underprediction GLY 19. Correctly predicted binding site residues shown as sticks and coloured in *blue* and incorrect predictions in *red*, the pyridoxal-5'-phosphate (PLP) ligand is shown as a sphere and coloured yellow

The reason why the BDT and MCC score were closer to a perfect prediction, than a random prediction is because there were a limited number of incorrect predictions; only three. In addition, one of the incorrect predictions (ALA 196), is close to the binding site and conserved residues. A further reason is the high level of molecular similarity between the predicted structure and the observed structure with a TM-score of 0.913 (normalised by the average length of two chains). A TM-score of <0.5 and <1.00, suggests high structural similarity. Figure 3.8, illustrates the TMalign superpostion of the observed structure (blue) and predicted model (red). The main difference between the two structures appears to be the flexible loop, which is present in the predicted protein model but not the observed protein model.

This high level of structural similarity meant large majorities of ligand binding site residues were correctly predicted. The observed protein model had 367 residues, whereas the predicted protein model has 373 residues.

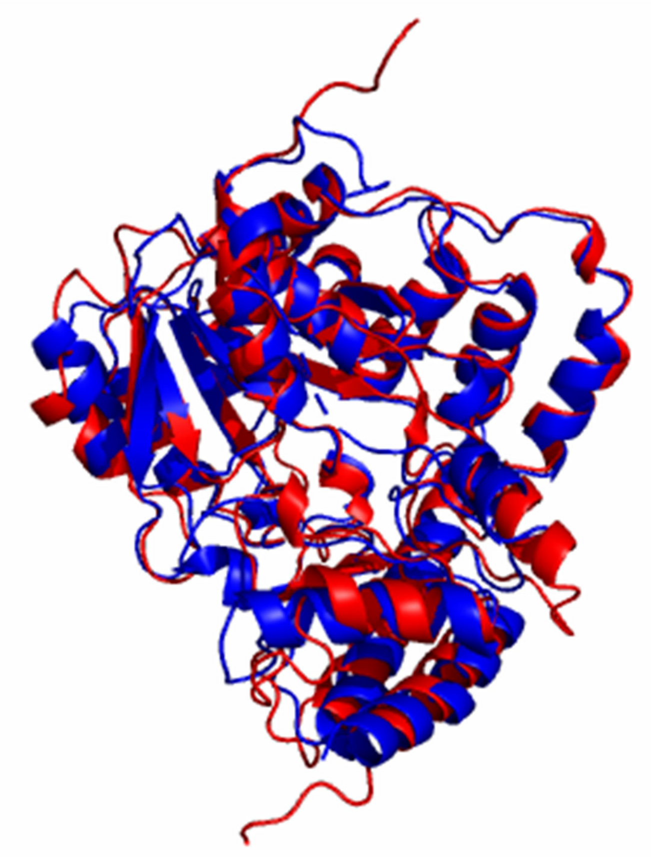


Figure 3.8. Comparison of TMalign (Zhang & Skolnick, 2005) structures for CASP11 target T0819 (PDB ID 4wbt). The structure in blue is the observed structure for T0819 and the structure in red is the predicted structure from IntFOLD3 (McGuffin *et al.*, 2015)

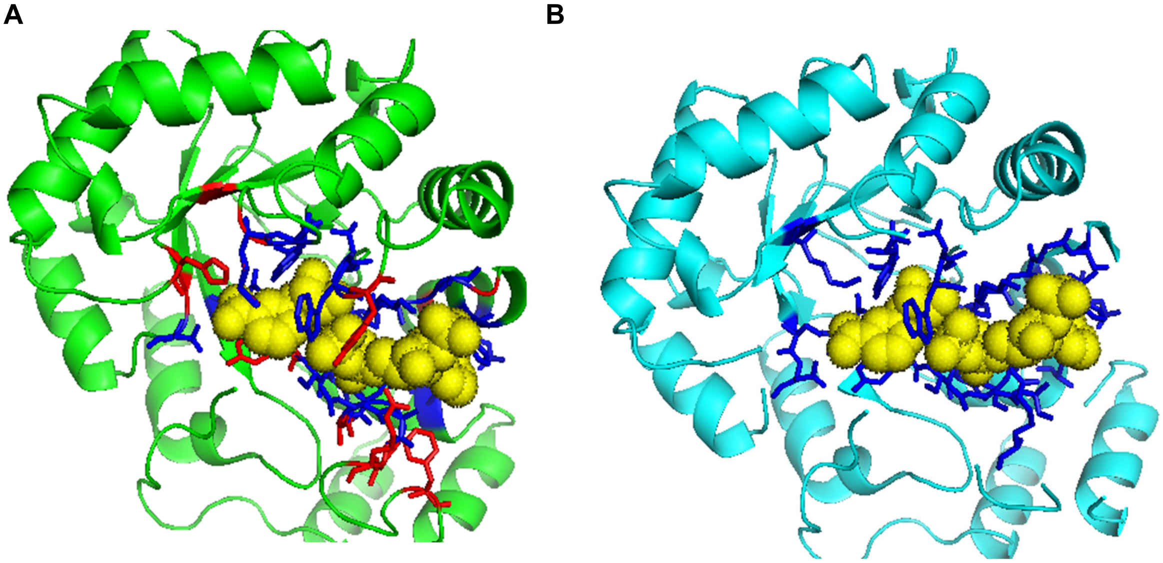


Figure 3.9. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0807 (PDB ID 4wgh). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the nicotinamide-adenine-dinucleotide (NAP) ligand is shown as a sphere and coloured yellow. BDT score of 0.849 and MCC score of 0.771. **(B)** The observed ligand binding site residues shown as sticks for T0807 (PDB ID 4wgh), with binding site residues coloured in blue and the ligand NAP coloured yellow

The second best predicted CASP11 target is aldo/keto reductase from *Klebsiella pneumoniae* (CASP 11 ID T0807 and PDB ID 4wgh). There were five underpredictions; (LYS 80, SER 142, ALA 243, ALA 245, ASN 252) and six overpredictions (TYR 23, MET 54, HIS 113, ALA 197, GLY 200, PHE 207).

Aldo/keto reductases (AKRs) are NAD(P)(H)-dependent oxidoreducatases that comprise a multigene superfamily (Marienhagen *et al.*, 2008). Aldo/keto reductases are capable of catalysing the reduction of aldehydes or carbonyl groups present in a variety of biochemicals (Hur *et al.*, 2009). Additionally, AKRs may play a role in the modification or detoxification of various biologically active compounds (Hur *et al.*, 2009). The ligand nicotinamide adenine dinucleaotide phosphate (NAP), is responsible for enzymatic reduction reactions of the AKR family (Hur *et al.*, 2009).

Despite the higher number of incorrect predictions compared to CASP 11 target T0819; 11 predictions as opposed to three, the BDT and MCC score are still closer towards a perfect prediction (0.849 and 0.771, respectively). Thereby, showing that the number of incorrect predictions has an impact on the MCC/BDT score, but does not necessarily correlate with a much lower score. The main reason for the incorrect predictions could be due to extension of the ligand-binding site by the inclusion of other ligands that are closely bound to NAP. Selenomethionine (MSE) is located in residue MET 54, which was one of the overpredictions and is next to residue TYR 55 for NAP and this could be a reason as to why residues around MSE could have been predicted as a ligand-binding site. As part of the FunFOLD3 algorithm ligands are considered part of the cluster if any of their atoms were in contact with the continuous mass – as MSE is close to NAP it could be that FunFOLD3 picked up residues related to MSE and this lead to overpredictions. Figure 3.10 below shows the amino acid MSE and the ligand NAP bound to aldo/keto reductase and illustrates how the close the binding site for NAP overlap is to the crystal structure of MSE. Selenomethionine is a modified amino acid that is used in single-wavelength anomalous diffraction (SAD) and multi-wavelength anomalous dispersion (MAD) X-ray crystallography to help determine the structure (Mayr & Nidetzky, 2002).

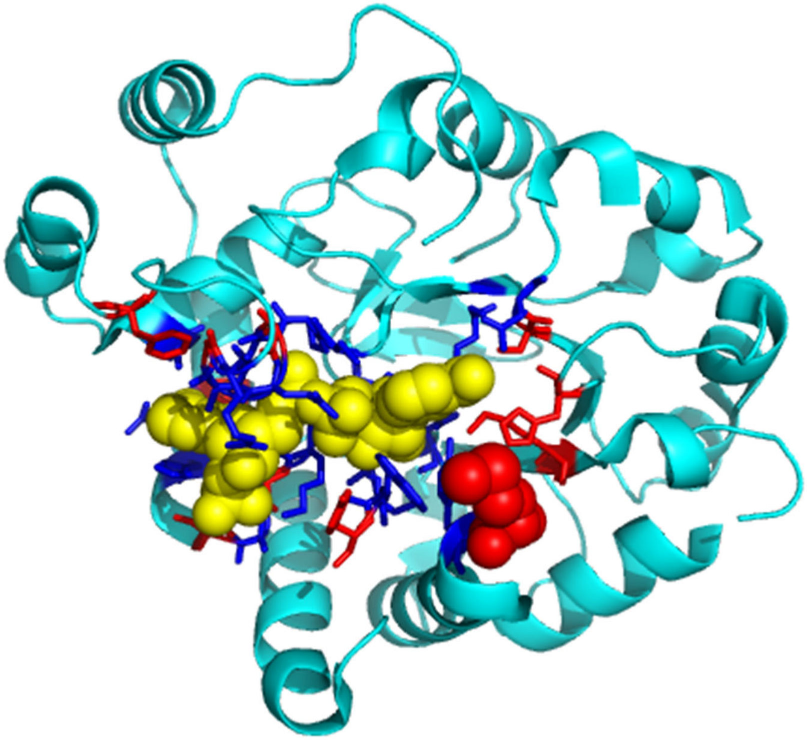


Figure 3.10 Comparison of the ligand binding site for NAP and MSE for CASP 11 target T0807 (PDB ID 4wgh). Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in *blue* and incorrect predictions in *red*, the NAP ligand is coloured yellow and MSE is coloured red, as it is an overprediction. The protein model is coloured cyan, as the structure has been obtained from PDB

A TM-score of 0.875 was obtained, showing a high level of molecular similarity between the observed and predicted molecular structures. Despite the lower structural similarity between the two protein molecules, the number of residues was closely matched with 283 residues for the observed protein and 284 residues for the predicted protein. An explanation for the lower TM-score for these models compared to T0819, despite having a closer number of residues could be due to the folding of the protein molecule. The flexible loop on the predicted model appears to be longer and extends further than the flexible loop on the observed protein. As with CASP11 target T0819, this high level of structural similarity resulted in BDT and MCC scores, which were more aligned with a perfect prediction. The TMalign superposition of observed and predicted structures is shown in Figure 3.11.

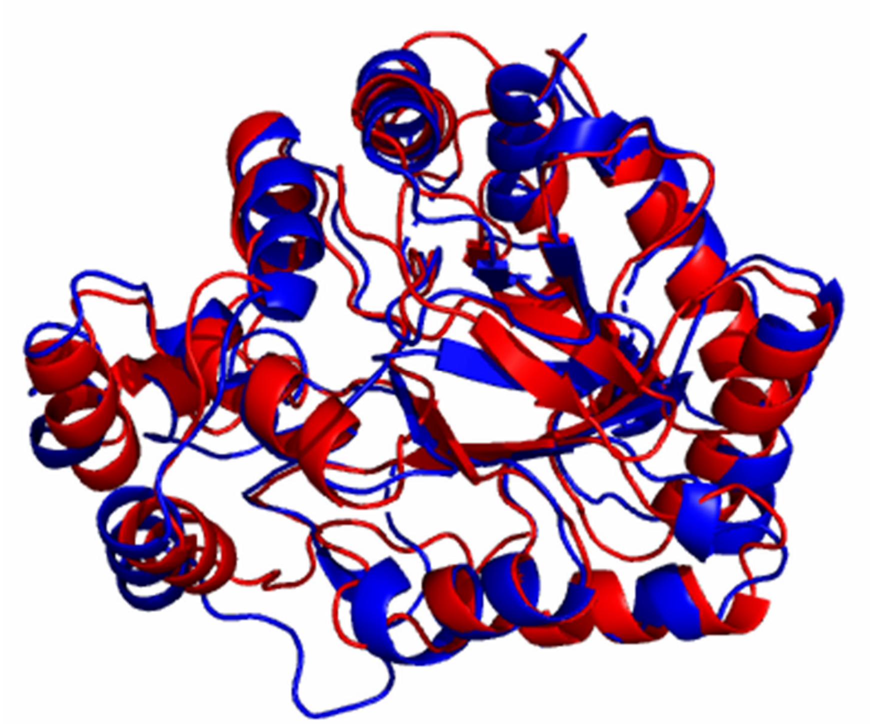


Figure 3.11. Comparison of TMalign (Zhang & Skolnick, 2005) structures for CASP11 target T0807 (PDB ID 4wgh). The structure in blue is the observed structure for T0807 and the structure in red is the predicted structure from IntFOLD3 (McGuffin *et al.*, 2015)

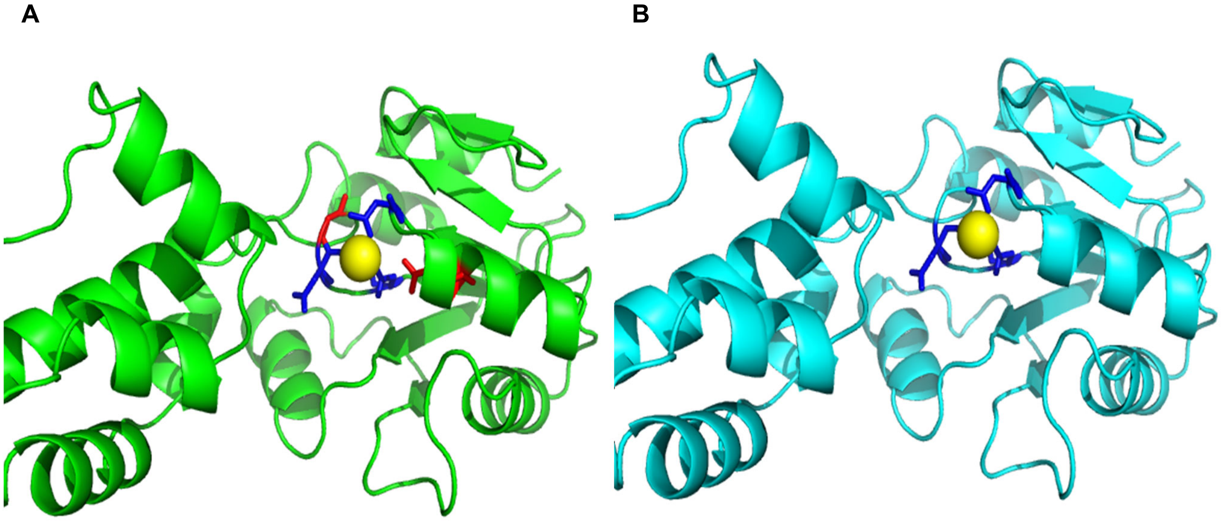


Figure 3.12. Comparison of FunFOLD3 ligand binding site predictions for CASP11 target T0854 (PDB ID 4rn3). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the magnesium (MG) ligand is shown as a sphere and coloured yellow. BDT score of 0.845 and MCC score of 0.7451. **(B)** The observed ligand binding site residues shown as sticks for T0854 (PDB ID 4rn3), with binding site residues coloured in blue and the ligand MG coloured yellow

The third best predicted CASP11 target is HAD-superfamily hydrolase, subfamily IA, variant 1 from *Geobacter sulfurreducens* (CASP 11 T0854 and PDB ID 4rn3). There were two incorrect predictions; with one underprediction (GLY 19) and one overprediction (ASP 177). The underprediction GLY 19 was located on a flexible loop of the protein.

HAD-superfamily hydrolase, as the name suggests, belongs to a large superfamily of hydrolases with a wide variety of substrate specificity. The superfamily consists of epoxide hydrolases and different types of phosphatases (Hendrickson, 1999). The FunFOLD3 server identified the ligand as magnesium and this was supported by the observed binding site. All HAD phosphoaspartyl transferases use magnesium as an obligatory cofactor. Magnesium aids in the correct positioning of the substrate phosphoryl group relative to the ASP nucleophile. Additionally, magnesium provides charge neutralisation of the transition state (Koonin & Tatusov, 1994).

A TM-score of 0.748 was obtained, showing a high level of molecular similarity between the observed and predicted molecular structures. The number of residues was also closely matched, with the observed structure having 210 residues and the predicted structure having 212 residues. As with CASP11 target T0807, despite the similar number of residues the variances in folding of the protein has resulted in a lower TM-score as compared with CASP11 target T0819. One of the big differences is with the flexible loop of the predicted model, which has not bonded with the α helix. Additionally, there are differences in the alignment of some of the α helices of the predicted and observed protein model, which could have resulted in a lower TM-score, despite the high number of residues.

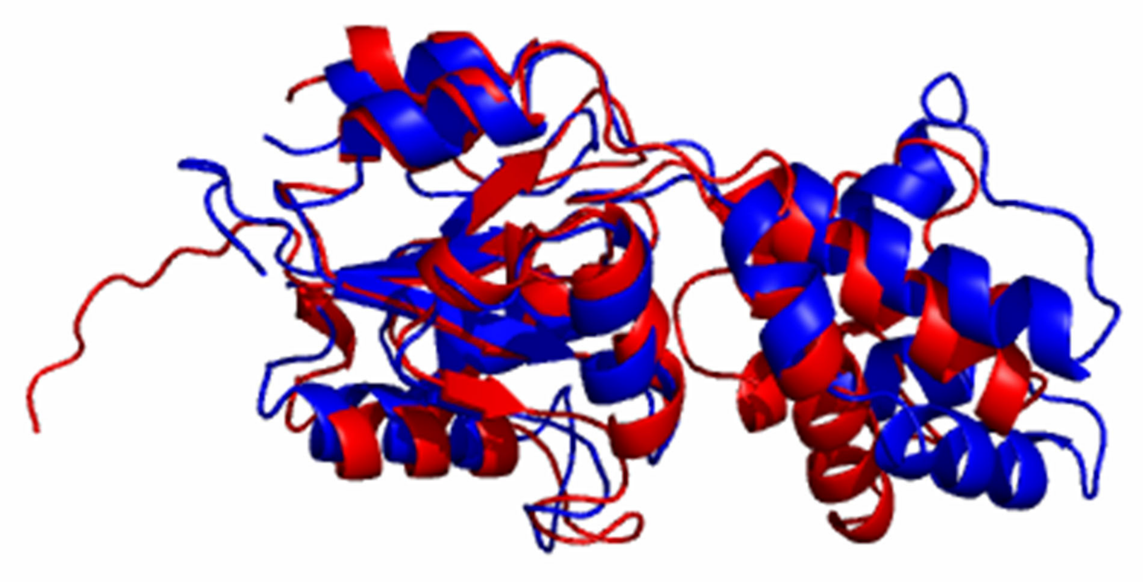


Figure 3.13. Comparison of TMalign (Zhang & Skolnick, 2005) structures for CASP11 target T0854 (PDB ID 4rn3). The structure in blue is the observed structure for T0854 and the structure in red is the predicted structure from IntFOLD3 (McGuffin *et al.*, 2015)

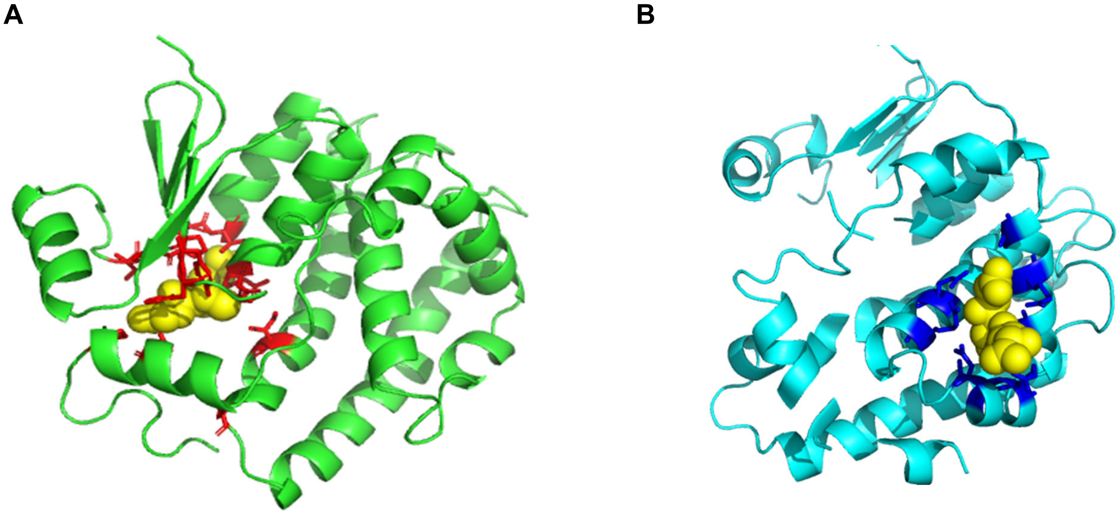


Figure 3.14. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0849 (PDB ID 4w66). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions red, the glutathione (GSH) ligand is shown as a sphere and coloured yellow. BDT score of 0.0375 and MCC score of -0.05. **(B)** The observed ligand binding site residues shown as sticks for T0849 (PDB ID 4w66), with binding site residues coloured in blue and the ligand GSH coloured yellow

The eighth predicted CASP11 target is Glutathione S-transferase domain protein from *Haliangium ochraceum* (CASP 11 T0849 and PDB ID 4w66). There were no correct predictions with this protein.

Glutathione S-transferase (GSTs), are a family of cytosolic enzymes, which are involved in the detoxification of exogenous and endogenous species. The protein is also prone to polymorphisms which can have an impact of the drug metabolism (Xie, Bonner and Jensen, 2000).

The FunFOLD3 server has correctly identified the glutathione ligand, however the predictions have failed to identify any of the correct ligand binding site residues. This may be due, in part to the large size of the ligand or could be as a result of the amino acid MSE being in close proximity to the GSH ligand and causing the extension of the ligand binding site in some parts of the protein.

A TM-score of 0.721 was obtained, which as with T0813 is quite surprising, given the low BDT and MCC score, which was obtained (0.0375 and 0.05, respectively). The number of residues for the observed protein was 236 and for the predicted protein was 240. This suggests, that the reason for the low score is to do with the ligand binding site and not poor structural similarity.

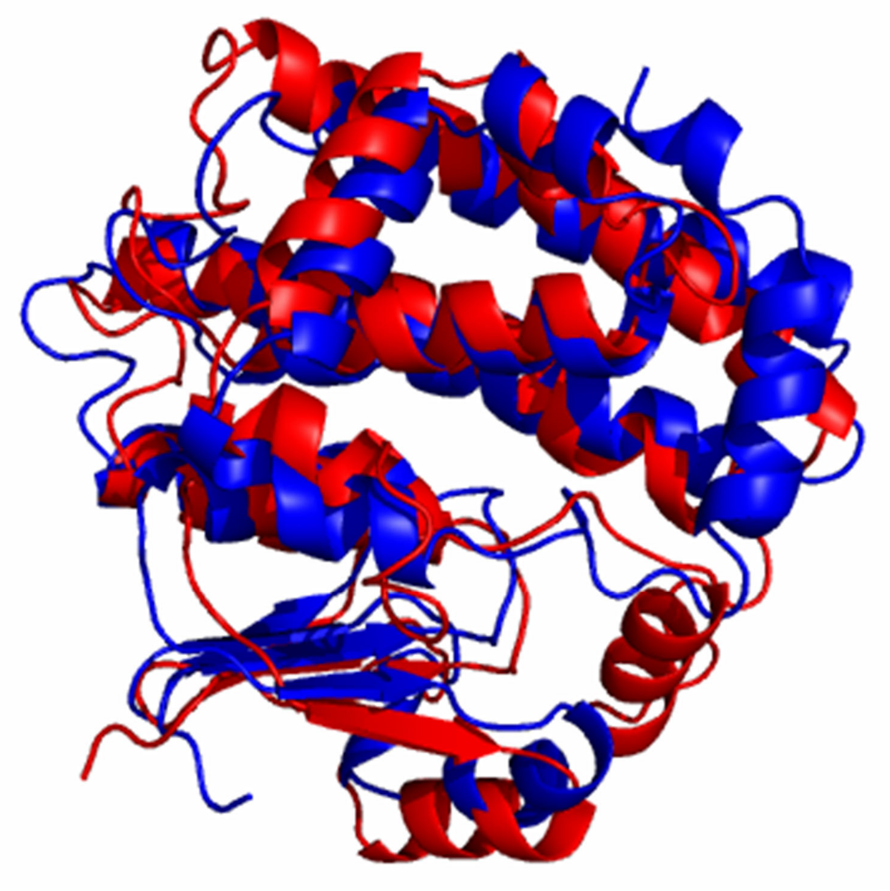


Figure 3.15. Comparison of TMalign (Zhang & Skolnick, 2005) superposition for CASP11 target T0849 (PDB ID 4w66). The structure in blue is the observed structure for T0849 and the structure in red is the predicted structure from IntFOLD3 (McGuffin *et al*., 2015)

Another consideration in the analysis of the CASP11 targets is the ligands that are bound to the protein. The following compares the ligands identified using the FunFOLD3 webserver and those identified using the PDB ID with the associated CASP11 protein as shown in Table 3.4. Comparison of ligands predicted by CASP11 targets T0783, T0786 and T0819 and the ligands associated with the PDB entry are shown in Figure 3.16.

Table 3.4 below, shows the comparison between the ligands predicted by FunFOLD3 and the associated ligands as per the PDB entry.

Table 3.4. Comparison of ligands predicted using FunFOLD3 and ligands identified on Protein DataBank

CASP 11 target ID	FunFOLD3 ligand	PDB ligand
T0783 (PDB ID 4cvh)	Cytidine-5'-triphosphate, , cytidine-5'-monophosphate, copper ion	Ethylene glycol (EDO), chloride ion, magnesium ion
T0786 (PDB ID 4qvu)	Zinc, Iron, adenosine monophosphate	Tetraethylene glycol (PG4), zinc, sodium
T0798 (PDB ID 4ojk)	Guanosine-5'-diphosphate (GDP)	Guanosine-5'-diphosphate (GDP)
T0807 (PDB ID 4wgh)	2'-monophosphoadenosine 5'-diphosphoribose (NAP)	2'-monophosphoadenosine 5'-diphosphoribose (NAP), acetate (ACT)
T0813 (PDB ID 4wji)	1,4-dihydronicotinamide adenine dinucleotide (NAI), 2'-monophosphoadenosine 5'-diphosphoribose (NAP), nicotinamide adenine dinucleotide (NAD)	Magnesium ion, chloride ion, tyrosine (TYR), 2'-monophosphoadenosine 5'-diphosphoribose (NAP)
T0819 (PDB ID 4wbt)	Vitamin B6 Phosphate (PLP)	Polyethylene glycol (PE4), vitamin B6 Phosphate (PLP), triethylene glycol (PGE), di(hydroxyethyl)ether (PEG), glycerol (GOL)
T0845 (PDB ID 4r5o)	Calcium ion, chloride ion	Calcium ion, chloride ion, polyethylene glycol (7PE), acetate (ACT)
T0854 (PDB ID 4rn3)	Magnesium ion	Ethylene glycol (EDO), Magnesium ion
T0849 **(**PDB ID 4w66)	Glutathione	Glutathione

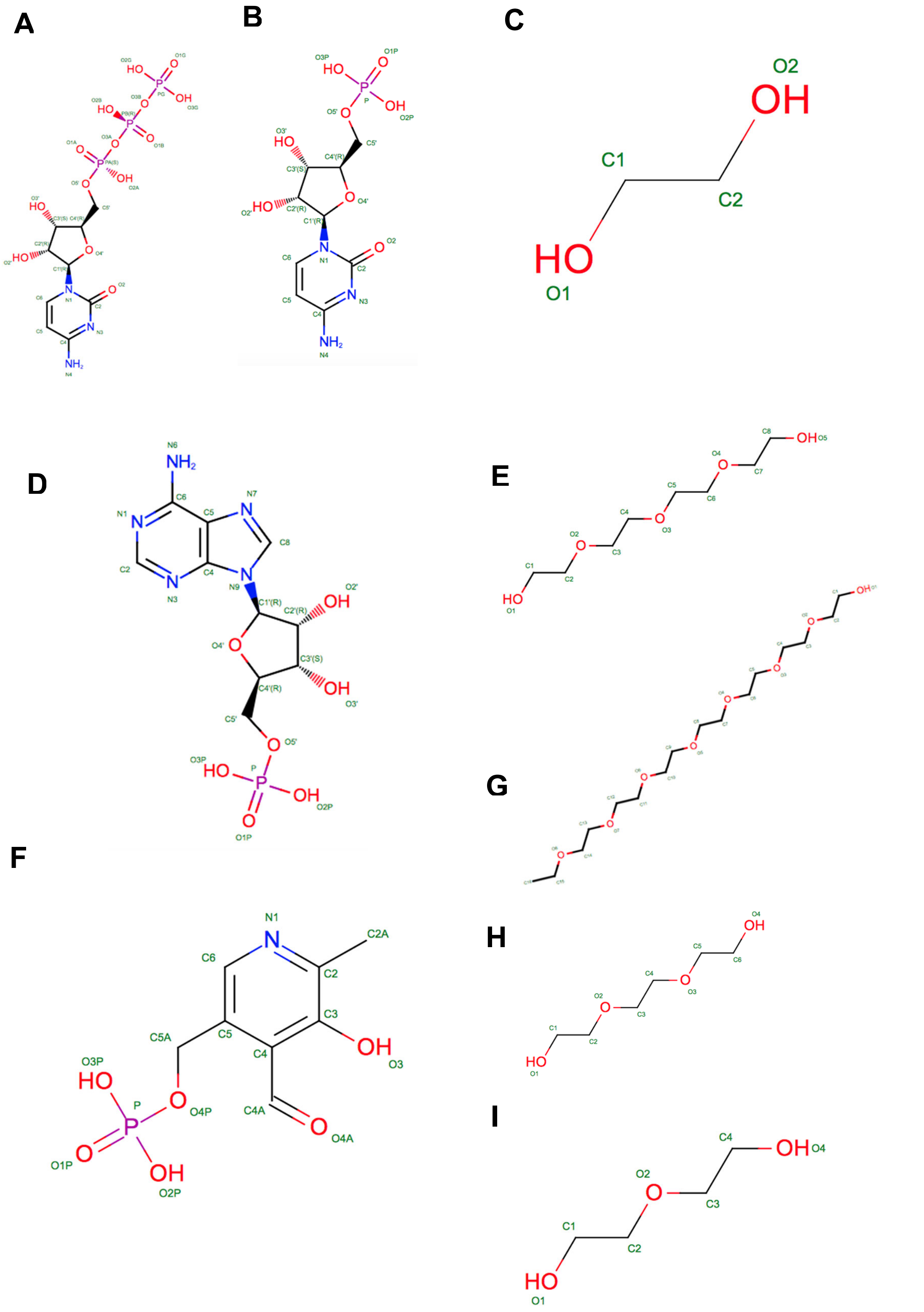


Figure 3.16. Comparison of predicted ligands by FunFOLD3 and ligands as per the PDB entry for CASP11 targets T0783 (PDB ID 4cvh), T0786 (PDB ID 4qvu) and T0819 (PDB ID 4wbt)
(A) CASP11 target T0783 predicted ligand cytidine-5'-triphosphate **(B)** CASP11 target T0783 predicted ligand cytidine-5'-monophophate **(C)** Ethylene glycol ligand associated with PDB ID 4cvh **(D)** CASP11 target T0786 predicted adenosine monophosphate ligand **(E)** Tetraethylene glycol ligand associated with PDB IS 4qvu **(F)** CASP11 target T0819 Predicted vitamin B6 phosphate ligand **(G)** Polyethylene glycol **(H)** Triethylene glycol and **(I)** Di(hydroxyethyl)ether ligands associated with PDB ID 4wbt. Images obtained from (Burley *et al.*, 2017)

3.3.3 Analysis of CASP12 Functional Prediction

As with CASP11, targets obtained from CASP12 were analysed. Ligand-binding site residues predictions are reported on regardless of if there is an associated PDB ID. It is likely all CASP12 targets will have a PDB ID associated in the foreseeable future by the CASP12 organisers and will be analysed as and when the PDB IDs become available.

The same rationale as per CASP11, was applied in CASP12 with biologically relevant ligands being ascertained as per BioLiP database and the ligand binding residues determined. In a change from CASP11, where all ligands were identified – this will also be part of CASP12 analysis – there is an additional focus on identification of the middle/centroid ligand of the protein.

A total of 61 CASP12 targets were released for analysis with some targets were in complex with other targets. Analysis using the FunFOLD3 server yielded a total of twelve proteins containing biologically relevant ligands and binding site residues. These were T0868 (PDB ID 5j4a), T0872 (PDB ID 5jmb), T0892 (PDB ID 5nv4), T0899, T0901, T0905, T0907, T0909 (PDB ID 5g5n), T0911, T0912 (PDB 5mqp), T0916 and T0919. CASP12 target IDs are given outside parenthesis and PDB IDs, where applicable within parenthesis. Protein-ligand interactions were predicted for all of the twelve FN targets. Once all of the predicted CASP targets have PDB IDs associated then a mean MCC score and BDT score will be calculated. Each of the predictions will be discussed in detail.

FunFOLD3 predicted the binding site residues for twelve CASP12 targets. The corresponding observed binding site residues are also provided, along with *under* and *over*-predictions. Correct residues are highlighted in red as illustrated in Table 3.5. For further results, please refer to Appendix 2. A comparison of the MCC, BDT and TM-scores for the targets are shown in Figure 3.3 and a summary of the MCC and BDT scores for each of the CASP12 targets is shown in Table 3.6

Table 3.5. Predicted and observed ligand-binding site residues for CASP12 targets
Correct ligand binding site residues are depicted in red and bold and presented in ascending CASP12 target ID. For CASP12 targets where there is no ligand predicted in the experimental structure the observed ligand-binding site residue column is blank

CASP12 target	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0868 **(PDB ID 5j4a)**	39,40,44,45,46,98	No biologically relevant ligands	N/A	N/A
T0872 **(PDB ID 5jmb)**	6,7,68	No biologically relevant ligands	N/A	N/A
T0899	79,178,214,215	No PDB ID released	N/A	N/A
T0901	54,57,130,131,169,170	No PDB ID released	N/A	N/A
T0905	58,239,240,241	No PDB ID released	N/A	N/A
T0907	82,83,84,85,112	No PDB ID released	N/A	N/A

T0909 **(PDB ID 5g5n)**	146,147,169,170,190,207,208, 232	CL ligand(1):
		289
		
		CL ligand(2):
		359,360
		
		CL ligand(3):
		85,87,126
		
		CL ligand(4):
		44,46,187,218,252
		
		CL ligand(5):
		211,218,223,241,242,243,252,286
			CL ligand(1):
			289
			
			CL ligand(2):
			359,360
			
			CL ligand(3):
			85,87,126
			
			CL ligand(4):
			44,46,187,218,252
			
			CL ligand(5):
			211,218,223,241,242,243,252,286
				46,147,169,170,190,207,208, 232

T0911 **(PDB ID 6e9n)**	44,160,164,165,393	68,123,126,358,371,374,375,377, 378	68,123,126,358,371,374,375,377, 378	44,160,164,165,393
T0912 **(PDB ID 5mqp)**	MAV ligand: 334,340,423,426,468,469,470 FRU: 207,208,209,490	155,262,264,268	155,262,264,268	207,208,334,340,423,426, 468,469,470,471,474,490
T0913	100,149,153,156,171,206,266, 267,318,358,359,364	64,65,66,67,209,210,273,274, 320,321,363,368,371	64,65,66,67,209,210,273,274,320, 321,363,368,371	100,149,153,156,171,206, 266,267,318,358,359,364
T0916 **(PDB ID 5tj4)**	**14**,**15**,16,17,18,51,59,60,62,**63**, 64,68,107*	GLC ligand(1) 62,63,65,66,153,154,155,340,344	N/A	N/A

T0919	39,203,204,207,271,272,273	GLC ligand(2): 12,14,15,63,111,153,155,156,230 Structure cancelled by organisers	N/A	N/A

*compared against GLC ligand(2)

Table 3.6. MCC and BDT scores for CASP12 targets
A list of CASP12 targets with associated MCC and BDT scores. The results are listed from ascending to descending order by CASP12 target ID. For CASP12 target T0916 MCC and BDT scores are compared to the GLC observed ligand, NAD(1) would be compared against GLC(1) observed ligand and so forth

CASP12 target ID	MCC score	BDT score
T0911	-0.0167	0.0265
T0912	MAV: -0.00892 FRU: -0.00672	MAV:0.0213 FRU: 0.0295
T0913	-0.0367	0.091
T0916	NAD(1): 0.162 NAD(2): 0.263	NAD(1):0.263 NAD(2):0.37

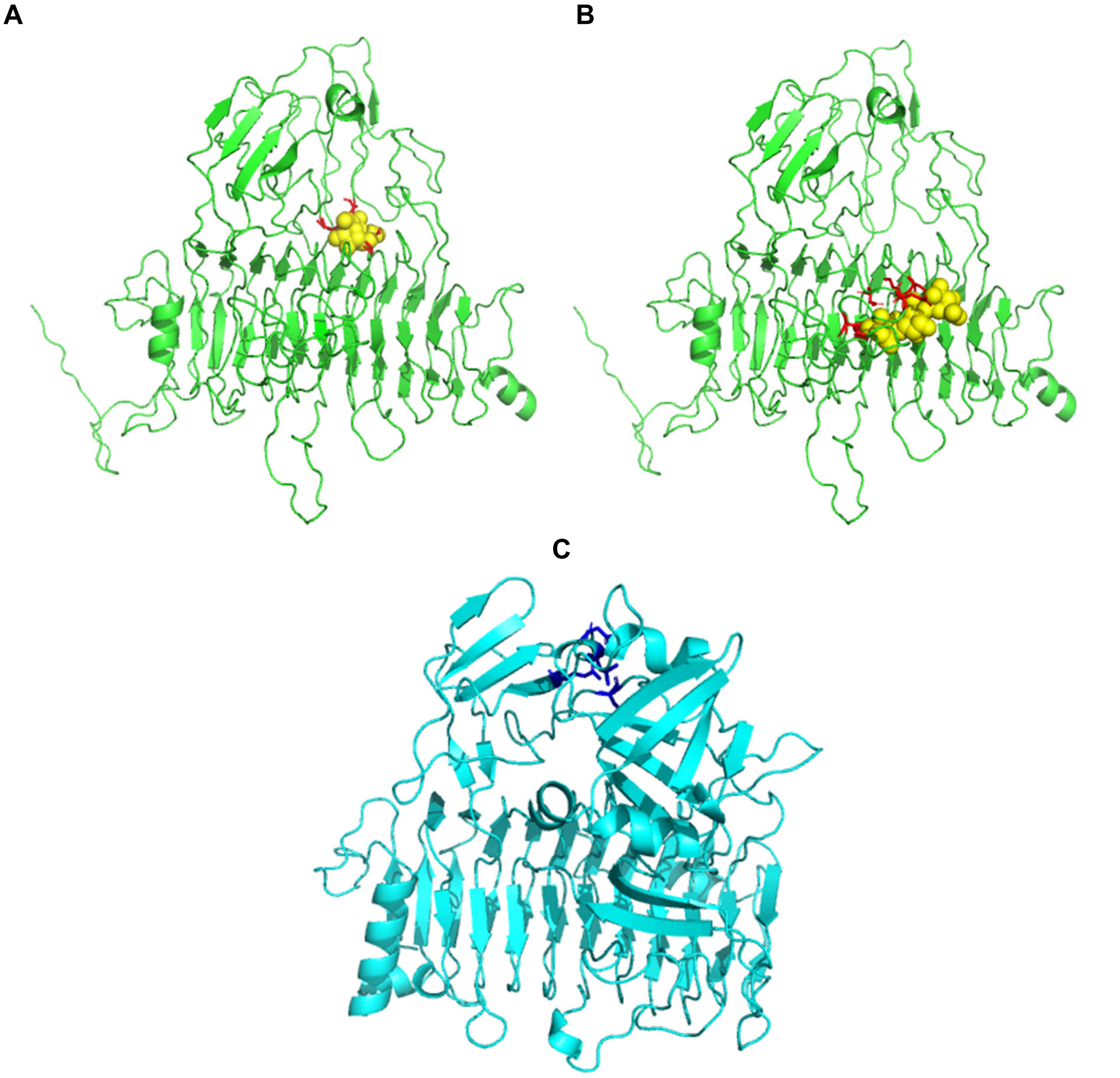


Figure 3.17. Comparison of FunFOLD3 ligand-binding site predictions for CASP12 target T0912 (PDB ID 5mqp)
(**A**) Predicted ligand binding site residues shown as sticks with incorrect predictions in *red* and the predicted fructose (FRU) ligand shown as sphere and coloured yellow. MCC score of -0.00672 and BDT score of 0.0295 was achieved **(B)** Predicted ligand binding site residues shown as sticks with incorrect predictions in *red* and the predicted alpha-D-mannopyranuronic acid (MAV) ligand shown as sphere and coloured *yellow.* MCC score of -0.00892 and BDT score of 0.0213 was achieved **(C)** The observed ligand binding site residues shown as sticks for T0912 (PDB ID 5mqp), with binding site residues coloured in *blue.* The calcium ligand has not been illustrated as it is not part of the sequence as shown in PyMOL

The ninth predicted CASP12 target is glycoside hydrolase (CASP T0912 and PDB ID 5mqp) as can be seen from Figure 3.17, there were no correct predictions for this target and FunFOLD3 had predicted two ligands (fructose and MAV) whereas, as illustrated in Figure 3.17C only one ligand is associated with glycoside hydrolase as confirmed with PDB (Laver, Lenz and Dulhunty, 2001).

Glycoside hydrolase are a group of enzymes responsible for the hydrolysis of glycosidic bonds in carbohydrates (Yang, Roy and Zhang, 2013). One of the predicted ligands as per FunFOLD3 was fructose and a rationale for why FunFOLD3 predicted this ligand could be confusion with other hydrolase such as fructan 1-exohydrolase and thiocyanate hydrolase, which may have some structural homology to glycoside hydrolase (Czjzek & Michel, 2017).

It is somewhat surprising, that with an enzyme such as glycoside hydrolase, only one ligand is observed and it would be a metal ion at that. A literature search on the role of calcium has identified that this metal ion has a catalytic role in GH97 inverting glycoside hydrolase (Yang, Roy and Zhang, 2013). Several glycoside hydrolases require the participation of a metal ion for catalysis, and in particular a divalent metal ion such as zinc or calcium (Okuyama *et al.*, 2014) and is potentially the case with this protein. In hydrolase enzymes, such as Bt GH97a (inverting α-glucoside hydrolase) and Bt GH97b (retaining α-glucoside hydrolase) both have one calcium ion in the active site, which plays an important role in the catalysis of both enzymes. Without further information on the role of this protein, it is difficult to say the exact function of the calcium ion.

A TM-score of 0.633 was obtained with 624 residues for the observed protein and 599 for the predicted protein, as Figure 3.18 shows, the structural similarity is contained within some of the molecule. The actual protein contains numerous β sheets in a portion of the protein and the predicted protein model has failed to form these β sheets in the correct region.

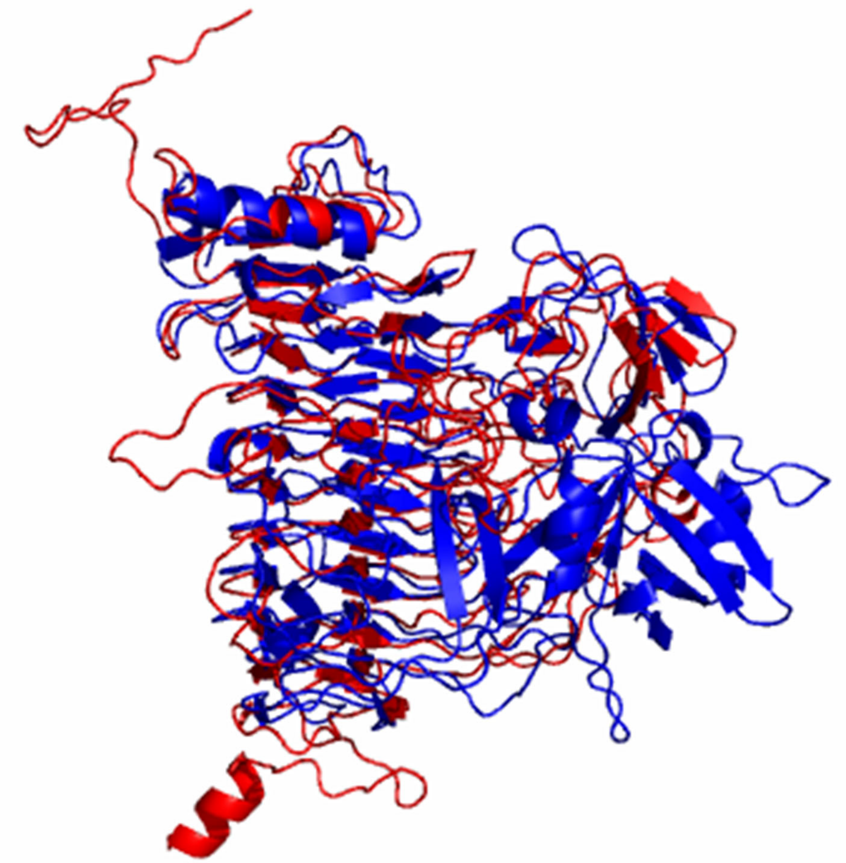


Figure 3.18. Comparison of TM-align (Zhang & Skolnick, 2005) structures for CASP12 target T0912 (PDB ID 5mqp)
The structure in blue is the observed structure for T0912 and the structure in red is the predicted structure from IntFOLD4. A TM-score of 0.633 was achieved. The score was normalised for the observed structure as it is the reference molecule.

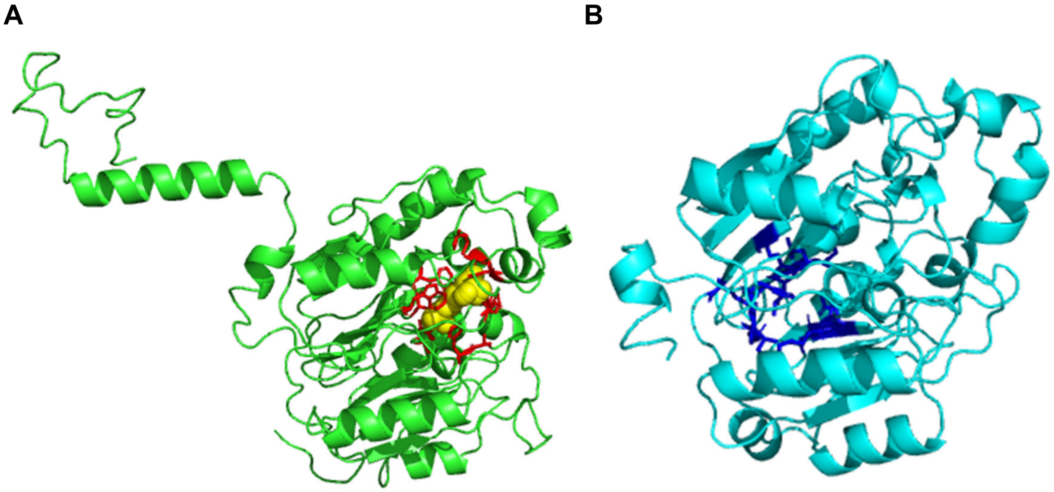


Figure 3.19. FunFOLD3 ligand-binding predictions for CASP12 target T0913
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in *blue* and *under*- and *over* predictions in *red*, the predicted ligand phosphate citruline shown as sphere and coloured *red.* **(B)** The observed ligand binding site residues shown as sticks for T0913 with binding site residues coloured in *blue* no ligand was identified on PDB

The tenth predicted CASP12 target is F4ZCI1 and the information obtained from UniProtKB identifies this target as protein 2-nitroimidazole nitrohydrolase. The annotation level on UniProtKB is three out of a possible level five annotation, meaning protein inferred from homology indicating that the existence of this protein is probable because clear othologs exist in closely related species (UniProt Consortium, 2019). This evidence is supported by the ability of IntFOLD4 to predict a structure, as the ability to predict a structure relies on evidence from known proteins.

As can be seen from Figure 3.19, a ligand in the form of citrulline was predicted. Citrulline or L-Citrulline is a ubiquitous, naturally occurring nonessential amino acid and about 80% of citrulline is converted in the kidney to arginine (ARG) (Okuyama *et al.*, 2014). A literature search of l-citrulline identified a therapeutic role of l-citrulline in improving erectile hardness in patients with mild erectile dysfunction (Kaore, Shilpa N. Kaore, 2014). Therefore the identified ligand could be biologically relevant. Based on this information, it is possible to gain more insights into the protein, a literature search identified argininosuccinate synthase as an enzyme that catalases the synthesis of argininosuccinate from citrulline and aspartate (Cormio *et al.*, 2011). Argininosuccinate synthase has the following EC number associated

6.3.4.5 indicating the enzyme belongs to the group related to other carbon-nitrogen ligases and specifically the argininosuccina synthase family. As this target is an enzyme, FunFOLD3 had predicted EC terms which were; 2.1.4.1 (glycine amindino transferase), 3.5.3.6 (arginine deiminase) and 3.5.3.18 (dimethylargininase). Although the EC numbers do not match exactly there is a similarity with enzyme argininosuccinate synthase, such as acting on carbon-nitrogen bonds (EC 3.5.).

Figure 3.20 below, shows the TMalign superposition of predicted structure and the observed structure. The TM-score was 0.81367 showing good structural homology.

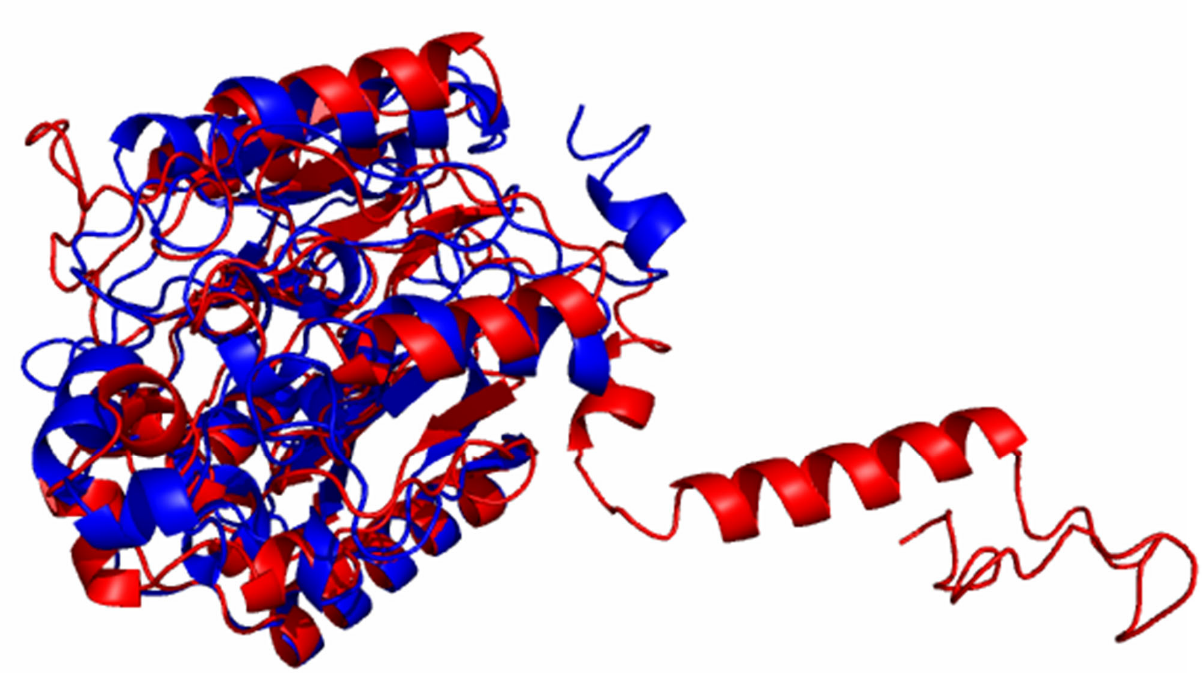


Figure 3.20. Comparison of TM-align (Zhang & Skolnick, 2005) structures for CASP12 target T0913
The structure in blue is the observed structure for T0913 and the structure in red is the predicted structure from IntFOLD4. A TM-score of 0.81367 was achieved. The score was normalised for the observed structure as it is the reference molecule.

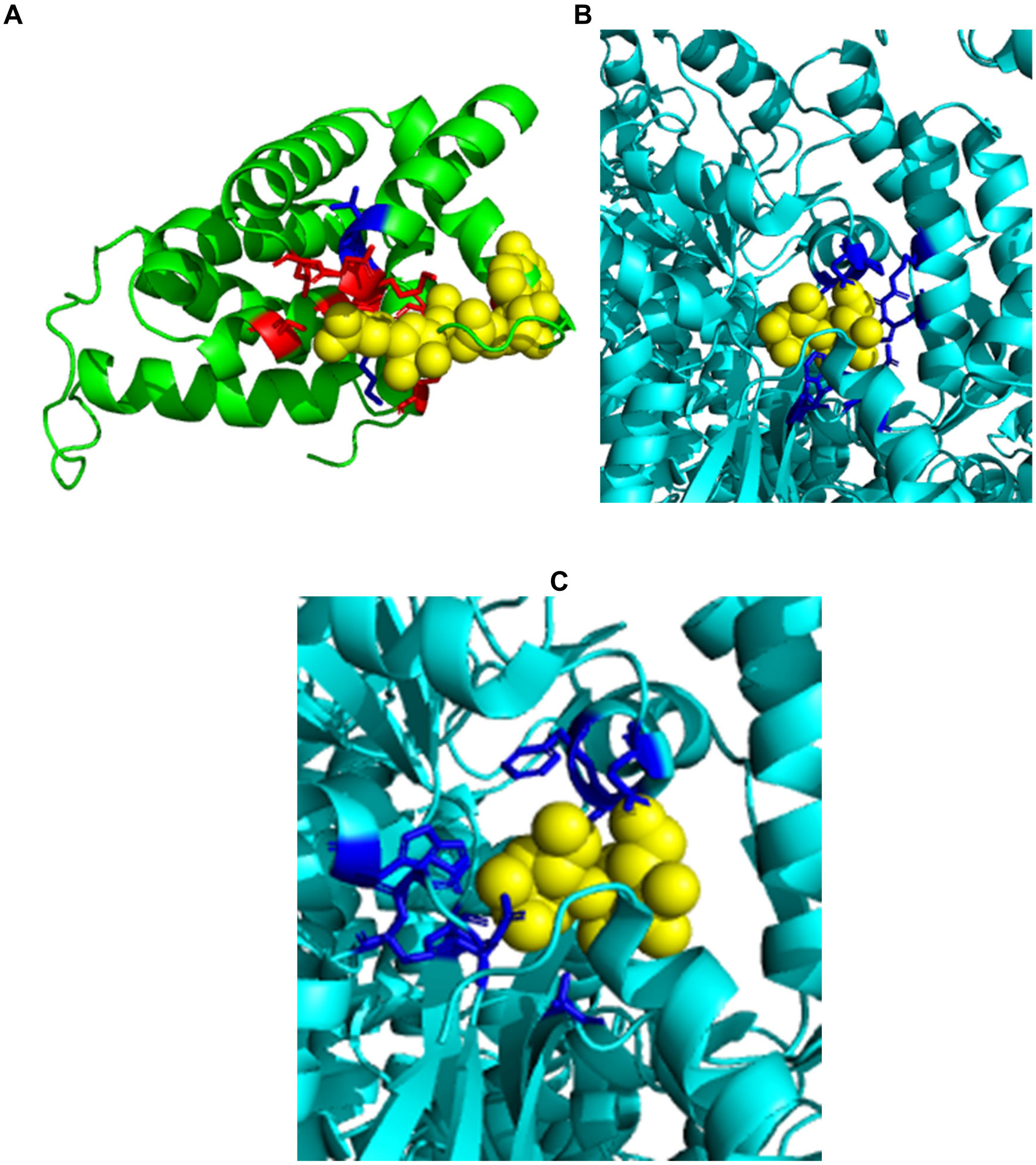


Figure 3.21. Comparison of FunFOLD3 ligand-binding site predications for CASP12 target T0916 (PDB ID 5tj4) **(A)** Predicted ligand binding site residues shown as sticks and coloured *red* and the predicted ligand NAD shown as sphere and coloured *yellow.* An MCC and BDT score of 0.263 and 0.370, respectively was achieved, compared to GLC(2) **(B)** and **(C)** The observed ligand binding site residues shown as sticks with binding site residues coloured in *blue* the alpha-D-glucopyranose (GLC) ligand is shown as sphere. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target. Given the location of the ligand in the observed structure it had to be shown in two different images for ease .

The eleventh predicted CASP12 target is Gasdermin B C-terminal domain (T0916 and PDB ID 5tj4) as can be seen from Figure 3.21A predicted ligand-binding site residues were obtained. The PDB entry classifies the protein as ligand binding protein. Additionally, the GLC ligand was predicted in two different locations within the observed structure. The observed structure consists of 10 chains, hence why only a portion of the observed structure has been illustrated.

The GLC ligand is not identified in the PDB entry, only sodium is predicted and this could be due to the use of sodium containing solvents in the crystallisation of the protein structure (Chao, Kulakova and Herzberg, 2017). In comparison NAD was predicted for the T0916 target and due to similarities in the ligand-binding site with one of the GLC ligands, an MCC and BDT score was obtained which was 0.263 and 0.370, respectively.

The CASP12 organisers cancelled the observed structure, so the prediction was based off the PDB file. The structure was cancelled as an apparent template (5b5r) appeared before the server deadline and was re-released as TBM target T0948. Of note, this template was not identified by FunFOLD3 to contain ligands nor was it one of the templates that was predicted in terms of structure similarity.

Gasdermin B is involved in pyroptosis, an inflammatory form of programmed cell death, which is critical for amplifying protective immunity during infection (Haines, Pendleton and Eichler, 2011). The exact mechanism of pyroptosis did not become clear until 2015 when researchers identified cleavage of a protein within this family (Gasdermin D) generates a 31kDa N-terminal fragment and a 22kDa C-terminal fragment. The role of the N-terminal has been identified alone induced pyroptosis when expressed ectopically (Jorgensen and Miao, 2015). Hergueta-Redondo and colleagues (Hergueta-Redondo *et al.*, 2014) identified gasdermin-b as a promoter in the invasion and metastasis in breast cancer cells. At this stage, it was known that all the four human proteins of this family (Gasdermin A, Gasdermin B, Gasdermin C and Gasdermin D) contain several conserved sequences in the N- and C-terminal regions, but to date no functional domains or motifs had been described (Shi *et al.*, 2015). Since the publication of this study in 2014, no further studies have been published into the function of Gasdermin B. Table 3.7, below shows the GO terms predicted for CASP12 target T0916 (PDB ID 5tj4).

Table 3.7. Predicted GO terms for CASP12 target T0916 (PDB ID 5tj4)
The predicted GO terms for CASP12 target T0916 (PDB ID 5tj4) and their associated term domains and function are shown below. Biological process is coloured red and molecular function coloured green

GO term	GO term domain	Function
GO:0000166	Molecular function	nucleotide binding
GO:0008106	Molecular function	alcohol dehydrogenase (NADP+) activity
GO:0004022	Molecular function	alcohol dehydrogenase (NAD) activity
GO:0016491	Molecular function	oxidoreductase activity
GO:0046872	Molecular function	metal ion binding
GO:0008912	Molecular function	acetaldehyde reductase activity
GO:0008198	Molecular function	ferrous iron binding
GO:0055114	Biological process	oxidation-reduction process
GO:0005975	Biological process	carbohydrate metabolic process
GO:0006004	Biological process	fucose metabolic process
GO:0019317	Biological process	fucose catabolic process
GO:0019301	Biological process	rhamnose catabolic process
GO: 0042355	Biological process	L-fucose catabolic process
GO:0042846	Biological process	glycol catabolic process
GO:0051143	Biological process	propanediol metabolic process

The GO terms listed above do not currently provide any additional insight into what has already been published, the GO terms mainly support the prediction of the ligands as per FunFOLD3, unsurprisingly.

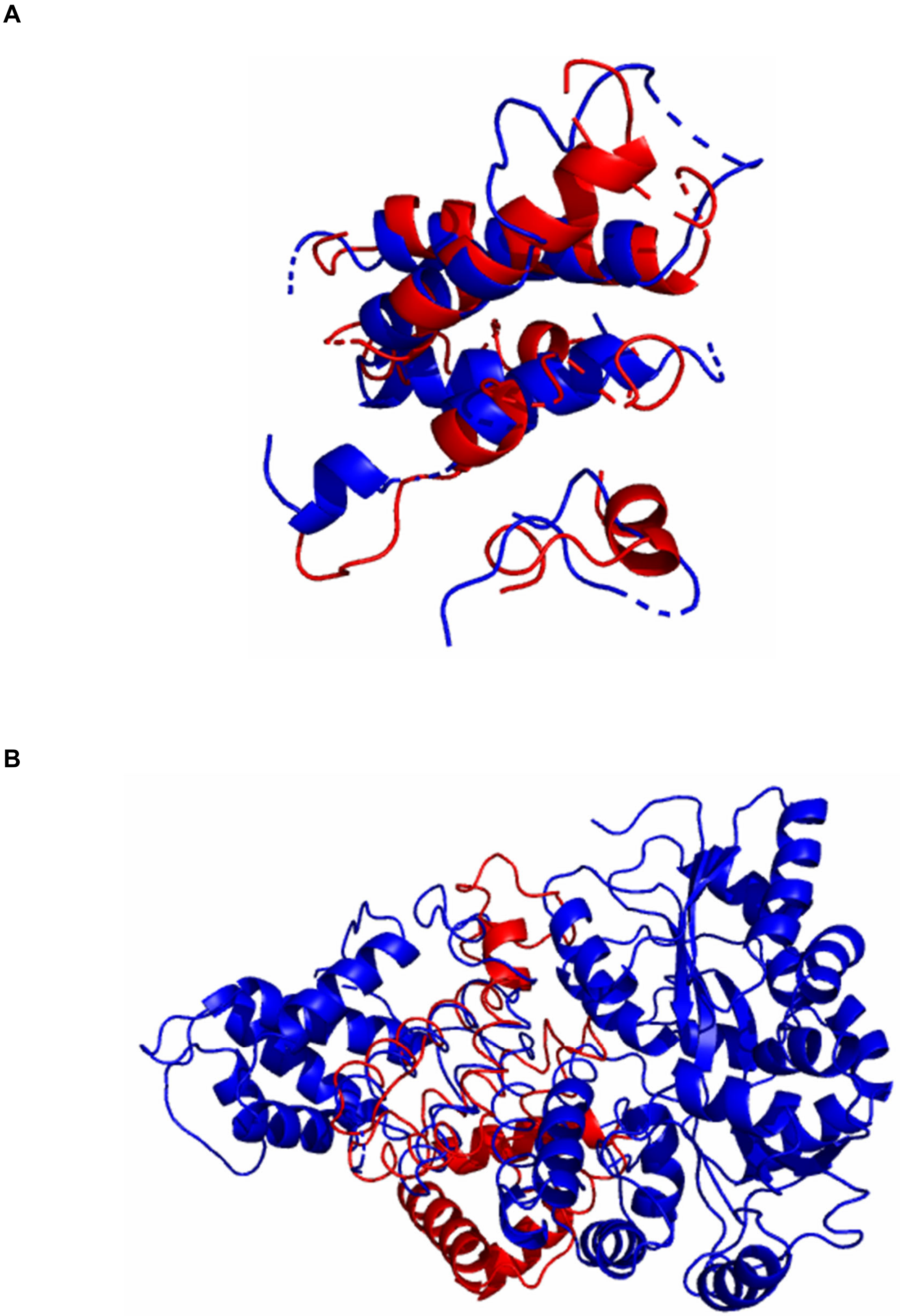


Figure 3.22. Comparison of TM-align (Zhang & Skolnick, 2005) structures for CASP12 target T0916 (PDB ID 5tj4)
(A) The structure in blue is the observed structure from PDB and the structure in red is the predicted structure from IntFOLD4. A TM-score of 0.14930 was achieved. The score was normalised for the PDB entry as it is the reference molecule and shows poor structure homology potentially. This superposition is of the aligned regions only whereas **(B)** is for all the domains of the observed structure compared to the domain released by the CASP organisers

Table 3.8 below shows the comparison between ligands predicted by FunFOLD3 and ligands associated with the targets as per the PDB entry.

Table 3.8. Comparison of ligands predicted using FunFOLD3 and ligands identified on Protein Databank
*ligand has been identified by FunFOLD3 following analysis on observed structure and is not a ligand on PDB (Quinten and Kuhn, 2012)

CASP 12 target	FunFOLD3 ligand	PDB ligand
T0868 (PDB ID 5j4a)	BCG	No ligands
T0872 (PDB ID 5jmb)	Calcium ion	No ligands
T0899	Magnesium ion	N/A
T0901	Magnesium ion	N/A
T0905	GLY	N/A
T0907	Calcium ion	N/A
T0909 (PDB ID 5g5n)	Carbonate ion (CO3)	Methyl mercury ion (MMC), glycerol (GOL) and chloride ion
T0911 (PDB ID 6e9n)	Dibromotyrosine (DBY)	Gluconic acid*
T0912 (PDB ID 5mqp)	Fructose (FRU) and alpha-D-mannopyranuronic acid (MAV)	Calcium ion
T0913	Citrulline (CIR)	N/A
T0916 (PDB ID 5tj4)	Nicotinamide-adenine-dinucleotide (NAD)	Maltose (MAL) and sodium ion
T0919	Beta-d-glucose (BGC)	N/A

3.3.4 Analysis of CASP13 Functional Prediction

As with previous CASP competitions, targets with biologically relevant and predicted ligand-binding site residues will be analysed. Following on from the analysis in CASP12, there will be additional focus on the middle/centroid ligand of the protein.

A total of 183 CASP13 targets were released for analysis, this consisted of 90 regular targets; 13 hetero-multimer targets; 31 refinement targets and 49 assisted modelling prediction targets. Analysis using the FunFOLD3 server identified 34 targets containing biologically relevant ligands and binding site residues. These were T0949, T0953s2 (PDB ID 6f45), T0954 (PDB ID 6cvz), T0955 (PDB ID 5w9f), T0957s2 (PDB ID 6cp8), T0958 (PDB ID 6btc), T0965 (PDB ID 6d2v), T0961, T0970 (PDB ID 6g57), T0972, T0973, T0974s1, T0975, T0980s1 (PDB ID 6gnx), T0980s2 (PDB ID 6gnx), T0983, T0985, T0986s2 (PDB ID 6d7y), T0992 (PDB ID 6xbd), T0993s1, T0994, T0995, T0997, T1001, T1003 (PDB ID 6hrh), T1008 (PDB ID 6msp), T1009 (PDB ID 6dru), T1012, T1013, T1014 (PDB ID 6qrj), T1016 (PDB 6e4b), T1017s1, T1018 (PDB ID 6n91) and T1023s2. CASP13 target IDs are given outside parenthesis and PDB IDs, where applicable within parenthesis. Protein-ligand interactions were predicted for all 34 of the FN targets.

For the predicted CASP targets which have PDB IDs associated and have an actual structure released by CASP organisers will have a MCC score and BDT score will be calculated. Predictions without an associated PDB ID and/or actual structure cannot be objectively assessed, however the results will still be included as part of the analysis to demonstrate the results from FunFOLD3 server.

Corresponding observed binding site residues are also provided, along with *under* and *over*-predictions. Correct residues are highlighted in red as illustrated in Table 3.9. Three top scoring predictions are presented (T0974s1, T0961 and T0983), along with two lower scoring predictions (T0953s2 and T1003). For the remaining CASP13 predictions, please refer to Appendix 2.

Table 3.9. Predicted and observed ligand-binding site residues for CASP13 targets
Correct ligand binding site residues are depicted in red and bold and presented in ascending CASP13 target ID. For CASP13 targets where there is no ligand predicted in the experimental structure the observed ligand-binding site residue column is blank

CASP13 target	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0949	CU ligand: 85,159,161,171 OXY ligand: 95,98,99,160	No structure released		
T0953s2 **(PDB ID 6f45)**	119,120,121,122,124,125,126,127,155,156,157,158,159,**164**,165,166,167,168,169,170,174,195,198	54,164	54	119,120,121,122,124,125,126,127,155,156,157,158,159,165,166,167,168,169,170,174,195,198
T0954 **(PDB ID 6cvz)**	LYS ligand: 77,119,273,274,275 DT ligand: 116,161	123,124,129,130,131	123,124,129,130,131	LYS ligand: 77,119,273,274,275 DT ligand: 116,161
T0955 **(PDB ID 5w9f)**	ZN ligand: 27 RG and RA ligand: 39	No biologically relevant ligands found		
T0957s2 **(PDB ID 6cp8)**	LEU ligand: 116,164 ALA ligand: 102 SF4 ligand: 40	No biologically relevant ligands found		
T0958 **(PDB ID 6btc)**	DC ligand: 47,48,51,73,74,75	No biologically relevant ligands found		

	SAH ligand: 32,36,38			
T0961 **(PDB ID 6sd8)**	**171**, 172, **173**, **174**, **178, 179, 180,** 209, **210, 211, 212, 260, 324, 326, 327, 331, 334**, 336, **402, 405, 406, 409, 428**, 429, **431, 434, 437**	171,173,174,178,179,180,210,211,212,260,268, 324,326,327,331,334,402,403,405,406,409,424, 428,431,433,434,437*	268,403,424,433	172,209,336,429
T0965 **(PDB ID 6d2v)**	30, 32, **33**, **34**, **35,** 53, **54**, **56**, 58, 75, 76, **77**, 97, 98, 99, 101, 103, 134, 135, 136, 165, 192, 193, 194, 195, 201, 204, 219, 220, 221, 226,264,286, 291	CL ligand(1): 81,144 CL ligand(2): 39,41,42 NDP 10,12,13,14,15,33,34,35,54,55,56,57,77,78,79,81, 114,115,116,145,149,172,173,174,175*	10,12,13,14,15, 57,78,79,81, 114,115,116,145,149,172,173,174,175	30,32,53,58, 75, 76,97,98 99, 101,103,134,135,136,165, 192,193,194,195 201,204, 219,220,221,226,264,286, 291
T0970 **(PDB ID 6g57)**	0FX ligand: 53,88,89 ZN ligand: 21,49,79	No biologically relevant ligands found		
T0972	DC ligand: 1,82,83,85,86,87 HEM ligand: 17,18,19,20,21,22,23,24,25,30,59,60,92,95	Structure cancelled by organisers		

T0973
(**PDB ID 6yfn**)
	RA ligand(1):
	29,48,50,51,52,73,77103
	
	RA ligand(2):
	57,71
		No structure released		

T0974s1 (**PDB ID 6tri**)	31	31		
T0975	DU ligand: 78,81,83,84,170,171 DT ligand: 174,175 SF4 ligand: 62,64,311,312,313,316,322	No structure released		

T0980s1 **(PDB ID 6gnx)**	PRO ligand: 68,69 SER ligand: 13,72,73,75,76,78,79,81,83	No biologically relevant ligands found*		
T0980s2 **(PDB ID 6gnx)**	SAH ligand: 1,8,36,37,38,39,40,41	No biologically relevant ligands found*		
T0983 **(PDB ID 6uk5)**	SAH ligand: **2**,14,17,**21**,**46,47,48**,**52**,**67,68,69,72,88,89,** **90**,**105**,**106**,107,**108**,**111**,141,147,150,178, 228	SAM ligand: 2,10,21,46,47,48,52,66,67,68,69,72,88,89,90,91,105,106,108,111,112*	10,66,91,112	14,17,107,141,147,150,178, 228

T0985	BGC ligand(1):
	386,401,402,437,450,532,534,685,695, 700, 702,745,746,764,765
	
	BGC ligand(2):
	208,210
		No structure released		

T0986s2 **(PDB ID 6d7y)**	57,60,61,62,64	No biologically relevant ligands found*		
T0992	15,16,17	No structure released		
T0993s1 **(PDB ID 6xbd)**	18,21,23,43,44,45,46,47,48,49	No biologically relevant ligands found*		
T0994	CAZ ligand: 390,391,423,426,439,478,528,529,530, 531, 533,538 HIS ligand: 446,450	Structure cancelled by organisers		
T0995	48,54,130,137,164,189,192	No structure released		
T0997	179,180,184,188,199, 200,201,202	No structure released		
T1001	32,34,45,60,61,63, 74,79,80,81,82,83,84, 85,87,88,92,97,113	No structure released		
T1003 **(PDB ID 6hrh)**	113,144,145,146,149,172,244,246, 247,275, 278, 284, 306, 308, 474			

		257,258,259,262,285,287,328,332,357,359,360, 388,391,419,420,421		

T1008
(PDB ID 6msp)
	FMN ligand(1):
	27,34
	
	FMN ligand(2):
	36,69
		No biologically relevant ligands found*		

T1009
(PDB ID 6dru)
	GLC ligand(1):
	682,684
	
	GLC ligand(2):
	257,**286**,**325**,**393**,**395**,396,**470**,**484**,**487**,**520**,
	557
		MAN ligand(1):
		84,100,108,110,113,114,115,116
		
		MAN ligand(2):
		261,300,303
		
		MAN ligand(3):
		493,523,525
		
		MAN ligand(4):
		564,565,568
		
		BMA ligand(1):
		260,301
		
		BMA ligand(2):
		565,566
		
		GAL ligand:
		411
		
		XYS ligand:
		257,285,286,325,393,395,470,484,487,520,557
		
		BGC ligand(1):
		286,357,396,404,406,409
		
		BGC ligand(2):
		527,534
		
		BGC ligand(3):
		257,520,559
			MAN ligand(1):
			84,100,108,110,113,
			114,115,116
			
			MAN ligand(2):
			261,300,303
			
			MAN ligand(3):
			493,523,525
			
			MAN ligand(4):
			564,565,568
			
			BMA ligand(1):
			260,301
			
			BMA ligand(2):
			565,566
			
			GAL ligand:
			411
			
			BGC ligand(1):
			286,357,
			404,406,409
			
			BGC ligand(2):
			527,534
			
			BGC ligand(3):
			559
			
			XYS ligand:
				GLC ligand(1):
				682,684
				
				GLC ligand(2);
				396

			285	
T1012	27,30,125,126,127,128,129,130,135,136,137,138,139,140,141,160,161,163,166,167,170,171,173	Structure cancelled by organisers		
T1013	CLR ligand: 449,453,462,468,471 MPG ligand: 50 RET ligand: 91,98,107,118,121,125,216,447,450,451,454,476,480 LEU ligand: 183,184,185	No structure released		
T1014 **(PDB ID 6qrj)**	57,58,60,61,85,88,90,98,103,104,105,114,115,116,117,118,119,120,121,144,146	MG: 139,142 ANP: 139,142,143,167,172,180,185,186,187,199 200,201,202,228	139,142,143,167,172,180,185,186,187,199 200,201,202,228	57,58,60,61,85,88,90,98,103,104,105,114,115,116,117,118,119,120,121,144,146
T1016 **(PDB ID 6e4b)**	**7**,**8**,**14**,19,20,21,**57**,**81**,84,**149**,**150**	7,8,14,57,81,149,150		19,20,21,84
T1017s1	30,31,32,33	No structure released		
T1018 **(PDB ID 6n91)**	**14**,56,59,98,170,**197**,**278**,279	PO4:* 124,162 SO4:* 135,162 ZN:* 12,14,197.278	12	56,59,98,170,279
T1023s3	LEU ligand: 192,200,223,230,231	Structure cancelled by organisers		

🟥	GDP ligand: 49,50,51,52,53,54,55,56,78,137,190,191,193,194,225,226,227 RC ligand: 68,262,278,322,324,338 ZN ligand: 109,110

*no structure released by CASP13 organisers. Predictions are made against the PDB file

A comparison of the MCC, BDT and TM-scores for the targets are shown in Figure 3.4. As can be seen from the figure , the targets with the lowest TM-scores (T0953s2 and T1014) also has the lower MCC and BDT scores.

Table 3.10 below lists the MCC and BDT scores for each of the CASP13 targets from ascending to descending order by MCC and BDT score. A total of 10 MCC and BDT scores could be obtained due to having the same ligands predicted or if not, similar ligand-binding site residues thereby inferring a degree of similarity.

Table 3.10. MCC and BDT scores for CASP13 targets
MCC and BDT scores for CASP13 targets from ascending to descending order of MCC and BDT score

CASP13 target	MCC Score	BDT Score
T0974s1	1.0	1.0
T1009	0.91	0.94
T0961	0.843	0.903
T0983	0.715	0.715
T1016	0.556	0.646
T1018	0.522	0.48
T0965	0.12	0.35
T0953s2	0.11	0.12
T1003	-0.04	0.06
T1014	-0.05	0.05

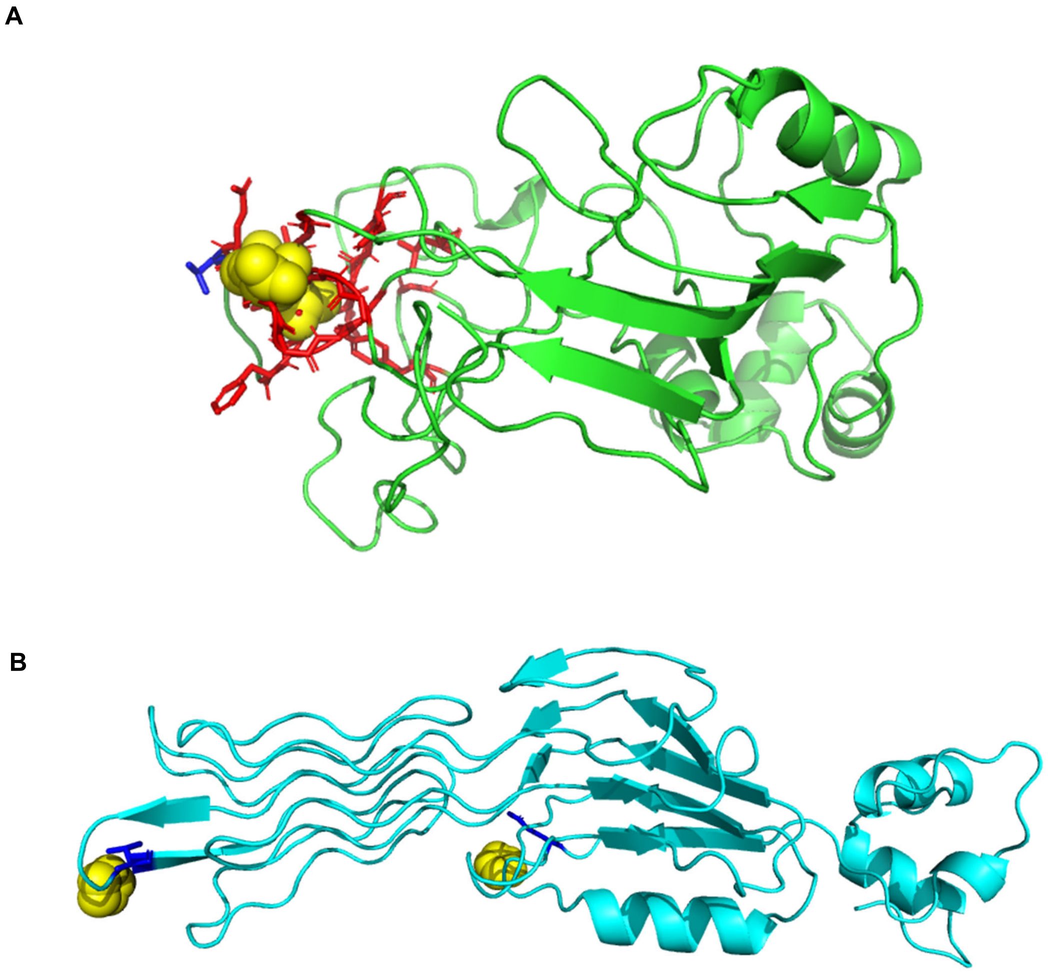


Figure 3.23. FunFOLD3 ligand-binding site predictions for CASP13 target T0953s2 (PDB ID 6f45)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand 3-O-acetyl-2-acetamido-2-deoxy-alpha-D-galactopyranuronic acid (DJB) shown as sphere and coloured yellow. BDT score of 0.11 and MCC score of 0.12 was achieved **(B)** The observed ligand binding site residues shown as sticks for T0953s2 with binding site residues coloured in blue and the correctly predicted ligand imidazole (IMD) shown as sphere and coloured yellow

The second predicted CASP13 target and the first which had an observed structure for comparison was adhesion tip from organism Salmonella phage vB_SenMS16. As Figure 3.23 shows, only one correct binding site was predicted. FunFOLD3 predicted DJB as the ligand, whereas the biologically relevant ligand identified in the observed structure was IMD and is at two locations within the structure. Adhesion tip is classified as a viral protein as part of the PDB entry. Cell adhesion is a central mechanism that drives the development of multicellular organisms and cells use adhesion to move, communicate and differentiate (Mateo *et al.*, 2015). Adhesion tips are of particular importance in phages, as this is how they recognise bacterial hosts to infect. This has led to their exploitation as bio-tools for bacterial remediation and detection (Dunne *et al.*, 2018).

The observed ligand in the predicted ligand is imidazole and imidazole nucleus is found in several categories of therapeutic agents such as anti-microbials, anti-virals and anti-cancer agents (Shalmali, Ali and Bawa, 2018). In contrast, DJB was the predicted ligand by FunFOLD3 for the predicted structure and despite the differences in the predicted ligands, an MCC and BDT score 0.12 and 0.11 was achieved, respectively for this target based on the one correct prediction.

A TM-score of 0.37895 was obtained and this is deemed as random structural similarity and this could explain the poor ligand-binding site prediction, as the overall structure was not predicted sufficiently

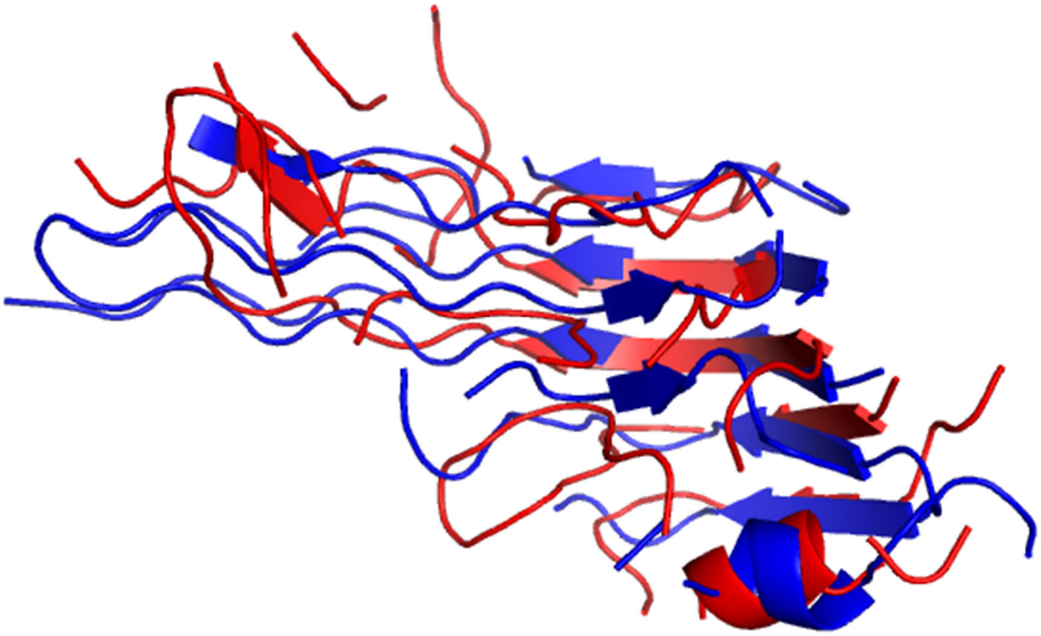


Figure 3.24. Comparison of TMalign (Zhang & Skolnick, 2005) structures for CASP13 target T0953s2 (PDB ID 6f45)
The structure in blue is the observed structure for 6f45 and the predicted structure for CASP13 target T0953s in red. A TM-score of 0.37895 was achieved for protein structures. The score was normalised for PDB ID 6f45 as it is the reference molecule

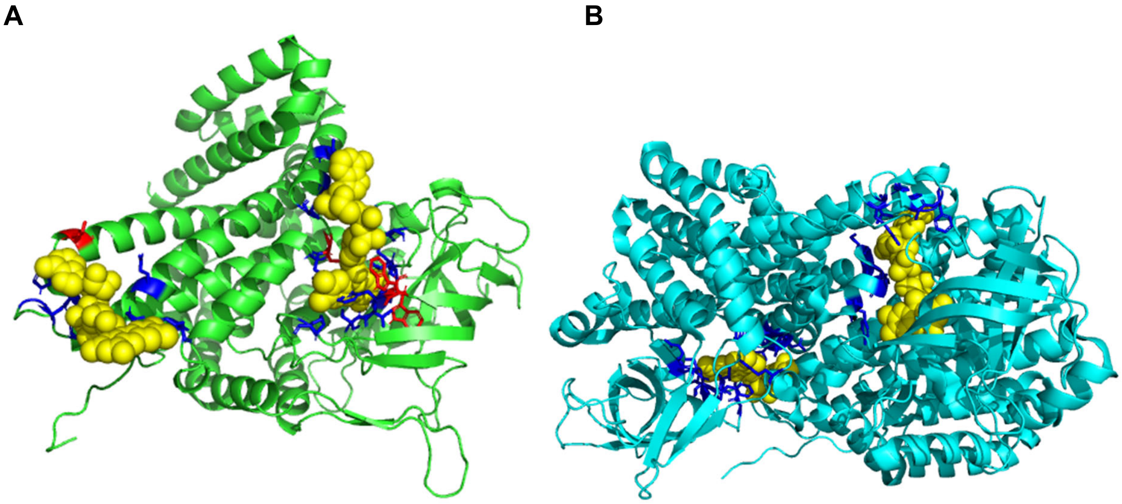


Figure 3.25. FunFOLD3 ligand-binding site predictions for CASP13 target T0961 (PDB ID 6sd8) **(A)** Predicted ligand binding site residues shown as sticks with correct predictions in blue and incorrect predictions in red, the predicted flavin-adenine dinucleotide (FAD) ligand shown as sphere. A BDT score of 0.903 and an MCC score of 0.843 was achieved **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in blue the FAD ligand is shown as sphere. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target

The seventh predicted CASP13 target is Q6MJ59 and is classified as oxidoreductase as per the PDB entry. No observed structure was released by the CASP organisers therefore, no comparisons can be made between the predicted and observed structure. However, the PDB entry identifies FAD as a ligand which demonstrates that FunFOLD3 has potentially correctly identified the ligand. Furthermore, data from the CASP13 competition identifies FAD as a cofactor with and without a C10 length acyl-CoA thioester ligand bound (Lepore *et al.*, 2019). Additionally, a single FAD molecule binds per monomer in a crevice located at the dimer interface. The C10-CoA ligand binds into a long narrow tunnel that runs deep into the protein beneath the bound FAD molecule which is similar to other ACAD structures (Lepore *et al.*, 2019).

Despite CASP organisers not releasing an observed structure, there is a PDB entry associated with the CASP target. In order to aid with analysis of the target, FunFOLD3 was analysed against the PDB structure. A high BDT score of 0.903 and MCC score of 0.843 was achieved for the target.

The TM-score for the predicted structure is 0.96964 showing very good structural alignment between the predicted and PDB protein structure. The TM-align superpositon of observed and predicted structures is shown in Figure 2.26.

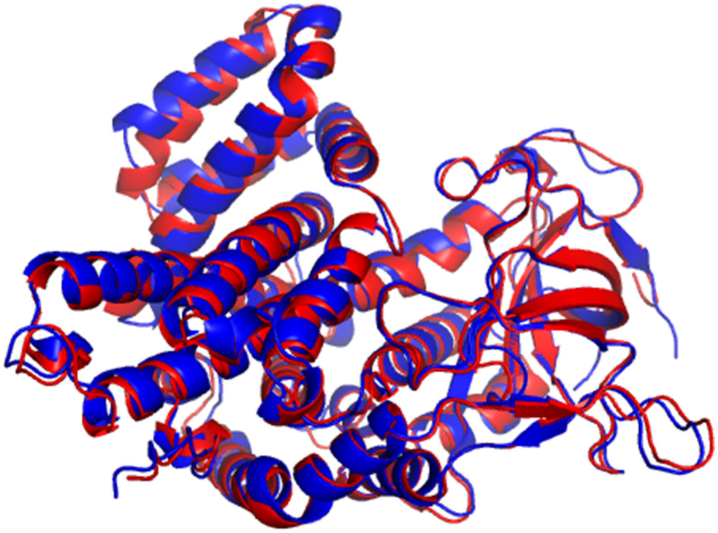


Figure 3.26. Comparison of TMalign (Zhang & Skolnick, 2005) structures for PDB ID 6sd8
The structure in blue is the observed structure for PDB ID 6sd8 and the predicted structure for CASP13 target T0961 in red. A TM-score of 0.96964 was achieved for protein structures. The score was normalised for PDB ID 6sd8 as it is the reference molecule

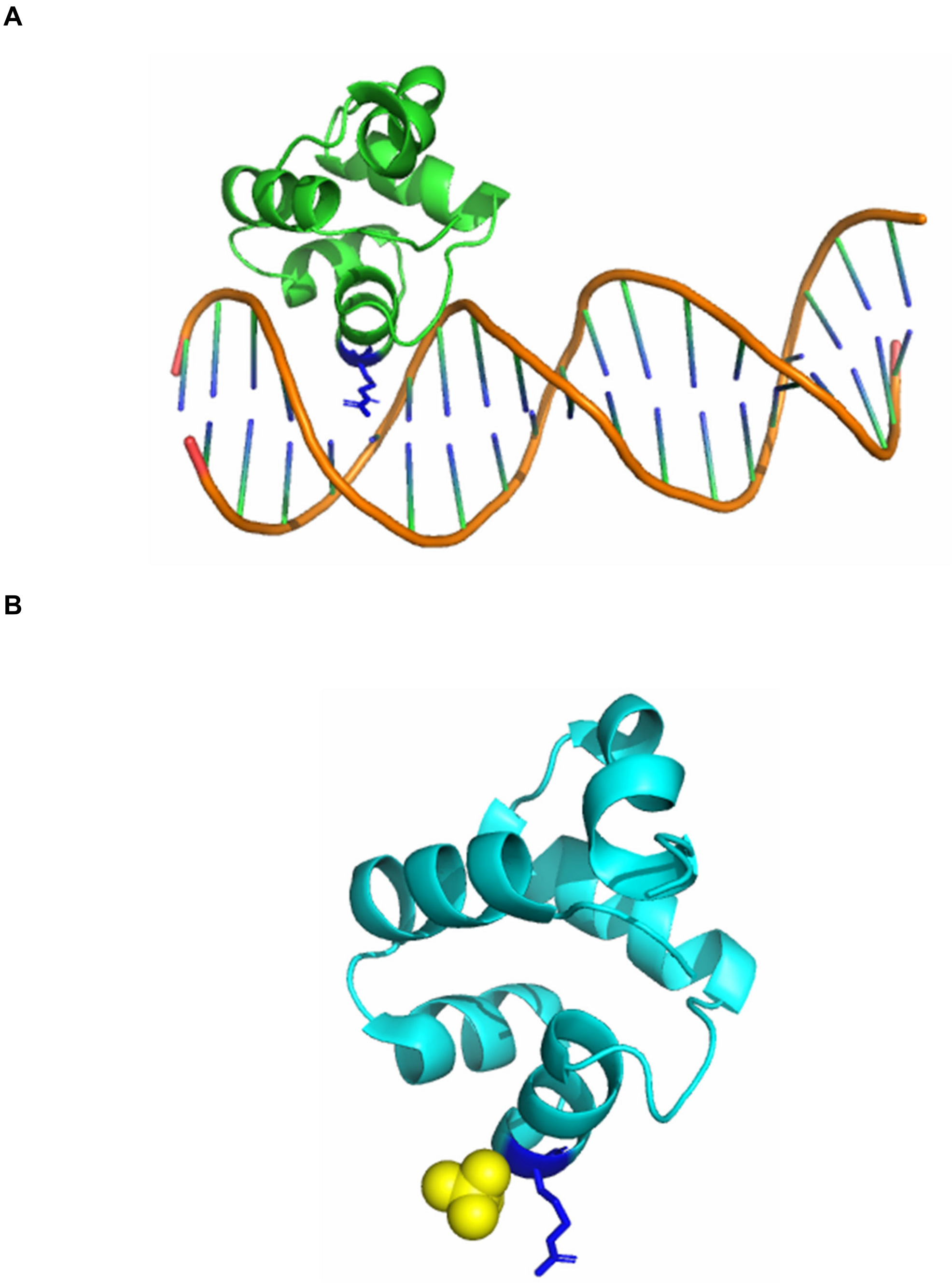


Figure 3.27. FunFOLD3 ligand-binding site predictions for CASP13 target T0974s1 (PDB ID 6tri)
(A) Predicted ligand binding site residues shown as sticks with correct prediction in blue the predicted DNA ligand shown as double helix. A BDT score of 1.0 and an MCC score of 1.0 was achieved, respectively **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in **blue** the sulphate (SO4) ligand is shown as sphere and coloured yellow. The actual structure is a dimer, whereas the predicted structure is a homodimer. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target. Only one chain in the observed structure was predicted, in line with the predicted structure and the section of the structure released by CASP13 organisers

The twelfth predicted CASP13 is O48503/O48504 and the PDB entry classifies the target as a viral protein from organism Lactococcus phage TP901-1 and is a temperate bacteriophage. No observed structure was released by CASP13 organisers, therefore analysis is made using the structure and information from the PDB database. The PDB entry identifies SO4 as a ligand, whereas in comparison FunFOLD3 predicted DNA as a ligand.

Literature information on temperature bacteriophages show that they can enter one of two life cycles following infection of a host; the lysogenic or lytic life cycles (Rasmussen *et al.*, 2020). The choice between the two life cycles is dependent upon regulation of promoters and their cognate regulatory proteins within the phage genome. The genetic switch is controlled by the CI repressor and the modulator of repression (MOR) antirepressor and their interactions with DNA (Rasmussen *et al.*, 2020). Solved crystal structures of MOR in complex with the N-terminal of CI, reveals the structural basis of MOR inhibition of CI binding to the DNA operator sites (Rasmussen *et al.*, 2020). This information potentially supports the prediction of DNA as a ligand by FunFOLD3. Despite the difference in predicted ligand in comparison to the observed ligand in the pdb file, a perfect BDT and MCC score was obtained of 1.0.

The GO terms predicted by FunFOLD3 are given below in Table 3.11 and the predictions are around DNA binding or specific DNA functions. With respect to templates 1lmb (transcription/DNA), 1y9q (transcription regulator), 3clc (transcription regulator/DNA), 3jxb (transcription regulator) and 3kxa (structural genomics, unknown function). From the predicted templates it is clear why DNA has been predicted as a ligand.

Table 3.11. Predicted GO terms for CASP13 target T0974s1

The GO terms for CASP13 target T0974s1 (PDB ID 6tri and their associated term domains and function are shown below. Bioloigcal process coloured red and molecular function coloured green

GO term	GO term domain	Function
GO:0006355	Biological Process	Regulation of transcription, DNA-templated
GO:0046872	Molecular function	Metal ion binding
GO:0043565	Molecular function	Sequence-specific DNA binding
GO:0003677	Molecular function	DNA binding
GO:0003700	Molecular function	DNA-binding transcription factor activity

Figure 3.28 below shows the TM-align superposition of the predicted model and observed structure from the PDB entry. A TM align score of 0.79965 was achieved, showing good structural homology.

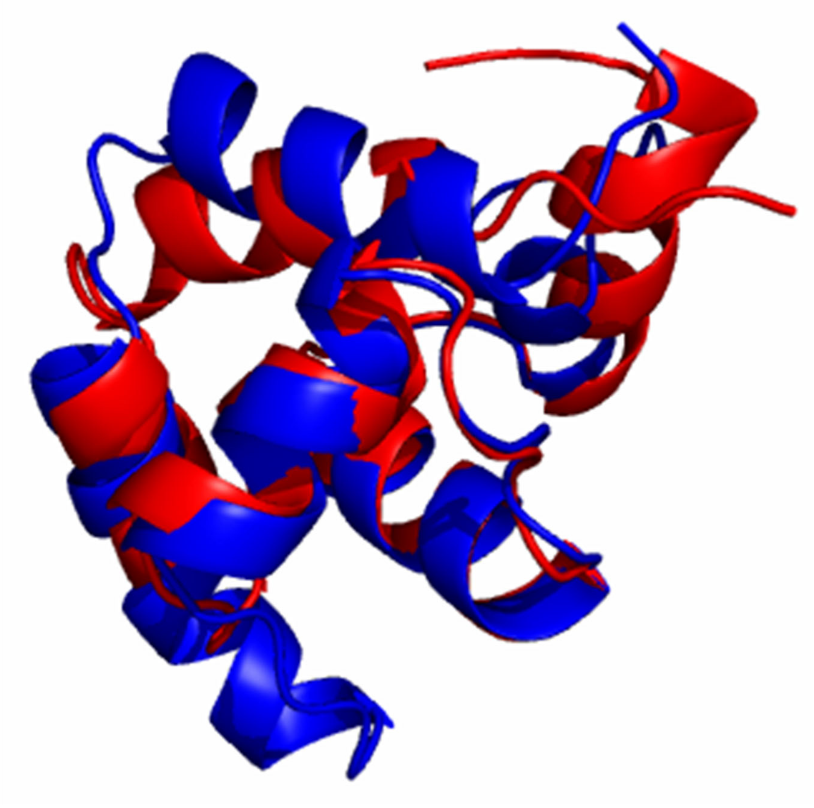


Figure 3.28. Comparison of TMalign (Zhang & Skolnick, 2005) structures for predicted T0974s1 and PDB ID 6tri
The structure in blue is the observed structure for PDB ID 6tri and the predicted structure for CASP13 target T0974s1 is in red . A TM-score of 0.79965 was achieved for protein structures. The score was normalised for PDB ID 6tri target as it is the reference molecule

A **B**

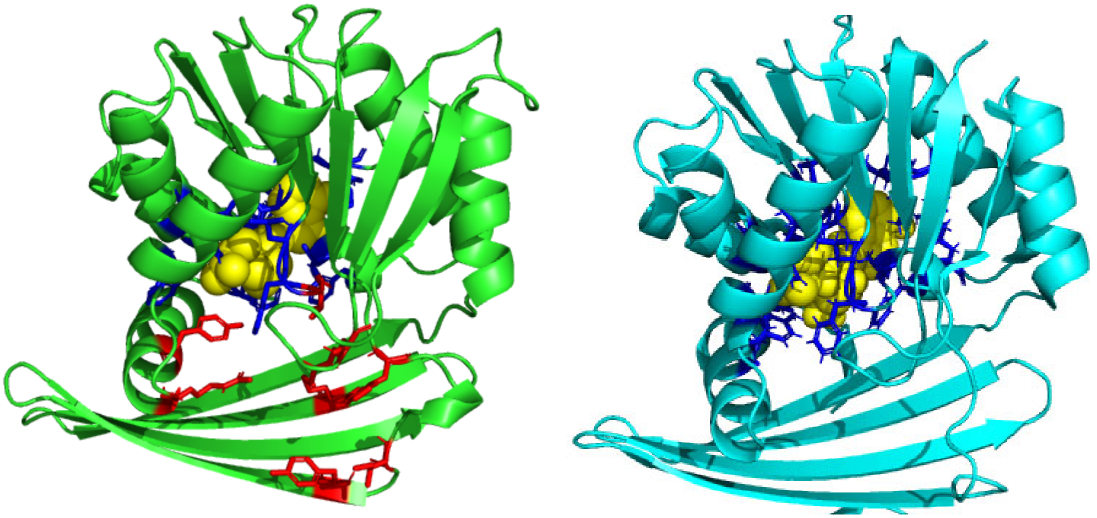


Figure 3.29. FunFOLD3 ligand-binding site predictions for CASP13 target T0983 (PDB ID 6uk5)
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red and correct predictions in blue, the predicted s-adenosyl-l-homocysteine (SAH) ligand shown as sphere and coloured yellow. MCC and BDT score was 0.715 and 0.715, respectively **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in blue the s-adenosylmethionine (SAM) ligand is shown as sphere and coloured yellow. The actual structure is a dimer, whereas the predicted structure is a homodimer. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target.

The sixteenth predicted CASP13 target was Cals10, an amino pentose methyltransferase and is a dimer, however, CASP13 organisers released part of the dimer, chain A for predictions to be made. Additionally, the PDB entry identifies SAM, pentaethylene glycol, (1PE), di(hydroxyethyl)ether (PEG) and acetate ion (ACT) as potential ligands. However, 1PE, PEG and ACT were not part of the observed structure so are likely to be solvents used in the crystallography process and FunFOLD3 did not identify these ligands as biologically relevant.

There is a difference in the predicted and observed ligand, despite a good MCC and BDT score. SAM is used by methyltransferase enzymes, such as this protein target, as a donor of a methyl group to a diverse range of substrates (Huang et al., 2020). The methyl group is directed bonded to the sulphur atom, which is therefore a sulphonium cation (Huang et al., 2020). Upon donation of the methyl SAM is converted to S-adenosylhomocysteine (SAH), an electrically-neutral thioether that is toxic to the cell (Huang et al., 2020). Based on this information, it is fair to assume that the ligands are identical and prediction is based upon the donated state of the enzyme.

Figure 3.30 below is of the TM-align superposition of observed and predicted structures with the observed structure from the PDB entry for the target. A TM-align score of 0.93352 was achieved showing very good structural homology between the predicted and observed structure.

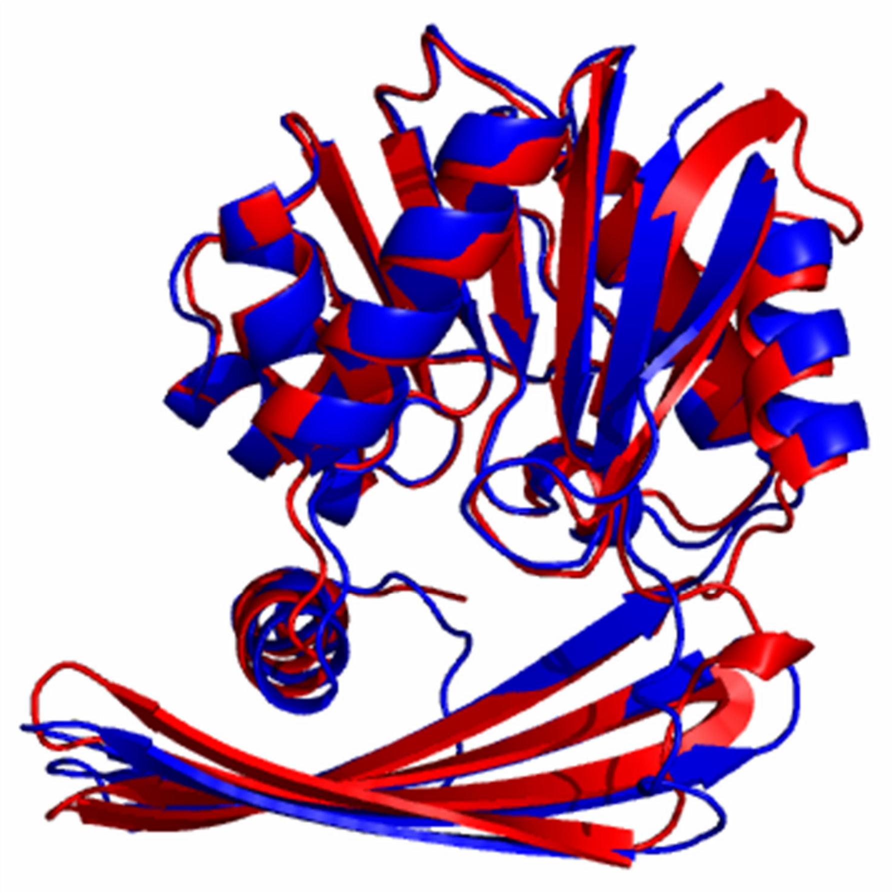


Figure 3.30. Comparison of TMalign (Zhang & Skolnick, 2005) structures for predicted T0983 and PDB ID 6uk5
The structure in blue is the observed structure for PDB ID 6uk5 and the predicted structure for CASP13 target T0983 is in red. A TM-score of 0.93352 was achieved for protein structures. The score was normalised for PDB ID 6uk5 target as it is the reference molecule

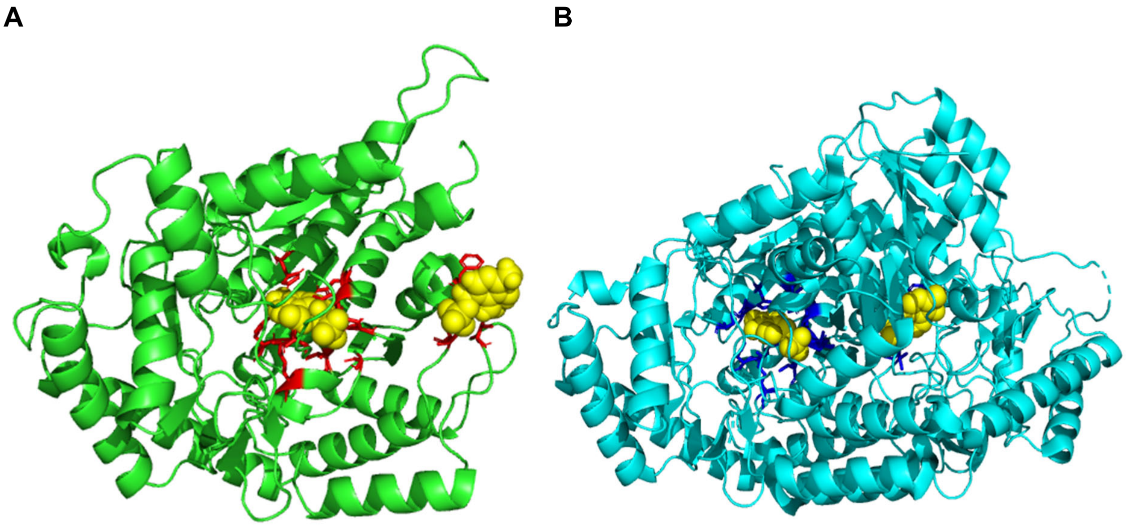


Figure 3.31. FunFOLD3 ligand-binding site predictions for CASP13 target T1003 (PDB ID 6hrh)
(**A**) Predicted ligand binding site residues shown as sticks with incorrect predictions in red the predicted pyridoxal-5'-phosphate (PLP) ligand shown as sphere and coloured yellow. MCC and BDT score was -0.04 and 0.06, respectively **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in blue the PLP ligand is shown as sphere and coloured yellow.

The twenty-fifth predicted CASP 13 target is ALAS2, a human erythroid-specific 5'-aminolevulinate synthase and is classified as an oxidoreductase as per the PDB ID entry. The protein is a dimer with two chains and the predicted structure released by the CASP organiser was part of the dimer. FunFOLD3 correctly identified the biologically relevant ligand, PLP. Despite the correctly predicted ligand the MCC and BDT scores were low, -0.04 and 0.06, respectively. Additionally, the information as per the PDB entry suggests that the ligand is present in two locations in both chains A and B of the molecule. However, the predictions from FunFOLD3 on the predicted protein suggest that the ligands are in the same location and this was the same for the observed structure. Potentially suggesting that the structure is a homodimer, which is further supported in literature (Na *et al.*, 2018).

Available information in literature, identifies the role of ALAS in the biosynthesis of haem. Biosynthesis of haem is a complex process that involves multiple stages controlled by different enzymes. The first protein in this stage is a pyridoxal 5'-phosphate (PLP)-dependent homodimeric enzyme, 5-aminolevulinate synthase (ALAS) (Na *et al.*, 2018). In eukaryotic ALAS from S.cereivisae, one ALAS subnit contains covalently bound cofactor, PLP, whereas the second is PLP-free, (Brown *et al.*, 2018) which is unlike the structures in figure 96. Comparisons of the subunits reveals PLP-couples reordering of the active site and of additional regions to achieve the active conformation of the enzyme (Brown *et al.*, 2018).

As ALAS is a member of a large family of enzymes that employ PLP, other members of the family are also homodimers with PLP-binding pockets located at the dimer interface (Brown *et al.*, 2018).

Figure 3.32 below, shows the TM-align, superposition of observed and predicted structures, with the observed structure as released from CASP13 organisers. A TM-align score of 0.91953 was achieved showing very good structural homology between the predicted and observed structure.

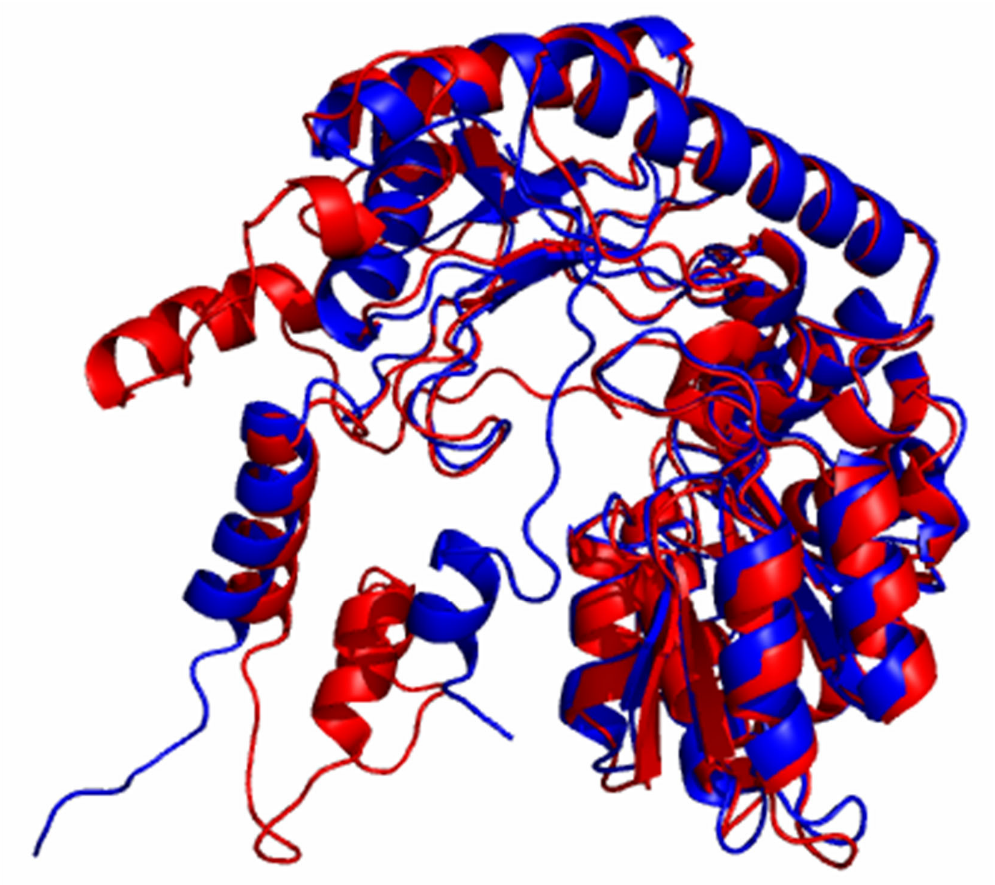


Figure 3.32. Comparison of TMalign (Zhang & Skolnick, 2005) structures for predicted and observed structure for T1003 (PDB ID 6hrh)
The structure in blue is the observed structure for T1003 and the predicted structure is in red. A TM-score of 0.91953 was achieved for protein structures. The score was normalised for the observed structure T1003 as it is the reference molecule

Table 3.12 below, is a list of the CASP13 targets which have a ligand as predicted by FunFOLD3 and where necessary the ligands according to the PDB entry is also listed.

Table 3.12. Comparison of ligands predicted using FunFOLD3 and ligands identified on the PDB

CASP 13 target	FunFOLD3 ligand	PDB ligand
T0949	Oxygen and Copper	N/A
T0953s2 (PDB ID 6f45)	DJB	MRD, MPD,IMD and MG
T0954 (PDB ID 6cvz)	Lysine and DNA	MG
T0955 (PDB ID 5w9f)	ZN and RNA	N/A
T0957s2 (PDB ID 6cp8)	SF4 and ALA	EPE and GOL
T0958 **(**PDB ID 6btc)	SAH and DNA	N/A
T0961 (PDB ID 6sd8)	FAD	FAD
T0965 (PDB ID 6d2v)	NAD	NDP, SCN and CL
T0970 (PDB ID 6g57)	0FX	N/A
T0972	HEM	N/A
T0973 (PDB ID 6yfn)	RNA	N/A
T0974s1 (PDB ID 6tri)	DNA	SO4
T0975	DNA and SF4	N/A
T0980s1 (PDB ID 6gnx)	PRO and SER	N/A
T0980s2 (PDB ID 6gnx)	SAH	N/A
T0983	SAH	SAM, 1PE, PEG and ACT

(PDB ID 6uk5)		
T0985	BGC	N/A
T0986s2 (PDB ID 6d7y)	DAL	N/A
T0992	CA	N/A
T0993s1 (PDB ID 6xbd)	ADP	PEF
T0994	CAZ and HIS	N/A
T0995	CDT	N/A
T0997	DGL	N/A
T1001	BLA	N/A
T1003 (PDB ID 6hrh)	PLP	PLP
T1008 (PDB ID 6msp)	FMN	N/A
T1009 (PDB ID 6dru)	GLC	NAG, XYS and GOL
T1012	ACO	N/A
T1013	CLR, MPG, RET and LEU	N/A
T1014 (PDB ID 6qrj)	ADP	ANP and MG
T1016 (PDB ID 6e4b)	PO4	1PE, GOL and CL
T1017	ZN	N/A
T1018 (PDB ID 6n91)	ZN	DCF, CXS, SO4, PO4, GOL, ZN, EDO, FMT and NA
T1023s3	LEU, GDP, RNA and ZN	N/A

3.4 Further discussion and summary of CASP11, CASP12 and CASP13

In this chapter FunFOLD3 was used to predict the ligands and ligand-binding site residues for CASP11, CASP12 and CASP13. In CASP11, a total of nine targets had ligand-binding sites and ligands predicted. The MCC score ranged from 0.877 to -0.05 and BDT score ranged from 0.853 to 0.035 this gave an average of 0.417 for MCC and a BDT score of 0.51 based on eight targets and the highest MCC and BDT achieved if multiple ligands were not predicted. The ninth target was excluded from the MCC and BDT score due to different predicted ligands and sharing no correct ligand-binding site residues between the observed and predicted proteins.

In CASP12, predictions were more complicated to analyse due to either observed structures not being released by CASP12 organisers and/or no PDB ID released for the targets. Furthermore, predictions were somewhat poorer, with only one target; T0916 (PDB ID 5tj4) having correctly predicted some of the same residues as in the observed structures. This was despite differences in the predicted (NAD) and observed ligand (GLC). None of the predicted and observed ligands matched for the targets. There was an increase in the number of targets which were predicted and a total of 12 targets had predictions.

CASP13 showed the biggest increased in predictions with a total of 34 targets, as with CASP12 some of the targets were harder to analyse due to not having an observed structure, however if there was a PDB ID associated then the file from the Protein Databank database could be used in the analysis. This CASP also saw a diverse range of ligands predicted from metal ions such as zinc to larger more complex molecules such as DNA/RNA molecules, highlighting the versatility of FunFOLD3 for the first time. Although, in the targets which predicted RNA/DNA (T0954, T0955, T0958, T0973, T0974s1, T0975, T1023s3) the observed structures did not have RNA/DNA. Only two targets; T0954 and T0974s1 had ligands with MG and SO4, respectively. Out of the 34 targets only two target had a correct prediction; T0961 (PDB ID 6sd8) and T1018 (PDB ID 6msp) with other targets arguably having the same ligand just in different states due to involvement in reactions (T0965, T0983 and T1014). Therefore, the “state” in which proteins are predicted needs to be considered, before an reaction or after, as this can clearly impact on the ligands predicted. For CASP13 the MCC score ranged from 1.0 to -0.05 and BDT score ranged from 1.0 to 0.05. Thus giving an average of 0.4686 and 0.5264 for MCC and BDT, respectively. This was based on ten targets with three targets having the same ligands (T0961, T1003 and T1018) and two targets having potentially the same ligands in a different state (T0983 and T1014). Interestingly, the target with the perfect score; T0974s1 did not have the same ligand with DNA being predicted and sulphate identified as the observed ligand.

For eight out of the nine CASP11 targets at least one of the ligands compared to the ligands in the PDB entry were identified, the exception was T0783 where none of the predicted ligands matched those identified on PDB. For some CASP11 targets (T0798, T0845 and T0849), all the ligands were identified. However, it is worth considering whether all ligands in PDB are biologically relevant, considering that the basis for predicting ligands with FunFOLD3 is whether or not the ligands are biologically relevant. As seen in Table 2.4, there was a difference in the predictions of ligands by FunFOLD3 and those in PDB, for example CASP11 targets T0783, T0807, T0813, T0819 and T0845.

For CASP11 target T0783, EDO is listed as a ligand and has the potential to be a ligand as it is an alcohol and widely used in antifreeze formulations. However, it is also used as a solvent so could potentially be an artifact and is not a biologically relevant ligand. The same can be of CASP11 target T0786 with PDB having PG4 as a ligand, T0807 with acetate ligand, T0819 with PE4, PGE, PEG and GOL and T0845 with 7PE and ACT. Glycerol (GOL) is quite interesting as with BioLiP, information in literature is used to filter out biologically irrelevant ligands, therefore FunFOLD3 will only predict ligands which are biologically relevant due to the filtering process. If a ligand is in the artifact list, the simplest way is to treat it as biologically irrelevant and discard it (Yang, Roy and Zhang, 2013). However, there is a flaw with this method as it can mean some ligands are missed and will be false negatives and these ligands could be biologically relevant. Glycerol is one of the most frequently used crystallisation additives and it is therefore regarded as biologically irrelevant by many existing databases (Yang, Roy and Zhang, 2013). However, glycerol can have a biological role in some proteins. An example being glycerol binds to the protein enzyme diol dehydratase (PDB ID: 3auj) with a biological role as glycerol is bound to the substrate binding site in the $(\beta/\alpha)_8$ or TIM barrel of the diol dehydratase α subunit (Yamanishi *et al.*, 2012). This ultimately led to glycerol being added as a biologically relevant ligand (Yang, Roy and Zhang, 2013).

In four of the highest scored predictions, the ligands were quite large. For example, for the CASP 11 T0819 prediction which has a BDT and MCC score of 0.853 and 0.877, respectively, the ligand was PLP, which has a molecular weight of 247 g/mol. For the CASP11 T0807 prediction, which has a BDT/MCC score of 0.849 and 0.771 respectively, the identified ligand was NAP which has a molecular weight of 743.9 g/mol. The prediction for CASP11 T0854, with a BDT and MCC score of 0.845 and 0.745 respectively, is the exception to this rule as the identified ligand is magnesium with a smaller molecular weight of 24.4 g/mol. Finally, the CASP11 T0798 prediction had a BDT/MCC score of 0.797 and 0.753, respectively the ligand was GDP with a molecular weight of 443.2 g/mol.

A finding with some of the CASP11 targets (T0807 and T0849) was extending the ligand-binding residue sites to potentially incorporate MSE. Upon inspection of the structure files for both the predicted and observed CASP11 targets, the MSE residues are not present in the predicted files, residues are MET, and in comparison with the observed structure files MSE residues have remained. Therefore, this only provides a possible explanation as to why the ligand binding site residues were over-extended (refer to Appendix 2: Predicted and observed structure files for CASP11 target T0807). This is potentially a bug within the method and can be recoded so that MSE residues in the templates, which are actually METs and part of the sequence, are not accidentally included as ligands.

Identification of the ligands that bind to a protein is important for understanding the function of the protein. For example, the CASP11 target protein T0849 has only one associated ligand, showing the protein has high specificity and selectivity. More importantly, the ligand that binds to this protein is glutathione; an important molecule for the phase II metabolism of xenobiotics (Ekici, Paetzel and Dalbey, 2008). On this basis, it is rational to assume that the protein is glutathione-S-transferase. These results suggest, that the FunFOLD3 server is better at predicting ligands, which have large molecular weights. An explanation could be that the ligand binding residues that cover a large part of the protein and are much more likely to be contained in a cluster. A universal finding, regardless of ligand size, is that there are usually a few key residues involved and more importantly, it is paramount to identify these key sites in order to fully understand function (*The Protein Prediction Center*, 2014) This pattern was not seen with CASP13. The prediction for T1009 had a MCC and BDT score of 0.91 and 0.94, respectively, the predicted ligand GLC has a molecular weight of 180.16 g/mol and in comparison the result for T1014 had a MCC and BDT score of -0.05 and 0.05, respectively and the predicted ligand ADP had a molecular weight of 427.20 g/mol.

The average TM-align scored obtained for all nine CASP 11 targets was 0.76, suggesting that IntFOLD is able to select protein structures towards the high structural similarity based on the top model being used for predictions. However, the average MCC and BDT score of 0.459 and 0.51, respectively, suggests that further refinement of the server is needed in order to better predict ligand-binding sites. Despite having highly similar structures to the observed protein models, the ligands contacting residues did not follow the same high level of prediction, this is because there is a difference between predicting the tertiary structure of a protein, compared to the active site which is very localised. A novice could assume that having highly similar structures of the predicted protein model to the observed model, would automatically translate to ligand residues also following the same high level of prediction. However, it is worth considering convergent evolution seen in serine proteases (Dukka, 2013). Proteases play a variety of roles within all living cells and are typically grouped into four mechanistic classes: cysteine, serine proteases, metallo and aspartic proteases. The best known class is the serine protease class which uses the classical Ser/His/Asp catalytic triad mechanism. Clans are formed from the active site arrangement and members of the same clan use different active site architectures, indicating that the tertiary structure is not always related to the active site configuration (Ekici, Paetzel and Dalbey, 2008). This potentially allows for varying activity in different cellular environments. Subtilisin is an example of a protease which like chymotrypsin and trypsin utilises a Ser/His/Asp triad but has no sequence similarity to chymotrypsin-like proteases. Thus demonstrating, quite excellently convergent evolution; the folds of these proteases are completely different although they both converged in a similar Ser/His/Asp mechanism to carry out proteolysis (Ekici, Paetzel and Dalbey, 2008).

An additional observation from CASP11 is the difference between the observed and predicted residue numbers. For CASP11 targets where there was no issues with ligand-binding prediction based on the MCC and BDT score, the difference in residues appears to not be an issue. However, for some CASP11 targets such as T0813, the extra residues were THR308, THR309, THR310, THR311, LEU312, TYR313, LEU314 and LEU315, an observation is these residues are located in the flexible loop of the protein. The inability of IntFOLD3 to correctly fold this region and thus for FunFOLD3 to identify it as a ligand could have led to the low MCC and BDT score (0.08 and 0.194, respectively) associated with this target.

In comparison to previous years, the format of the CASP11 experiment, participation statistics and number of targets are very similar. Almost 60,000 models on 100 prediction targets were collected from 207 modelling groups which represented 100 research labs across the globe (Feller & Lewitzky, 2012). In a publication about the progress and new directions in CASP11(Moult *et al.*, 2016), the authors state the most exciting result in CASP11 was the generation of an accurate three dimensional model of a large (256 residue) protein. The authors believe this was due to more accurate prediction of contacts between protein residues (Moult *et al.*, 2016). This was CASP target T0806 (PDB ID 5caj). As this protein did not bind any ligands, it did not form part of the analysis.

CASP12 saw an apparent change in ligand-binding site predictions compared to CASP11. In CASP11, all of the CASP targets with ligands had an associated PDB ID, that meant there was a clear method of analysing results. However, there was a shift in CASP12 where six of the twelve targets had an associated PDB ID. This did make analysis of the results quite limiting due to sparse data, as some proteins had limited annotations. This suggests there could potentially be a shift in elucidating the importance of ligand-binding site from the CASP organisers. By a way of example, in CASP11 the aim of ligand-binding site predictions could be seen as “bench-marking” as results could be confirmed with data available from PDB or BioLiP, allowing participating groups to assess the quality of predictions with relative ease. In CASP12 with six targets not having associated PDB IDs, at this stage, there is currently – without further investigation – there is no definite way of determining if this is correct. Rather than this being a hurdle, it means there is a window of opportunity to start investigating minimal/low annotation proteins. CASP12 targets T0899, T0901, T0905 and T0907 illustrates this perfectly; there are no data on whether the predicted MG ligand is correct or indeed the ligand-binding residues. From a research perspective, this is intriguing as it potentially means that new information has been gathered on this protein and could be a contribution to available data. Thus, it would need to be supported by the use of other computational methods such as protein-ligand docking experiments.

As discussed previously, FunFOLD3 predicts ligands based on the concept that protein structure superposition of distantly related templates to a modelled protein can aid in identification of ligand binding sites (Moult *et al.*, 2016). Output files following a FunFOLD3 prediction include, but are not limited to, a FN file and ligand-binding residue locations in a PDB file. The FN prediction file provides information, where applicable on GO terms, EC numbers, ligand-binding residues and predicted ligands. For proteins with no PDB IDs, information on related proteins could prove useful in determining the protein's function. This was of particular interest in CASP12 due to the lack of PDB IDs associated with targets and also with CASP13 for the same reason. In order to determine further information on the low annotation proteins, a Pfam search was performed. Pfam is a database of protein families, where families are sets of protein regions that share a significant degree of sequence similarity, thereby suggesting homology (Roche, Tetchner and McGuffin, 2011). Pfam contains two types of families: high quality, manually curated Pfam-A families and automatically generated Pfam-B families . Members of the same family are expected to share a common evolutionary history and thus at least some functional aspect. However, Pfam does stress that homology is no guarantee of functional similarity and transfer of functional annotation based solely on family membership should always be undertaken with caution (Punta *et al.*, 2012). On the other hand, additional data from Pfam, such as conservation of common domain architectures can increase confidence in a given functional hypothesis (Punta *et al.*, 2012). Six CASP12 targets with little/no annotation had a Pfam search performed and no information was found for four of the targets; T0899, T0901, T0905 and T0919. One of the targets; T0911 was identified as being part of a major facilitator family which ties in with the description provide by the CASP12 assessors (D-galactonate transport) and a GO term prediction related to transmembrane transporter (GO:0055085). CASP12 target T0916 was identified as gasdermin pore forming domain and is aligned with the CASP12 description of gasdermin. The Pfam entry for this protein states the precise function is unknown, however it is thought that this protein plays a role as a secretory or metabolic product involved in the secretory pathway and includes gasmerdin amongst non-syndromic hearing impairment protein 5 (DFNA5) and pejvakin. The information from Pfam ties in with the information from the GO term prediction by FunFOLD3, as three of the 15 predictions related to metabolic process (GO:0006004, GO:0005975, GO:0051143). The information from Pfam has not provided ay additional information which can aid in determining function of the poorly annotated proteins in the CASP12 competition but did slow a link between the results from Pfam and FunFOLD3 GO term prediction albeit, quite weak. As a result of Pfam not providing any further information into the function of proteins, it will no longer be used as part of an analysis or summary but has been mentioned in this section to illustrate findings and why no further reference is required.

In contrast to CASP11, where only one target (CASP11 target T0783) had predicted ligands which did not match the PDB ligands. For the CASP12 predictions, all of the predicted ligands, were not associated with the protein’s PDB entry. This potentially suggests a flaw with the predictions or that not all the ligands associated with the proteins as per the PDB entry are biologically relevant.

In CASP13, 75 proteins and protein complexes were suggested as modelling targets by 36 structure determination groups from 14 countries.(Lepore *et al.*, 2019) As previously alluded, some targets had their structures cancelled, this amounted to a total of eight. This was due to: lack of structure by the time of the assessment or release of relevant structural information before the end of the time of assessment, or release of relevant structural information before the end of the target prediction session (Lepore *et al.*, 2019). Of the remaining 67 entries, 58 were solved by X-ray crystallography, seven with cryo-EM and two by NMR,(Lepore *et al.*, 2019) of which FunFOLD3 identified 34 as containing ligands. T0958 (PDB ID 6btc) was called LP1413, as it was a little protein (96 amino acids) and annotated as containing DUF (domain of unknown function 1413) (Lepore *et al.*, 2019). Whilst no enzymatic activities in the purified protein was found, it does bind single-stranded DNA with high affinity (Lepore *et al.*, 2019) and this somewhat supports the prediction by FunFOLD3, despite both the observed structure and the PDB entry not supporting this prediction. Furthermore, the protein adopts a winged helix-turn-helix fold, so it would be wise to guess that it binds double- rather than single-stranded DNA binding as its function (Lepore *et al.*, 2019) and could be why FunFOLD3 predicted double-stranded DNA. T0961 correctly identified the observed ligand with the single FAD ligand binding per monomer in a crevice located in the dimer interface (Lepore *et al.*, 2019).

CASP13 brought a diverse range of ligand-binding site predictions and due to lack of PDB IDs being associated with the targets, this meant we had to use the predicted GO terms by FunFOLD3, in conjugation with the predicted ligand information contained within UniProtKB in order to provide insights into the potential function of the protein. An example of this is CASP13 target T0973 (PDB ID 6fyn), where the predicted ligand was RNA and the available information in literature supports this as the protein encapsulates ssRNA. Additionally, and unsurprisingly, RNA binding was a GO term that was predicted. Thus, using information potentially provided insight into the function of this target. Despite alignment across several different literature resources the PDB ID entry identified calcium as a ligand. The most interesting target in terms of prediction with no available PDB ID is T0975. FunFOLD3 identified two DNA ligands and an iron/sulphur cluster. The UniProtKB entry and the ligands align perfectly with the function being ssDNA bidirectional exonuclease and FunFOLD3 predicted DNA in two different locations, with the protein exhibiting both 5’-3’ and 3’-5’ activities which could be why two different DNA locations were predicted. Additionally, the iron/sulphur cluster also has a key role in structural linking to create a cavity that encircles the ssDNA. Currently, as there is no PDB ID associated with this target and nor was an observed structure released it is difficult to confirm these findings. However, a target such as this would be an ideal opportunity to explore further with support from crystallisation experiments. Targets such as this, potentially highlight how *in silico* modelling is useful for determining the ligands and ultimately function of a protein before exploration and confirmation with more expensive crystallisation techniques.

Another interesting observation was the prediction of ligands in different states due to involvement with reactions. For example, CASP13 targets T0965 (PDB ID 6d2v) and T0983 (PDB ID 6uk5) have predicted the correct ligand predicted, however both are in the reduced state following a reaction. This highlights another consideration in protein-ligand prediction, the structure of a protein can change following a reaction and so can a ligand, although it will be the same ligand but in a different state and therefore will not be ruled out.

Several CASP13 targets were discussed in a publication and has provided further insight and understanding into some of the CASP13 predictions. T0953 (PDB ID 6cvz) was predicted to have DNA as a ligand with magnesium being the observed ligand. Residues involved in DNA damage repair are not part of distinct residues which were identified (467-477, 594-606, and 656-664) in the disordered loops in the top and bottom faces (Lepore *et al.*, 2019). TRP-543 and ILE-639 bind RPA32 and the QKMDF consensus motif that mediates the interaction with proliferating cell nuclear antigen during DNA replication (Lepore *et al.*, 2019). This could provide one explanation as to why DNA was predicted by FunFOLD3.

As with CASP12, a Pfam search was conducted on the eleven targets with no PDB ID associated and no/little annotation as per the UniProtKB entry, despite the mixed results obtained with the CASP12 targets. For target T0949 there were no significant matches but there were four insignificant matches. Two related to Cupredoxin_1 and two for DUF5060. Cupredoxin-like fold consists of beta-sandwich with seven strands in two beta-sheets in a Greek-key beta-barrel and contains copper bound within the structure. This has similarities with the predicted structure as it contains seven beta-sheets but also one alpha-helix and predicted copper in the structure. For DUF5060 no similarities were seen with the target (El-Gebali *et al.*, 2019). T0972 has two insignificant matches, one of which is DNA_pol_B_3 and is DNA polymerase family B viral insert (El-Gebali *et al.*, 2019) showing similarity with the CASP13 target as DNA was predicted by FunFOLD3.

T0985 had a significant match related to Glyco_hyfdro_36 which is the exact same protein which UniProtKB identifies the protein additionally several PDB entries are associated with the Pfam entry of which seven were identified by FunFOLD3 as templates containing biologically relevant ligands (1v7v, 1v7w, 2cqt, 3act, 3qde, 3qfy and 5h3z) (El-Gebali *et al.*, 2019). T0994 had eight matches, two of which were significant and six were insignificant. The significant matches were peptidase_M56 and described as a BlaR1 peptidase M56 and transpeptidase which is described as penicillin binding protein transpeptidase. The description of BlaR1 peptidase M56 is the production of beta-lactamase and penicillin-binding protein 2a is regulated by a signal-transducing integral membrane protein and a transcriptional repressor. The signal transduced is a fusion protein with penicillin-binding and zinc metalloprotease domains. None of the associated PDB entries were templates that contained biologically relevant ligands (El-Gebali *et al.*, 2019). In comparison, the penicillin binding protein transpeptidase domain has seven of the templates which FunFOLD3 predicted with biologically relevant ligands associated (1ax1, 1xkz, 2iwb, 3q81, 3vma, 4jf4 and 5e2f). The entry also states that the active site serine (residue 337) is conserved in all members of this family (El-Gebali *et al.*, 2019). Of note, this residues was not predicted as a ligand-binding residues by FunFOLD3.

T0995 had one significant match, CN_hydrolase which is a carbon-nitrogen hydrolase and does appear to be quite different to the UniProtKB entry which identified the protein as cyanide dehydratase however, two of the templates identified as having ligands by FunFOLD3 (1ems and 1uf5) are also associated with the Pfam entry (El-Gebali *et al.*, 2019).

T0997 has a significant match with YkuD_2 which is a L,D-transpeptidase catalytic domain however no information is provided in the entry (El-Gebali *et al.*, 2019). T1012 had a significant match with Acetyltransf_1 with a description of acetyltransferase (GNAT) family and information contained on the InterPro entry states that the n-acetyltransferases are enzymes that use acetyl coenzyme A (CoA) (El-Gebali *et al.*, 2019) and this matches the ligand predicted by FunFOLD3 and the molecular function in the entry GO:0008080 (N-acetyltransferase activity) matches one of the GO terms predicted by FunFOLD3.

T1013 was classified as an unknown protein by CASP13 and on Pfam two significant matches were identified 7tm_1 and Glyco_hydro_11. The Pfam entry for 7tm_1 expands on the protein and that it is a seven transmembrane receptor (rhodopsin family) and contains amongst other G-protein-coupled receptors (GCPRs), members of the opsin family, which have been considered to be typical members of the rhodopsin superfamily (El-Gebali *et al.*, 2019). The GO terms associated with the InterPro entry are GO:0016021 (integral component of membrane), GO:0004930 (G protein-coupled receptor activity) and GO:0007186 (G protein-coupled receptor signalling pathway) all of which were predicted by FunFOLD3 (El-Gebali *et al.*, 2019).

T1017s1 had a significant Pfam match with DUF596 and is described as a protein of unknown function and there is no further information available, (El-Gebali *et al.*, 2019) which supports the information as per the UniProtKB entry. There were no matches were found for T0992 and T1001 so these targets have not been discussed.

The average TM-score obtained for the CASP13 targets was 0.643 across 24 targets which had an observed structure from CASP13 or the structure from the PDB entry. This was lower than the TM-align score obtained for CASP11. However, this score did not seem to impact the average MCC and BDT scores which were slightly higher than those obtained for CASP11. This is most likely aided by a perfect prediction for T0974s1, despite a different ligand being predicted.

When comparing the observed and predicted structures using TM-align, there appears to be the greatest visible difference in the flexible loops of the predicted structure (e.g. CASP13 target T0953s2 see Figure 3.23). Flexible loops are less ordered and ideally situated to form binding sites for other molecules. This is well demonstrated in folds on immunoglobulin antibodies (Rang et al., 2015), these loops fold around each of the antigen molecules and this is made easier by the flexibility of the loops (Alberts et al., 2002). When determining protein structure, it is important remember that a protein’s structure is not static but undergoes conformational changes to undergo ligand-binding and this is worth noting when a protein binds to a ligand it can undergo conformational changes (Alberts et al., 2002). Loop regions of proteins occur in inter-domain segments of otherwise well-folded where they can serve multiple functions: short loop sometimes feature as mere linkers or may also provide the required flexibility for the movement of the neighbouring domains (linker loops) (Feller & Lewitzky, 2012). Other loops serve as linker regions but also allow proteins to interact intra-molecularly when undergoing shape changes (intramolecular docking loops). Short loops localised within a well-folded protein domain can also work together to form binding pockets for proteins and a range of other biomolecules (binding pocket loops) (Feller & Lewitzky, 2012). Hence, this provides a possible reason as to why the flexible portions of the predicted proteins contain additional residues and can be difficult to order with the rest of proteins (Feller & Lewitzky, 2012).

CASP13 organisers concluded that accurate prediction of loops is still a challenging task (Lepore *et al.*, 2019). Due to flexible loops often being involved in protein-protein interactions, their incorrect prediction can compromise the accuracy of the interacting surface and overall structure of the complex (Lepore *et al.*, 2019).

A key question in protein modelling is whether a structure is accurate enough to answer a specific biological question. An example of this can be seen with the CASP11 target T0783 (PDB ID 4cvh). This target had the lowest TM-score of 0.54529 within this CASP and despite the low structural similarity and potentially all the biologically relevant ligands were predicted. Thus, part of the biological question; “what are the biologically relevant ligands?”, was answered. This question becomes of paramount importance in drug development (Moult *et al.*, 2016).

The results across the three CASP experiments highlight the wide range of ligands that the FunFOLD3 server can predict; ranging from simple metal ions such as magnesium and sodium, to larger more complex molecules, such as glutathione. For the total 57 protein targets that had ligands predicted across the three CASP competitions, 15 (27%) of the targets had at least one metal ion predicted. Metalloproteins make up some 30% of proteins in known genomes (Hasnain, 2004). Metalloproteins are a special class of proteins that utilise the unique properties of metal atoms in conjugation with the macromolecular assembly to perform life-sustaining processes (Hasnain, 2004). Metals may play structural roles (e.g. zinc in zinc-finger domains) or enzymatic roles (e.g. zinc in carbonic anhydrase).

Metal coordination has been found to both stabilise and destabilise the folded states of their corresponding proteins *in vitro* (Palm-Espling, Niemiec and Wittung-Stafshede, 2012). Therefore, the ability of FunFOLD3 to predict metal ions should not be underestimated and can also provide insights to the function of a protein.

In conclusion, the results from the CASP11, CASP12 and CASP13 analyses suggest the FunFOLD3 server is useful for predicting binding site residues. Additionally, the results also support the use of FunFOLD3 for determining ligands. Based on the ligand prediction results, FunFOLD3 has potentially assisted in determining the relevance of ligands as per the PDB entry for proteins. If knowledge of protein-ligand binding was relied upon solely using PDB it could potentially lead to misunderstanding about the function of a protein, however that it not to say to completely rule out the information on PDB but to be cautious in relying on this information alone. Therefore, FunFOLD3 has potentially contributed to knowledge of proteins to the wider scientific community. However, further validation is necessary and this can be obtained from protein-docking and wet lab experiments.

GO terms were also reported for CASP targets where no observed structure was available or where there were discrepancies between a predicted and observed ligand. This was helpful in order to provide understanding of the protein’s function, in particular when no observed structure was available. The benchmarking of GO terms will be explored in Chapter 5.

This chapter has focused on functional predictions by FunFOLD3 across three CASP competitions (CASP11, CASP12 and CASP13). Chapter 4 will analyse predictions specifically from IntFOLD4. The FunFOLD3 component of IntFOLD4 outputs ligand-binding predictions and as with FunFOLD3, the purpose of the next chapter will be to objectively compare the observed predictions against observed proteins and potentially provide insight into the strengths and weaknesses of the predictions.

Chapter 4: Analysis of protein targets from CASP12 by IntFOLD4

4.1 Background to IntFOLD4

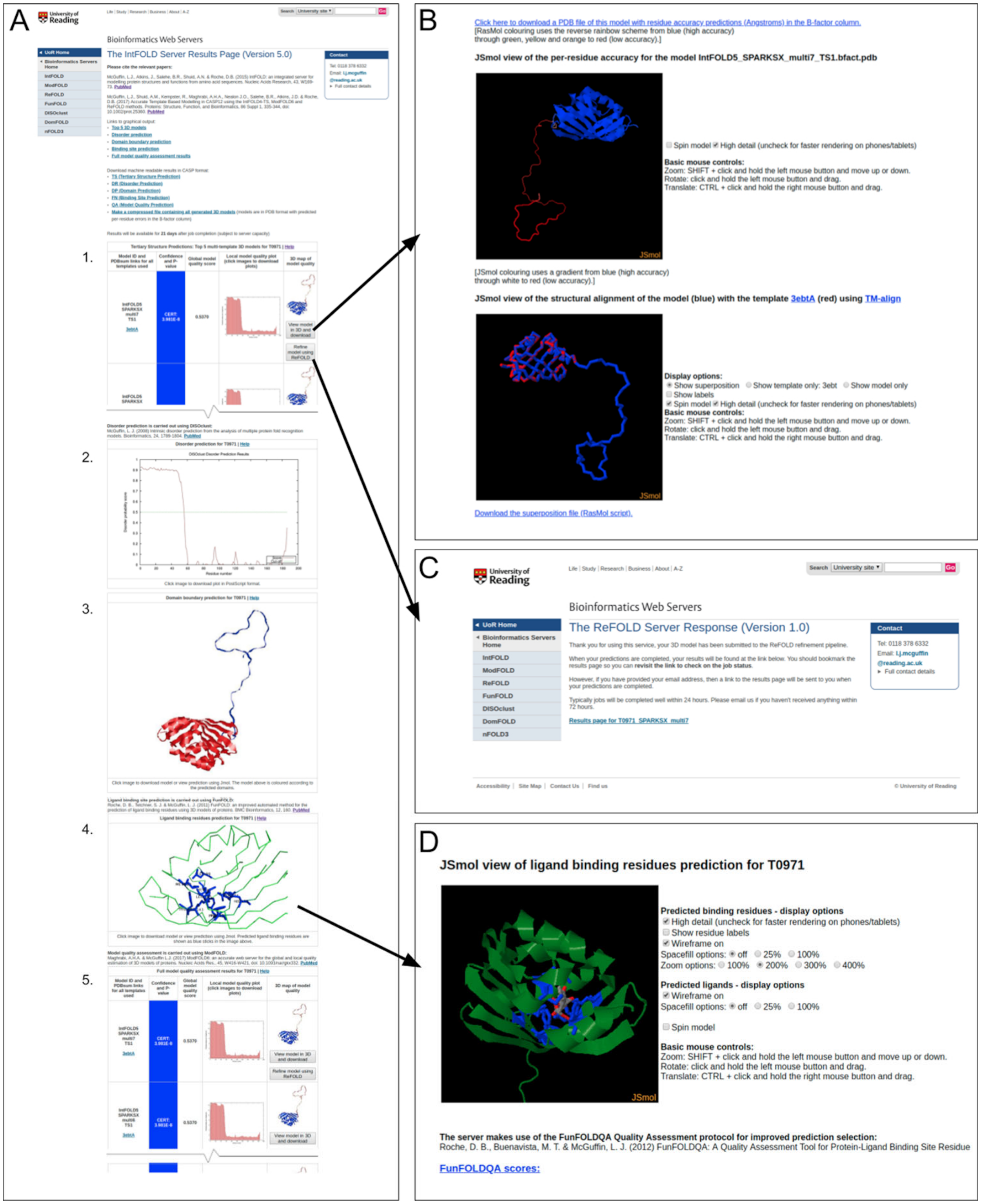
As mentioned in Chapter 1, the IntFOLD server is a fully integrated pipeline incorporating the latest methods for tertiary structure prediction, domain boundary prediction, prediction of intrinsically disordered regions, prediction of protein-ligand interactions and the global and local quality assessment of predicted models of proteins (McGuffin *et al.*, 2019).

Tertiary structure prediction using IntFOLD-TS produces full atom models and are subsequently ranked using the ModFOLDclust model quality assessment method. Disorder prediction is performed by DISOclust and depends on the ModFOLDclust QMODEL output in order to identify the regions of high variability occurring in 3D models generated for the nFOLD stack. Domain prediction is performed using DomFOLD and utilises the PDP method in order to identify structural domains in the top model obtained from the IntFOLD-TS method. Function prediction using FunFOLD is the basis of this thesis and is utilised to produce ligand-binding site residue predictions. The FunFOLD algorithm works by performing model-to-template superpositions of the top ranked IntFOLD 3D model and related templates with bound ligands to identify putative contacting residues (Roche *et al.*, 2011).

The methods within the IntFOLD server are interdependent, with output from one algorithm becoming the input for another (Roche *et al.*, 2011). The outputs from IntFOLD have been tested in the community wide experiment on the critical assessment of methods for protein structure prediction (Roche *et al.*, 2011).

IntFOLD is not the only server in existence which is able to generate results using the above mentioned methods. Distill is a suite of servers for the prediction of protein structures and features include, secondary structure, relative solvent accessibility, contact density, backbone structural motifs, residue contact maps at 6, 8 and 12 Ångstroms and course protein topology (Roche *et al.*, 2011). At the start of the development of the IntFOLD server was unique as it provided an integrated underlying methodology, unified graphical output and a single point for submission (Baú *et al.*, 2006). Currently, other methods have also applied protein model quality assessment such as, MULTICOM (Cao, Wang and Cheng, 2014)which encompasses four automated methods (MULTICOM-REFINE, MULTICOM-CLUSTER, MULTICOM-NOVEL and MULTICOM-CONSTRUCT) and was developed during CASP10 (Roche *et al.*, 2011). The foundation to all the methods in the server is ModFOLD, a model quality assessment tools, which us used to rank all models in terms of their global quality, as well as providing estimates of local quality as a distance in Ångstroms (Cao, Wang and Cheng, 2014).

The only requirement for input is a protein sequence in single letter code. The results for each submission to the IntFOLD server is then formatted into a single table which summarises all the prediction data graphically through thumbnail images of plots and annotated 3D models. The sections consist of; top five 3D models, disorder prediction, domain boundary prediction, binding site prediction and full model and quality assessment results. Figure 4.1 is an example of the output for CASP13 target T0971


A
B
C
D
1.
2.
3.
4.
5.
University of Reading
Life | Study | Research | Business | About | A-Z
Search University site ▾ Go
UoR Home
Bioinformatics Servers Home
IntFOLD
ModFOLD
ReFOLD
FunFOLD
DISOclust
DomFOLD
nFOLD3
Contact
JSmol
DISOclust Disorder Prediction Results
Disorder probability score
Residue number
Score
Cut-off
View model in 3D and download
Refine model using ReFOLD

Bioinformatics Web Servers

The IntFOLD Server Results Page (Version 5.0)

Please cite the relevant papers:

McGuffin, L.J., Adkins, J., Salehe, B.R., Shuid, A.N. & Roche, D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic Acids Research, 43, W169-73. PubMed

McGuffin, L.J., Shuid, A.M., Kempster, R., Maghrabi, A.H.A., Nealon J.O., Salehe, B.R., Adkins, J.D. & Roche, D.B. (2017) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. Proteins: Structure, Function, and Bioinformatics, 86 Suppl 1, 335-344, doi: 10.1002/prot.25360. PubMed

Tel: 0118 378 6332
Email: l.j.mcguffin@reading.ac.uk
► Full contact details

Links to graphical output:
- Top 5 3D models
- Disorder prediction
- Domain boundary prediction
- Binding site prediction
- Full model quality assessment results

Download machine readable results in CASP format:
- TS (Tertiary Structure Prediction)
- DR (Disorder Prediction)
- DP (Domain Prediction)
- FN (Binding Site Prediction)
- QA (Model Quality Prediction)
- Make a compressed file containing all generated 3D models (models are in PDB format with predicted per-residue errors in the B-factor column)

Results will be available for **21 days** after job completion (subject to server capacity)

Tertiary Structure Predictions: Top 5 multi-template 3D models for T0971 | Help

Model ID and PDBsum links for all templates used	Confidence and P-value	Global model quality score	Local model quality plot (click images to download plots)	3D map of model quality
IntFOLD5 SPARKSX multi7 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download Refine model using ReFOLD
IntFOLD5 SPARKSX				

Disorder prediction is carried out using DISOclust:
McGuffin, L. J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics, 24, 1799-1804. PubMed

Disorder prediction for T0971 | Help

Click image to download plot in PostScript format.

Domain boundary prediction for T0971 | Help

Click image to download model or view prediction using Jmol. The model above is coloured according to the predicted domains.

Ligand binding site prediction is carried out using FunFOLD:
Roche, D. B., Tetchner, S. J. & McGuffin, L. J. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics, 12, 160. PubMed

Ligand binding residues prediction for T0971 | Help

Click image to download model or view prediction using Jmol. Predicted ligand binding residues are shown as blue sticks in the image above.

Model quality assessment is carried out using ModFOLD:
Maghrabi, A.H.A. & McGuffin L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. Nucleic Acids Res., 45, W416-W421, doi: 10.1093/nar/gkx332. PubMed

Full model quality assessment results for T0971 | Help

Model ID and PDBsum links for all templates used	Confidence and P-value	Global model quality score	Local model quality plot (click images to download plots)	3D map of model quality
IntFOLD5 SPARKSX multi7 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download Refine model using ReFOLD
IntFOLD5 SPARKSX multi8 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download

Click here to download a PDB file of this model with residue accuracy predictions (Angstroms) in the B-factor column.
[RasMol colouring uses the reverse rainbow scheme from blue (high accuracy) through green, yellow and orange to red (low accuracy).]

JSmol view of the per-residue accuracy for the model IntFOLD5_SPARKSX_multi7_TS1.bfact.pdb

☐ Spin model ☑ High detail (uncheck for faster rendering on phones/tablets)

Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

[JSmol colouring uses a gradient from blue (high accuracy) through white to red (low accuracy).]

JSmol view of the structural alignment of the model (blue) with the template 3ebtA (red) using TM-align

Display options:
● Show superposition ○ Show template only: 3ebt ○ Show model only
☐ Show labels
☑ Spin model ☑ High detail (uncheck for faster rendering on phones/tablets)
Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

Download the superposition file (RasMol script).

Bioinformatics Web Servers

The ReFOLD Server Response (Version 1.0)

Thank you for using this service, your 3D model has been submitted to the ReFOLD refinement pipeline.

When your predictions are completed, your results will be found at the link below. You should bookmark the results page so you can **revisit the link to check on the job status**.

However, if you have provided your email address, then a link to the results page will be sent to you when your predictions are completed.

Typically jobs will be completed well within 24 hours. Please email us if you haven't received anything within 72 hours.

Results page for T0971_SPARKSX_multi7

Tel: 0118 378 6332
Email: l.j.mcguffin@reading.ac.uk
► Full contact details

Accessibility | Site Map | Contact Us | Find us
© University of Reading

JSmol view of ligand binding residues prediction for T0971

Predicted binding residues - display options
☑ High detail (uncheck for faster rendering on phones/tablets)
☐ Show residue labels
☑ Wireframe on
Spacefill options: ● off ○ 25% ○ 100%
Zoom options: ○ 100% ● 200% ○ 300% ○ 400%

Predicted ligands - display options
☑ Wireframe on
Spacefill options: ● off ○ 25% ○ 100%

☐ Spin model

Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

The server makes use of the FunFOLDQA Quality Assessment protocol for improved prediction selection:
Roche, D. B., Buenavista, M. T. & McGuffin, L. J. (2012) FunFOLDQA: A Quality Assessment Tool for Protein-Ligand Binding Site Residue

FunFOLDQA scores:

Figure 4.1. IntFOLD results for CASP13 target T0971
(A) Graphical output from the main results page showing (from top to bottom): 1. The table with the top 5 selected 3D models and scores (table truncated here to fit); 2. The prediction of natively unstructured/disordered regions; 3. The predicted structural domain boundaries; 4. The ligand binding site prediction; 5. The full model quality rankings for all generated models (table truncated here to fit). The arrows point to additional pages that are linked to when users click on images/buttons on the main page. **(B)** Clicking the button titled 'View model in 3D and download' leads to dynamically generated pages showing interactive views of the model, and structural superpositions of the model with relevant template/s, which can be manipulated in 3D using the JSmol/HTML5 framework (http://www.jmol.org/) and/or downloaded for local viewing. **(C)** Clicking the button titled 'Refine model using ReFOLD' submits the 3D model to the ReFOLD service for refinement guided by accurate quality estimates. **(D)** Clicking on the image of the ligand binding site prediction links to a dynamically generated page that provides numerous options for interactively viewing the likely protein–ligand interactions in 3D with JSmol. Figure taken from McGuffin et al., 2019

4.1.1 Analysis of biological and functional relevance of CASP12 predictions from IntFOLD4

The CASP12 assessors performed a systematic assessment to compare the ensembles of predictions of a target protein from different modelling algorithms to quantify the utility of perditions for inferring or recognising function (Roche *et al.*, 2011). The question which was addressed was: to what extent do the CASP predictions accurately provide function information – compared to experimental structures. Regions or sites for assessment were based on experimentalists motivation to solve structures and ultimately help define the term "protein function". The defined regions/sites for assessment were separated into three categories of functional sites: (1) *Holo sites*: pockets based on observed ligand-binding in experimental structures (2) *Apo sites:* sites based on (a) critical residues provided by experimental authors or (b) known motifs relevant to ligand or substrate binding and/or (c) site finding algorithms and (3) critical patches: patches centered at the key residues provided by experimental authors, including functionally critical residues, loops and mutations (Liu *et al.*, 2018)

The assessors evaluated the physical features of the predicted structure sites and the degree to which they share similarity with experimental structure sites (Liu *et al.*, 2018).

Aim: The aim of this section is to analyse the FunFOLD3 predictions from the IntFOLD4 server used in the CASP12 competition in each of the biologically relevant category and determine by critical analysis why the IntFOLD4 server was ranked highly in this category.

4.2 Materials & Methods

4.2.1 Materials

As with functional prediction in CASP11, CASP12 and CASP13, amino acid sequences for the prediction of biologically relevance were provided by the CASP12 organisers and are double-blinded, so neither the predictors nor assessors are aware of the structure at the time of prediction. Amino acid sequences failing into holo, apo or motif, key residues and mutation will form part of the analysis in this section and are presented in Table 4.1. The amino acid sequences below were deemed to be of significance and the targets were published by the CASP committee, hence why the CASP12 targets may differ from those presented in Chapter 3 (Liu *et al.*, 2018).

Table 4.1. Amino acid sequences for biologically relevant categories in CASP12

CASP 12 target ID	Amino acid sequence
T0860	VSYSDGHFLTKSGGVINFRKTRVTSITITILGNYGLRVVNGELQNTPLTFKGADFKSSTLKDELLIPLEGAVQLNTAPSTALCIFITTDHVYRELCMMQFLTDVDKTPFLVVLRSESKHETIQYMHIVTVHPFLSLT
T0861	MGKIFEDNSLTIGHTPLVRLNRIGNGRILAKVESRNPSFSVKCRIGANMIWDAEKRGVLKPGVELVEPTSGNTGIALAYVAAARGYKLTLTMPETMSIERRKLLKALGANLVLTEGAKGMKGAIQKAEEIVASNPEKYLLLQQFSNPANPEIHEKTTGPEIWEDTDGQVDVFIAGVGTGGTLTGVSRYIKGTKGKTDLISVAVEPTDSPVIAQALAGEEIKPGPHKIQGIGAGFIPANLDLKLVDKVIGITNEEAISTARRLMEEEGILAGISSGAAVAAALKLQEDESFTNKNIVVILPSSGERYLSTALFADLFTEKELQQ
T0863	MSAETVNNYDYSDWYENAAPTKAPVEVIPPCDPTADEGLFHICIAAISLVVMLVLAILARRQKLSDNQRGLTGLLSPVNFLDHTQHKGLAVAVYGVLFCKLVGMVLSHHPLPFTKEVANKEFWMILALLYYPTLYYPLLACGTLHNKVGYVLGSLLSWTHFGILVWQKVDCPKTPQIYKYYALFGSLPQIACLAFLSFQYPLLLFKGLQNTETANASEDLSSSYYRDYVKKILKKKKPTKISSSTSKPKLFDRLRDAVKSYIYTPEDVFRFPLKLAISVVVAFIALYQMALLLISGVLPTLHIVRRGVDENIAFLLAGFNIILSNDRQEVVRIVVYYLWCVEICYVSAVTLSCLVNLLMLMRSMVLHRSNLKGLYRGDSLNVFNCHRSIRPSRPLVCWMGFTSYQAAFLCLGMAIQTLVFFICILFAVFLIIIPILWGTNLMLFHIIGNLWPFWLTLVLAALIQHVASRFLFIRKDGGTRDLNNRGSLFLLSYILFLVNVMIGVVLGIWRVVITALFNIVHLGRLDISLLNRNVEAFDPGYRCYSHYLKIEVSQSHPVMKAFCGLLLQSSGQDGLSAQRIRDAEEGIQLVQQEKKQNKVSNAKRARAHWQLLYTLVNNPSLVGSRKHFQCQSSESFINGALSRTSKEGSKKDGSVKEPNKEAESAAASN
T0864	GHMASGPWKLTASKTHIMKSADVEKLADELHMPSLPEMMFGDNVLRIQHGSGFGIEFNATDALRCVNNYQGMLKVACAEEWQESRTEGEHSKEVIKPYDWTYTTDYKGTLLGESLKLKVVPTTDHIDTEKLKAREQIKFFEEVLLFEDELHDHGVSSLSVKIRVMPSSFFLLLRFFLRIDGVLIRMNDTRLYHEADKTYMLREYTSRESKISSLMHVPPSLFTEPNEISQYLPIKEAVCEKLIFPE
T0873	MGSSHHHHHHSQDPNSMKRLKDLREYLAVLEAHQDVREIDEPVDPHLEAGAAARWTYENRGPALMLNDLTGTGRFCRILAAPAGLSTIPGSPLARVALSLGLDVSATAHEIVDSLAAARTREPVAPVVVDSAPCQDNVLLGDDANLDRFPAPLLHEGDGGPYLNTWGTIIVSTPDGSFTNWAIARVMKIDGKRMTGTFIPTQHLGQIRKLWDNLGQPMPFAIVQGTEPGIPFVASMPLPDGIEEVGFLGAYFGEPLLVRAKTVDLLVPASAEIVIEGHVMPGRTAVEGPMGEYAGYQPRHTSMQPEYVVDAITYRDDPIWPISVAGEPVDETHTAWGLVTAAEALALLRAAKLPVATAWMPFEAAAHWLIVCLTEDWRERMPGLSRDGICLRISQVLAATRIEAMMTRVFVLDDDVDPSDQTELAWAIATRVSPAHGRLVRHGMINPLAGCYSAEERRLGYGPKAVLNGLLPPMAERSRRSSFRHTYPEPVRQRVIELLA

T0879	SVYDPAATADTVNPGNKIIYLTFDDGPGKYTQGLLDVLDKYNVKATFFVTNTHPDYQNMIAEEAKRGHTVAIHSASHKYNQIYTSEQAFFDDLEQMNSIIKAQTGNDASIIRFPGGSSNTVSKDYCPGIMTQLVNDVTARGLLYCDWNVSSGDANPKPISTEQVVQNVISGVQSHNVSVVLQHDIKEFSVNAVEQIIQWGQANGYTFLPLTTSSPMSHHRVN
T0880-0	FFTAAPLSYNTGNSTISLDYRSPQLRVSGGALALTSPVFVYQTPFNTPMRLRNGTYNEYADAHIQMVRFGTTVLFNIDVTGETNATGTQTWELQFDGTLGSCLTGRMQVMGGTGEELDVTPTFILPTSDKSVYKQGFMPIVCSENGEFKQSTYCSYALTYRLGNFYITLKSTTSGCKPIFQMSFMYESQIGIV
T0880-1	FFTAAPLSYNTGNSTISLDYRSPQLRVSGGALALTSPVFVYQTPFNTPMRLRNGTYNEYADAHIQMVRFGTTVLFNIDVTGETNATGTQTWELQFDGTLGSCLTGRMQVMGGTGEELDVTPTFILPTSDKSVYKQGFMPIVCSENGEFKQSTYCSYALTYRLGNFYITLKSTTSGCKPIFQMSFMYESQIGIV
T0882	SMTSRPKLRILNVSNKGDRVVECQLETHNRKMVTFKFDLDGDNPEEIATIMVNNDFILAIERESFVDQVREIIEKADEMLSEDVSVEPE
T0889	MARELEGKVAAVTGAASGIGLASAEAMLAAGARVVMVDRDEAALKALCNKHGDTVIPLVVDLLDPEDCATLLPRVLEKACQLDILHANAGTYVGGDLVDADTMAIDRMLNLNVNVVMKNVHDVLPHMIERRTGDIIVTSSLAAHFPTPWEPVYASSKWAINCFVQTVRRQVFKHGIRVGSISPGPVVSALLADWPPEKLKEARDSGSLLEASDVAEVVMFMLTRPRGMTIRDVLMLPTNFDL
T0891	ENMAVQSPKKHVFDAVIKAYKDNSDEESYATVYIKDPKLTIENGKRIITATLKDSDFFDYLKVEDSKEPGVFHDVKVLSEDKRKHGTKVIQFEVGELGKRYNMQMHILIPTLGYDKEFKIQFEVNMRTFV
T0893	LSQAQKMQAIGQLAGGVAHDFNNLLTAIQLRLDQLLHRHPVGDPSYEGLNEIRQTGVRAADLVRKLLAFSRKQTVQREVLDLGELISEFEVLLRRLLREDVKLITDYGRDLPQVRADKSQLETAVMNLAVNARDAVRAAKGGGVVRIRTARLTRDEAIQLGFPAADGDTAFIEVSDDGPGIPPDVMGKIFDPFFTTKPVGEGTGLGLATVYGIVKQSDGWIHVHSRPNEGAAFRIFLPVYEA
T0894	MVDNNYLSVSEKTELEIAKQTLKNSKNPAEREKAQQKYDALLEKDIASDKEVIAACGNGNAGSSACASARLKVIASKEGYEDGPYNSKYSQQYADAYGQIVNLLDITSVDVQNQQQVKDAMVSYFMATLGVDQKTAQGYVETTQGLEIAAASMTPLFGQAVANKITALVDKANKYPSGIGFKINQPEHLAQLDGYSQKKGISGAHNADVFNKAVVDNGVKIISETPTGVRGITQVQYEIPTKDAAGNTTGNYKGNGAKPFEKTIYDPKIFTDEKMLQLGQEAAAIGYSNAIKNGLQAYDAKAGGVTFRVYIDQKTGIVSNFHPK
T0895	MNKYLFELPYERSEPGWTIRSYFDLMYNENRFLDAVENIVNKESYILDGIYCNFPDMNSYDESEHFEGVEFAVGYPPDEDDIVIVSEETCFEYVRLACEKYLQLHPEDTEKVNKLLSKIPSAGHHHHHH
T0896	MNGDYMKASLVGVAAAVLMSVLAGSPVSAQVADASAQVVSKSQLIVGKRYYISVDTLNVRSSNSTTANNVIGKLSKNDVVEVYDVLNEATPLVQVKIIKSTTVSPYISSDFFVSKDYLSERELTLPTSRYFVVQNIATEKTRIYERCTATPGCAHKMVMETDMVVGRPEEGDGQDDNAYKTWVGHSRISEWVKFYQDGKAFYPRWYTPGQNIKDIPDPVTDSMSLYMGARKWLRKNEQGKTSNYGAFGWYAAKLTPAGENGGVNYQWIHGTMGWGKDGSKPIEITRMKMINFFSNPGSHGCTRLENQAVAYMRHLLGPGTDIYRVYAREASREAAPFSRYRDSQRPLPWEWMLLTNGAAQSNGLTADAATIRAQGISAVPGVNLIERGVYQVDRYPTVMPLNYSKSAASGLSGDRYEIDKNLKKGQGSNFRGYFLVDEGRFVSYSHPNYNATGGAIRVGGMADFMDSVPALLQAGAGNYYPPAIIK
T0910	GLDDVSNKAYEDAEAKAKYEAEAAFFANLKLSDFNIIDTLGVGGFGRVELVQLKSEESKTFAMKILKKRHIVDTRQQEHIRSEKQIMQGAHSDFIVRLYRTFKDSKYLYMLMEACLGGELWTILRDRGSFEDSTTRFYTACVVEAFAYLHSKGIIYRDLKPENLILDHRGYAKLVDFGFAKKIGFGKKTWTFCGTPEYVAPEIILNKGHDISADYWSLGILMYELLTGSPPFSGPDPMKTYNIILRGIDMIEFPKKIAKNAANLIKKLCRDNPSERLGNLKNGVKDIQKHKWFEGFNWEGLRKGTLTPPIIPSVASPTDTSNFDSFPEDNDEPPPDDNSGWDIDF
T0911	MVSGFAMPKIWRKLAMDIPVNAAKPGRRRYLTLVMIFITVVICYVDRANLAVASAHIQEEFGITKAEMGYVFSAFAWLYTLCQIPGGWFLDRVGSRVTYFIAIFGWSVATLFQGFATGLMSLIGLRAITGIFEAPAFPTNNRMVTSWFPEHERASAVGFYTSGQFVGLAFLTPLLIWIQEMLSWHWVFIVTGGIGIIWSLIWFKVYQPPRLTKGISKAELDYIRDGGGLVDGDAPVKKEARQPLTAKDWKLVFHRKLIGVYLGQFAVASTLWFFLTWFPNYLTQEKGITALKAGFMTTVPFLAAFVGVLLSGWVADLLVRKGFSLGFARKTPIICGLLISTCIMGANYTNDPMMIMCLMALAFFGNGFASITWSLVSSLAPMRLIGLTGGVFNFAGGLGGITVPLVVGYLAQGYGFAPALVYISAVALIGALSYILLVGDVKRVG
T0913	MHHHHHHMTTVDKRPSSRGYGDWRLSDIPQYKDGISTYEFVRATHEADYRTHQAEPVAGRTFGFNGIGRLTEVALHMPTKYTLHDQSSQYKESPSFFQGLMGVPDRGPVDLAAFQRETEELATAFENNGIKVHWVDYPEEPANPYGPLMGHVFLSWGSIWRGGSVISRFGFLPGMVGVSEYLAKWAWNTLNIPPLVAITEGAMEPGACNMIADEVLVTCLSASYDQRGTDQ

	LVAAISKTSGTEEFHNLQLRPAVEGFFNKATGACAHPDININAIDVGKLVVSPAALDWDARTWLYDNNFELIEADPDEQREFLAPCNVLLLEPGKVIAHADCHKTNQKIRDAGVEVIEVTGTEIRKACGGIKCRVMQINREPGPTLADVRNRVWR
T0914	VENNYLSVSEKTELEIAKQKLKNSKDPAEREKAQQKYDALLEKDISSDKAVITACSNGQAASAACAGERLKVIAAKGGYETGHYNNQVSDMYPDAYGQIVNLLNITSVDAQNQQQVKDAMVNYAMVQFGVDRATAQAYVETYDGMKVVAASMAPVIGAAAASKIEVLAGKQRLSNSFEVSSLPDANGKNHITAVKGDAKIPVDKIELYMRGKASGDLDSLQAEYNSLKDARISSQKEFAKDPNNAKRMEVLEKQIHNIERSQDMARVLEQAGIVNTASNNSMIMDKLLDSAQGATSANRKTSVVVSGPNGNVRIYATWTILPDGTKRLSTVNTGTFK
T0915	MINVNSTAKDIEGLESYLANGYVEADSFNDPEDDALECLSNLLVKDSRGGLSFCKKILKSNNIDGVFIKGSALNFLLLSEQWSYAFEYLTSNADNITLAELEKALFYFYCAKNETDPYPVPEGLFKKLMKRYEELKNDPDAKFYHLHETYNDFSKAYPLNN
T0917	MHHHHHHMRMEFRHNLPSSDIIFGSGTLEKIGEETKKWGDKAILVTGKSNMKKLGFLADAIDYLESAGVETVHYGEIEPNPTTTVVDEGAEIVLEEGCDVVVALGGGSSMDAAKGIAMVAGHSAEERDISVWDFAPEGDKETKPITEKTLPVIAATSTSGTGSHVTPYAVITNPETKGKPGFGNKHSFPKVSIVDIDILKEMPPRLTAITGYDVFSHVSENLTAKGDHPTADPLAIRAIEYVTEYLLRAVEDGEDIKAREKMAVADTYAGLSNTISGTTLRHAMAHPISGYYPDISHGQALASISVPIMEHNIENGDEKTWERYSRIAVALDASKPVDNTRQAASKAVDGLKNLLRSLDLDKPLSELGVEEEKIPEMTEGAFIYMGGGIEANPVDVSKEDVKEIFRKSL
T0920-0	KPVVGVILPFSSAFEDIAVEQQRAVELALAESGSAFEIVFKDGGADVDTAVQAFQDLVRSQENLAAVVSCSSWASSAIHPLAAEKDIFHVAIGSAALKRTEPGHTIRLTVGVQQEQEQLAAYLTDFERIAVLAMDNNLGSSWIRMLEDRFPKQVVAAQEYNPQQMDIAAQLATIKARDSEALVLISAGEAATIAKQARQAGIKAQLVGTRPIQRAEVLAASAFTNGLVYTYPSYNQDHPFMSAFTDRYGLEPGFFGVEAYDLCTTLSRALEQGRQTPKALFEWYAGNTFTGALGKVTFANDGDASYPYIFKKVTESGFRVAEFQFPMLLTQTAQELNAIFKDMDRSVAAAAEQLSTTGLRGDRASAILETLFNENQYAYNCVTVDATGTIVNVAPKQYSSVIGEDISGQEQIIRLHETHQPVLSQAIKMVEGFVGIDLEHPVFDQDGGFIGSVSVLTQPDFFGSIISRKVHNFPVEIFVLQRDGTTIYDVNAEEIGKNAFADPIYDAFPSLKRIARKMVSQAEGEGTYRFQDRHMEHAVAKQLLWTSIGLHGTNYRLALTYGAGEIED
T0920-1	KPVVGVILPFSSAFEDIAVEQQRAVELALAESGSAFEIVFKDGGADVDTAVQAFQDLVRSQENLAAVVSCSSWASSAIHPLAAEKDIFHVAIGSAALKRTEPGHTIRLTVGVQQEQEQLAAYLTDFERIAVLAMDNNLGSSWIRMLEDRFPKQVVAAQEYNPQQMDIAAQLATIKARDSEALVLISAGEAATIAKQARQAGIKAQLVGTRPIQRAEVLAASAFTNGLVYTYPSYNQDHPFMSAFTDRYGLEPGFFGVEAYDLCTTLSRALEQGRQTPKALFEWYAGNTFTGALGKVTFANDGDASYPYIFKKVTESGFRVAEFQFPMLLTQTAQELNAIFKDMDRSVAAAAEQLSTTGLRGDRASAILETLFNENQYAYNCVTVDATGTIVNVAPKQYSSVIGEDISGQEQIIRLHETHQPVLSQAIKMVEGFVGIDLEHPVFDQDGGFIGSVSVLTQPDFFGSIISRKVHNFPVEIFVLQRDGTTIYDVNAEEIGKNAFADPIYDAFPSLKRIARKMVSQAEGEGTYRFQDRHMEHAVAKQLLWTSIGLHGTNYRLALTYGAGEIED
T0942	MFRQLKKNLVATLIAAMTIGQVAPAFADSADTLPDMGTSAGSTLSIGQEMQMGDYYVRQLRGSAPLINDPLLTQYINSLGMRLVSHANSVKTPFHFFLINNDEINAFAFFGGNVVLHSALFRYSDNESQLASVMAHEISHVTQRHLARAMEDQQRSAPLTWVGALGSILLAMASPQAGMAALTGTLAGTRQGMISFTQQNEQEADRIGIQVLQRSGFDPQAMPTFLEKLLDQARYSSRPPEILLTHPLPESRLADARNRANQMRPMVVQSSEDFYLAKARTLGMYNSGRNQLTSDLLDEWAKGNVRQQRAAQYGRALQAMEANKYDEARKTLQPLLAAEPGNAWYLDLATDIDLGQNKANEAINRLKNARDLRTNPVLQLNLANAYLQGGQPQEAANILNRYTFNNKDDSNGWDLLAQAEAALNNRDQELAARAEGYALAGRLDQAISLLSSASSQVKLGSLQQARYDARIDQLRQLQERFKPYTKM
T0943-1	SFADMMKHGLTEADVGITKFVSSHQGFSGILKERYSDFVVHEIGKDGRISHLNDLSIPVDEEDPSEDIFTVLTAEEKQRLEELQLFKNKETSVAIEVIEDTKEKRTIIHQAIKSLFPGLETKTEDREGKKYIVAYHAAGKKALANPRKHSWPKSRGSYCHFVLYKENKDTMDAINVLSKYLRVKPNIFSYMGTKDKRAITVQEIAVLKITAQRLAHLNKCLMNFKLGNFSYQKNPLKLGELQGNHFTVVLRNITGTDDQVQQAMNSLKEIGFINYYGMQRFGTTAVPTYQVGRAILQNSWTEVMDLILKPRSGAEKGYLVKCREEWAKTKDPTAALRKLPVKRCVEGQLLRGLSKYGMKNIVSAFGIIPRNNRLMYIHSYQSYVWNNMVSKRIEDYGLKPVPGDLVLKGATATYIEEDDVNNYSIHDVVMPLPGFDVIYPKHKIQEAYREMLTADNLDIDNMRHKIRDYSLSGAYRKIIIRPQNVSWEVVAYDDPKIPLFNTDVDNLEGKTPPVFASEGKYRALKMDFSLPPSTYATMAIREVLKMDTSIKNQTQLNTTWLR
T0943-2	SFADMMKHGLTEADVGITKFVSSHQGFSGILKERYSDFVVHEIGKDGRISHLNDLSIPVDEEDPSEDIFTVLTAEEKQRLEELQLFKNKETSVAIEVIEDTKEKRTIIHQAIKSLFPGLETKTEDREGKKYIVAYHAAGKKALANPRKHSWPKSRGSYCHFVLYKENKDTMDAINVLSKYLRVKPNIFSYMGTKDKRAITVQEIAVLKITAQRLAHLNKCLMNFKLGNFSYQKNPLKLGELQGNHFTVVLRNITGTDDQVQQAMNSLKEIGFINYYGMQRFGTTAVPTYQVGRAILQNSWTEVMDLILKPRSGAEKGYLVKCREEWAKTKDPTAALRKLPVKRCVEGQLLRGLSKYGMKNIVSAFGIIPRNNRLMYIHSYQSYVWNNMVSKRIEDYGLKPVPGDLVLKGATATYIEEDDVNNYSIHDVVMPLPGFDVIYPKHKIQEAYREMLTADNLDIDNMRHKIRDYSLSGAYRKIIIRPQNVSWEVVAYDDPKIPLFNTDVDNLEGKTPPVFASEGKYRALKMDFSLPPSTYATMAIREVLKMDTSIKNQTQLNTTWLR
T0947	MIAILESEVPCVTKDLMKKLIFPLIALTLVSLQSFAGSKATDRSYKYCDDMTQIDKLLDRSRESVERIQREGLNIERIVVSKDKRQLYLVSGDTLLRTYTVAFGWNFIGHKQFEGDGKTPEGIYSIDYKNPKSQFTKSLHVDYPNKADIAYAKSQGRSPGGDIMIHGLPSNPQKYERISKIHPYDWTLGCIAVTNKEIEEIYALVKERTLVEVCKISPTK
T0948	SRNMKEKLEDMESVLKDLTEEKRKDVLNSLAKCLGKEDIRQDLEQRVSEVLISRELHMEDSDKPLLSSLFNAAGVLVEARAKAILDFLDALLELSEEQQFVAEALEKGTLPLLKDQVKSVMEQNWDELASSPPDMDYDPEARILCALYVVVSILLELAEGPTSVSS

4.2.2 Methods

The IntFOLD method uses a single-template local consensus fold recognition approach to predict protein tertiary structure from sequence (Liu *et al.*, 2018). The original IntFOLD-TS method integrates ModFOLDclust2 into the core of its pipeline (McGuffin *et al.*, 2015). The model quality assessment tool ranks models based on global quality scores. The generated models are assessed using the TM-score programs to generate structural alignment score as TM-score (Buenavista, Roche and McGuffin, 2012), GDT_TS (global distance test total score) scores (Xu and Zhang, 2010) and MaxSub scores (Zemla *et al.*, 2001). These scores show how similar two protein structures are to one another e.g. the model and the experimentally determined native structure (Siew *et al.*, 2000).

As part of this methodology, the ligand-binding residues from the FunFOLD3 component of the IntFOLD server outputs will be analysed. Using the top-ranked IntFOLD model and the functional prediction by FunFOLD3, models with bound ligands the location of the binding site residues and putative interacting ligands will be compared against the actual binding site residues and the ligands as identified by CASP12 and using MCC and BDT scores it will be determined how accurately IntFOLD predicted the ligand binding site predictions in an objective manner.

For targets in which no PDB ID has been provided by the CASP12 organisers, at the time of writing, the ligand-binding residues will be based on the supplementary data provided by the CASP organisers from the publication (Liu *et al.*, 2018).

IntFOLD is a freely available webserver and can be accessed at: http://www.reading.ac.uk/bioinf/IntFOLD/. The only requirement from the users is an amino acid sequence of the protein in question, as single letter codes (McGuffin *et al.*, 2019). Users have the option to provide a name for their prediction and an email address, which is used to provide a notification of the link to the results when predictions are completed (usually within 24 hours but can take up to 72 hours). Alternatively, if users prefer not to be notified via email, the link can be bookmarked to view the results later. Outputs will be a graphical and are shown in Figure 4.1. The graphical output is a table which summaries results as a thumbnail (McGuffin *et al.*, 2019). For functional prediction, downloadable coordinates and interactive 3D views of the protein-ligand interaction can be accessed via the FunFOLD results summary page (refer to Figure 4.1D) (McGuffin *et al.*, 2019). All of the raw data files for the predictions are available to download via the results page (McGuffin *et al.*, 2019).

Table 4.2. CASP12 targets falling into three biologically relevant categories
Twenty-eight sites with nine known ligand-binding site (holo and coloured yellow), nine putative ligand-binding sites (apo and coloured blue) and 10 critical patches surrounding key residues, motifs or mutations (coloured purple). The number of functional centres are in column 2, The types of sites are noted in column 3 as well as ligand IDS were applicable. Table adapted from Buenavista, Roche and McGuffin, 2012

CASP12 target ID	# point	Type
T0861	28	Holo/LLP
T0863	8	Holo/CLR
T0873	24	Holo/FMN
T0879	7	Holo/ZN/B
T0889	21	Holo/SOR
T0891	11	Holo/HEM
T0893	22	Holo/ADP
T0910	27	Holo/ANP
T0911	10	Holo/GCO
T0880-0	14	Apo
T0880-1	13	Apo
T0894	19	Apo
T0895	21	Apo
T0896	23	Apo
T0913	13	Apo
T0917	73	Apo
T0942	10	Apo
T0947	25	Apo
T0860	17	Motif
T0864	11	Key residues
T0882	11	Key residues
T0914	26	Key residues
T0915	14	Key residues
T0920-0	31	Key residues
T0920-1	14	Key residues
T0943-1	9	Motif
T0943-2	10	Motif
T0948	15-19	Mutation

4.3 Results

4.3.1 Summary of results

The main findings of the Chapter are given below:

- Out of the three functional site categories; holo, apo and critical patches. IntFOLD4 performed the best in the holo category based on the mean MCC and BDT scores. For comparison, the mean MCC and BDT scores for the holo category were 0.345 and 0.39, respectively. The mean MCC and BDT scores for the apo category were 0.14 and 0.094, respectively finally for the critical patches category the MCC and BDT scores were 0.17 and 0.11, respectively
- The highest MCC and BDT scores was achieved for CASP12 protein target T0879 (PDB ID 5jmu) a peptidoglycan N-acetylglucosamine deacetylase and was 0.571 and 0.750, respectively
- Of 28 categorised protein targets, six were classified as FM with the rest being TBM. IntFOLD4 made functional predictions for one of the FM target (T0863 PDB ID STRA6 receptor) with the remaining 18 predictions for TBM targets

Table 4.3. Predicted and observed ligand-binding site residues for holo CASP12 targets
Correct ligand-binding site residues are depicted in red and bold and presented in ascending holo CASP12 target ID with associated PDB IDs in parenthesis, where applicable

CASP 12 target ID	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0861 **(PDB ID 5j5v)**	**42**, 69, 70, **72**, 120, 147, 150, 154, **175**, **176**, **177**, **178**, **179**, **180**, **181**, 187, 228, **229**, **273,300, 301**	38, 40, 41, 42, 43, 45, 46, 71, 72, 73, 76, 143, 149, 153, 175, 176, 177, 178, 179, 180, 181, 182, 229, 230, 273, 300, 301, 306	38, 40, 41, 43, 45, 46, 71, 73, 76, 143, 149, 153, 182, 230, 306	69, 70, 120, 147, 150, 154, 187, 228
T0863 **(PDB ID 5syl)**	76, 77, 80,116,117, 120, 123, 157, 158, 159, 220	416, 419, 420, 423, 424, 516, 519, 520	416, 419, 420, 423, 424, 516, 519, 520	76, 77, 80,116,117, 120, 123, 157, 158, 159, 220
T0873 **(PDB ID 6da6)**	**165, 166, 180**, **183, 185, 199, 202, 203, 234, 235,** 236, **237, 244, 324**, 332, **337, 402,** 406	165, 166, 180, 182, 183, 184, 185, 197, 199, 202, 203, 204, 234, 235, 237, 244, 324, 326, 333, 336, 337, 341, 402	182, 184, 197, 204, 326, 333, 336, 341	236, 332, 406
T0879 **(PDB ID 5jmu)**	**24, 25, 73, 77**	24, 25, 73, 77, 79, 114, 183	79, 114, 183	N/A
T0889 **(PDB ID 5jo9)**	14, 16, 17, 18, 19, 38, 39, 40, 60, 61, 62, 63, 88, 89, 90, 91, 111, 138, 139, **140, 153,** 157, **183, 184, 185, 186, 188,** 189, **190**	92, 140, 141, 142, 147, 149, 150, 153, 183, 184, 185, 186, 188, 190,194, 240, 242	92, 141, 142, 147, 149, 153, 194, 240, 242	14, 16, 17, 18, 19, 38, 39, 40, 60, 61, 62, 63, 88, 89, 90, 91, 111, 138, 139, 157, 189,
T0891 **(PDB ID 4ymp)**	**21, 28, 29,** 32, **33,** 57, 58, 105, 107, 109, **114,** 118, 120	21, 24, 26, 28, 29, 33, 114, 116	24, 26, 29, 116	32, 57, 58, 105, 107, 109, 118, 120
T0893 **(PDB ID 5ldj)**	62, 127, **131, 132, 135, 179, 181, 197,** 199, 201, **203, 204, 205, 207, 231, 233**	131, 132, 135, 176, 178, 179, 180, 181, 189, 194, 195, 196, 197, 202, 203, 204, 205, 206, 207, 208, 231, 233	176, 178, 180, 189, 194, 195, 196, 202, 208,	62, 127, 199, 201
T0910	31, 331	40, 41, 42, 43, 44, 45, 46 ,48 ,62, 64, 83, 96, 112, 114, 115, 116, 119, 158, 160, 163, 165,175, 176, 178, 179, 195, 323	40, 41, 42, 43, 44, 45, 46 ,48, 62, 64, 83, 96, 112, 114, 115, 116, 119, 158, 160, 163, 165, 175, 176, 178, 179, 195, 323	31, 331
T0911 **(PDB ID 6e9n)**	44, 160, 164, 165, 168, 271, 272, 301, 337, 349, 353; 366, 393	68, 123, 126, 358, 371, 374, 375, 377, 378	68, 123, 126, 358, 371, 374, 375, 377, 378	44, 160, 164, 165, 168, 271, 272, 301, 337, 349, 353; 366, 393

Table 4.4. Predicted and observed ligand-binding site residues for apo CASP12 targets
Correct ligand-binding site residues are depicted in red and bold and presented in ascending apo CASP12 target ID with associated PDB IDs in parenthesis, where applicable

CASP 12 target ID	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0880-0	No ligands-binding site residues predicted	647,650,652,679,680,681,682	N/A	N/A
T0880-1	No ligands-binding site residues predicted	53,56,58,83,85,86,87,88,89,90,91,173	N/A	N/A
T0894 (PDB ID 5hkq)	No ligands-binding site residues predicted	188,189,193,199,201,202,203,204,205,261,262,263,308,310,320,321,322,323	N/A	N/A
T0895 (PDB ID 5hkq	No ligands-binding site residues predicted	46,47,48,49,50,51,52,52,63,70,71,72,73,74,76,77,78,81,82,83	N/A	N/A
T0896	206,261,262,263,264,265,266,274,299,301	131,144,145,146,147,148,149,150,151,153,154,155,156,339,437,438,439,440,471,472,473	131,144,145,146,147,148,149,150,151,153,154,155,156,339,437,438,439,440,471,472,473	206,261,262,263,264,265,266,274,299,301
T0913	100,149,153,156,171,206,266,267,318,358,359,364	64,65,66,67,209,210,273,274,320,321,363,368,371	64,65,66,67,209,210,273,274,320,321,363,368,371	100,149,153,156,171,206,266,267,318,358,359,364
T0917	**48,50,106,107,108,157,158,161,163,166,168,198,201,202,206,213,217,282,297**	46,47,48,49,50,51,52,54,57,73,74,75,76,77,78,79,80,81,82,84,85,88,104,105,106,107,108,109,111,136,157,158,160,161,163,166,168,170,172,179,181,198,201,202,203,206,209,210,213,217,273,276,277,278,279,282,286,293,294,295,296,297,298,383,384,385,386,387,388,391	46,47,49,51,52,54,57,73,74,75,76,77,78,79,80,81,82,84,85,88,104,105,109,111,136,160,170,172,179,181,203,209,210,273,276,277,278,279,286,293,294,295,296,298,383,384,385,386,387,388,391	N/A
T0942 (PDB ID 5xi8)	317,320,348,351,354	53,56,58,83,85,86,87,88,89,90,91,173	136,137,139,140,145,196,200,201,204,246	317,320,348,351,354
T0947	No ligand-binding site residues predicted	188,189,193,199,201,202,203,204,205,261,262,263,308,310,320,321,322,323	N/A	N/A

Table 4.5. Predicted and observed ligand-binding site residues for motif, key residues and mutated CASP12 targets
Correct ligand-binding site residues are depicted in red and bold and presented in ascending motif, key residue and mutated CASP12 target ID with associated PDB IDs in parenthesis, where applicable

CASP12 target ID	Predicted ligand-binding site residue	Observed ligand-binding site residue	Under-predictions	Over-predictions
T0860 (PDB ID 5fjl)	93	33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,128	33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,128	93
T0864 (PDB ID 5d9g)	No ligand-binding site residues predicted	1,2,3,9,10,11,12,13,45,46,47	N/A	N/A
T0882 (PDB ID 5g3g)	34	45,47,54,59,62,63,68	45,47,54,59,62,63,68	34
T0914 (PDB ID 6cp8)	48,49,50,51,55,56,58,59,60,62,63,64,65,66,67,68,69,70,71,72,73,109,112,113,115,116,117,119,126	220,221,222,223,224,225,226,227,228,229,231,238,244,245,246,247,248,249,250,251,252,253,254,255,256,259	220,221,222,223,224,225,226,227,228,229,231,238,244,245,246,247,248,249,250,251,252,253,254 255,256,259	48,49,50,51,55,56,58,59,60,62,63,64,65,66,67,68,69,70,71,72,73,109,112,113,115,116,117,119,126
T0915 (PDB ID 6cp8)	64,65,66,67,70,100,103,107	13,14,17,18,23,35,39,53,56,57,60,63,68	13,14,17,18,23,35,39,53,56,57,60,63,68	64,65,66,67,70,100,103,107
T0920-0 (PDB ID 5ere)	**17,70,71,72, 93,94,95,138**,142, 210,**211,255**	9,12,14,17,44,45,46,47,70,71,72,73,93,94,95,96,109,110,115,134,135,136,138,161,162,163,187,210,211,255	9,12,14, 44,45,46,47,73,96,109,110,115,134,135,136,161,162,163,187	142, 210
T0920-1 (PDB ID 5ere)	No ligand-binding site residues predicted	380,382,390,406,409,428,429,430,431,432,437,454,456,439	N/A	N/A
T0943-1 (PDB ID 5kkp)	No ligand-binding site residues predicted	98,99,101,106,107,108,109,110,111	N/A	N/A
T0943-2 (PDB ID 5kkp)	No ligand-binding site residues predicted	128,129,132,133,134,135	N/A	N/A
T0948-0 (PDB ID 5kkp)	No ligand-binding site residues predicted	47,48,49,50,51,52,53,54,55,56,69,70,74,75,76,77,78,81	N/A	N/A
T0948-1 (PDB ID 5kkp)	No ligand-binding site residues predicted	29,46,56,57,58,59,60,61,62,63,64,65,66,67,68	N/A	N/A
T0948-2 (PDB ID 5tj4)	No ligand-binding site residues predicted	46,47,48,49,50,51,52,53,54,55,56,57,60,70,74,75,76,77,78	N/A	N/A
T0948-3 (PDB ID 5tj4)	No ligand-binding site residues predicted	29,46,50,56,57,58,59,60,61,62,63,64,65,66,67,68	N/A	N/A

A summary of the MCC and BDT scores achieved for the holo, apo and motif/key residues/mutations are provided in Tables 4.6, 4.7 and 4.8, respectively.

Table 4.6. MCC and BDT scores for holo structure CASP12 targets
A list of CASP12 targets with associated MCC and BDT scores. Results are listed from ascending to descending order by MCC and BDT score

CASP12 target	MCC Score	BDT Score
T0879	0.571	0.750
T0873	0.673	0.648
T0893	0.584	0.610
T0861	0.497	0.543
T0891	0.448	0.510
T0889	0.29	0.36
T0911	0.03	0.05
T0863	0.014	0.002
T0910	0.004	0.02

Table 4.7. MCC and BDT scores for apo structure CASP12 targets
A list of CASP12 targets with associated MCC and BDT scores. Results are listed from ascending to descending order by MCC and BDT score

CASP12 target	MCC Score	BDT Score
T0917	0.488	0.275
T0913	0.03	0.09
T0896	0.034	0.009
T0942	0.0186	0.0039

Table 4.8. MCC and BDT scores for key, motifs and mutation structure CASP12 targets
A list of CASP12 targets with associated MCC and BDT scores. Results are listed from ascending to descending order by MCC and BDT score

CASP12 target	MCC Score	BDT Score
T0920-0	0.57	0.37
T0914	0.09	0.015
T0915	0.013	0.06
T0882	0.0035	0.008
T0860	0.0031	0.0055

Figure 4.2 on the next page shows the relation between the MCC and BDT scores to the TM-align score for the different CASP12 targets across the three categories for ligand presence within a structure. Only CASP12 targets with an observed structure have been depicted.

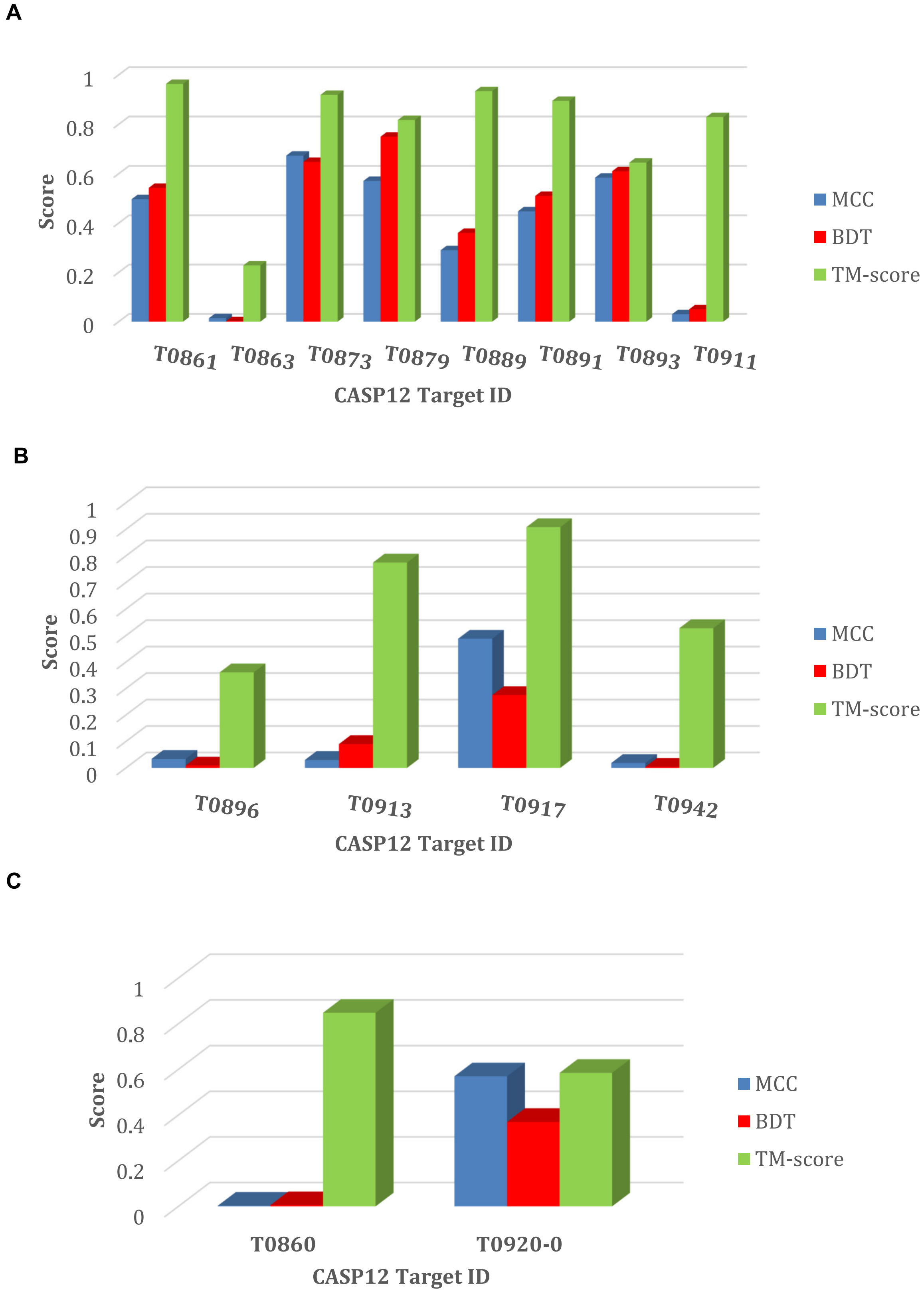


Figure 4.2. Comparison between BDT, MCC and TM-score for CASP12 targets using IntFOLD4
(A) MCC. BDT and TM-scores for the holo targets **(B)** MCC, BDT and TM-scores for Apo targets and **(C)** key, motifs and mutatation targets. Only targets with an observed structure as released by CASP12 organsiers have been included

4.3.2 Holo structures: ligands present in crystal

CASP12 identified nine targets which had a ligand present in crystal. These were T0861 (PDB ID 5j5v), T0863 (PDB 5sy1), T0873 (PDB ID 6da6), T0879 (PDB ID 5jmu), T0889 (PDB ID 5jo9), T0891 (PDB ID 4ymp), T0893 (PDB ID 5idj), T0910 and T0911 (PDB ID 6e9n). CASP12 target IDs are given outside parenthesis and PDB IDs, where applicable within parenthesis.

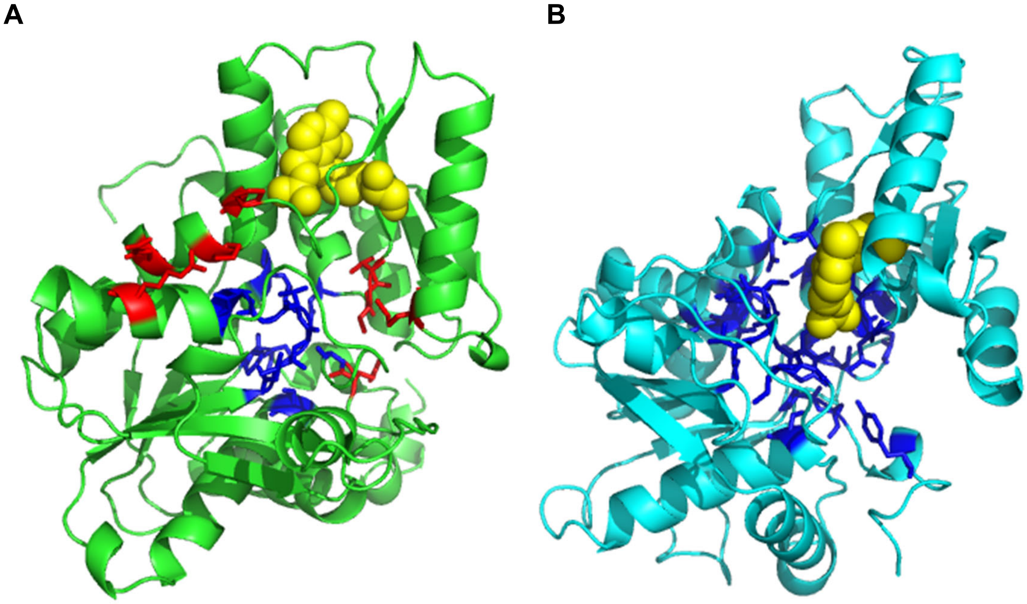


Figure 4.3. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0861 (PDB ID 5j5v)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand pyridoxal-5'-phosphate (PLP) is shown as sphere and coloured yellow. BDT score of 0.543 and MCC score of 0.497, respectively **(B)** The observed ligand binding site for T0861 with binding site residues shown as sticks and coloured blue and the ligand (2S)-2-amino-6-[[3-hydroxy-2-methyl-5-(phosphonooxymethyl)pyridin-4-yl]methylideneamino]hexanoic acid (LLP) shown as sphere and coloured yellow

The first predicted CASP12 target with a ligand in the crystal structure is T0861 and the MCC and BDT score was 0.497 and 0.543, respectively. There were 14 under predictions and eight over predictions, with the IntFOLD4 server predicting 21 ligand binding residues and the observed structure containing 28 ligand binding residues (as shown in Figure 4.3). Of these 21 residues, 13 were correct predictions. Additionally, IntFOLD4 predicted PLP and protoporphyrin IX containing Fe (HEM) and the observed biologically relevant ligand is LLP.

Figure 4.4 below shows the TM-align superposition of the observed and predicted structure. A TM-score of 0.96346, showing very good structural homology, with the main difference being seen in the alpha helices which has been modelled in the predicted 3D model (red) but is not present in the observed 3D model (blue).

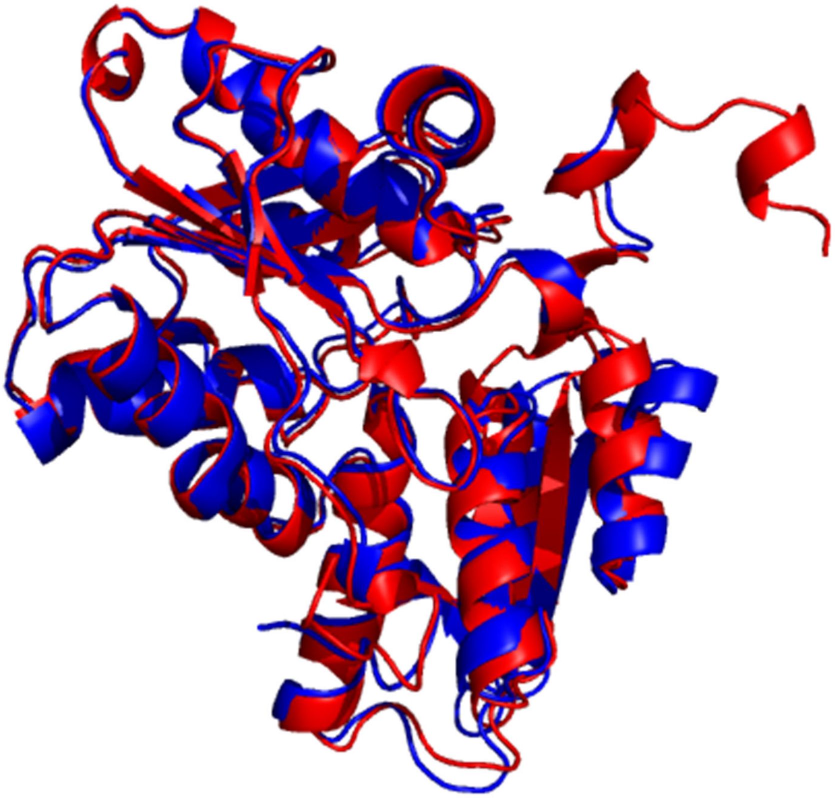


Figure 4.4. Comparison of TM-align (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0861 (PDB ID 5j5v)
The structure in blue is the observed structure for T0861 and the predicted structure is in red. A TM-score of 0.96346 was achieved for the protein structures. The score was normalised for the observed structure T0861 as it is the reference molecule

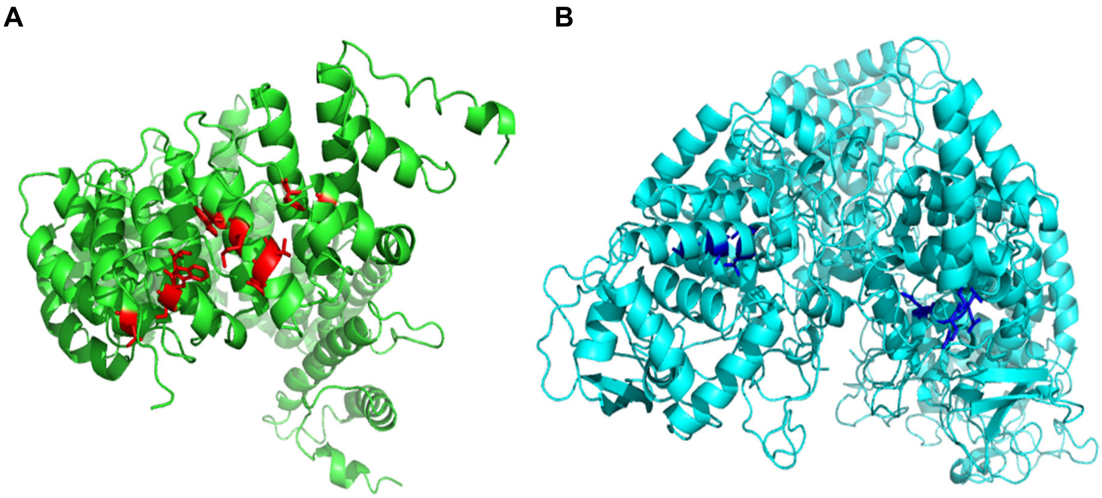


Figure 4.5. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0863 (PDB ID 5syl) **(A)** Predicted ligand binding site residues shown as sticks with under or incorrect predictions shown in red. The predicted ligands were ASN, THR and LEU but are not shown in the figure. BDT and MCC score 0.002 and 0.014, respectively **(B)** The observed ligand binding site for T0863 with binding site residues shown as sticks and coloured blue, no ligand is shown as the ligand CLR is missing from the PDB file

The next predicted CASP12 target is T0863 and scored quite poorly with regards to BDT and MCC with a score of 0.002 and 0.014, respectively. None of the residues were correctly predicted and IntFOLD4 predicted 11 residues with there being eight residues to predict. Furthermore, no correct biologically relevant ligands were predicted with IntFOLD4 predicting ASN, THR and LEU as biologically relevant ligands but are indeed amino acids. The recognised biologically relevant ligand is cholesterol, however this was not in the PDB file as a HETATM and is therefore not shown in Figure 4.5B.

Figure 4.6 below shows the TM-align, superposition of the observed and predicted structure. A TM-score of 0.22810 was achieved showing poor structural homology and this could explain why a poor and MCC and BDT score was achieved.

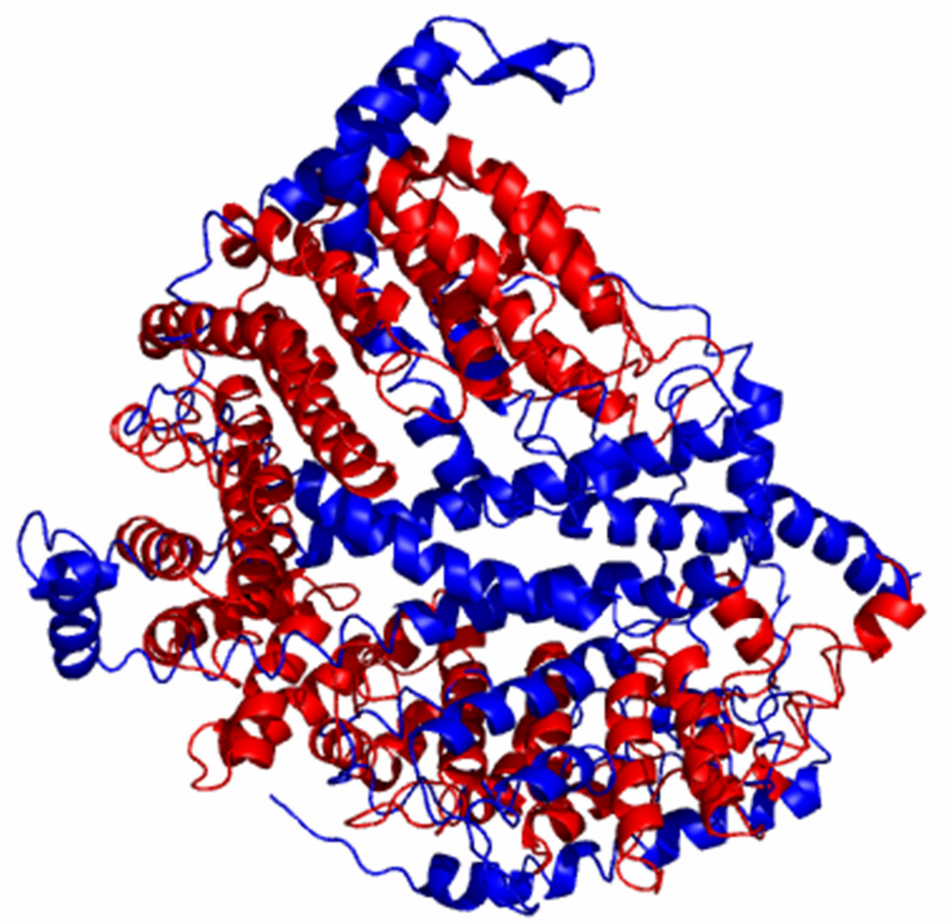


Figure 4.6. Comparison of TM-align (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0863 (PDB ID 5syl)
The structure in blue is the observed structure for T0863 and the predicted structure is in red. A TM-score of 0.22810 was achieved for the protein structures. The score was normalised for the observed structure T0861 as it is the reference molecule

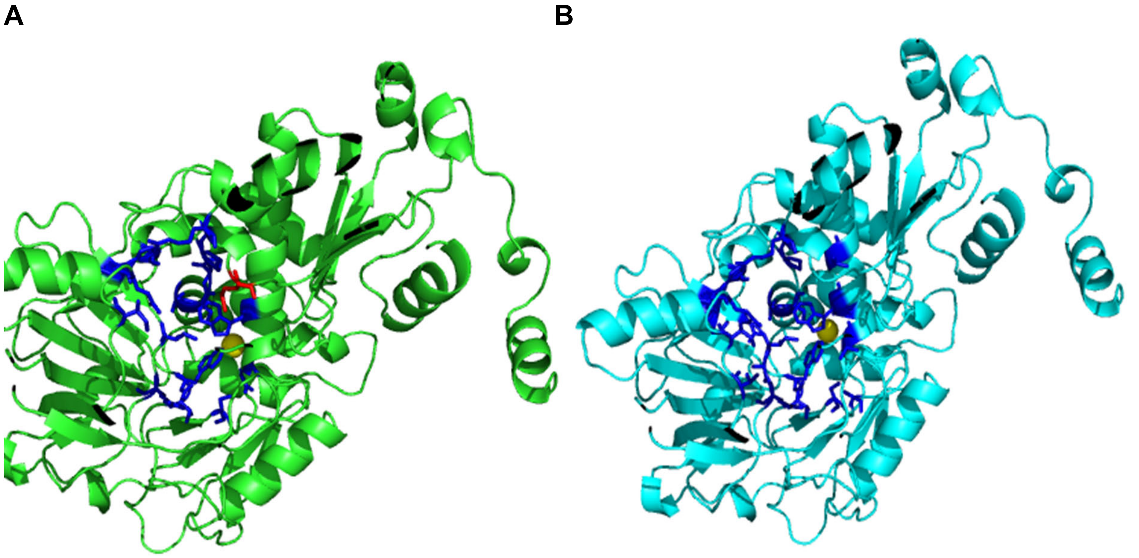


Figure 4.7. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0873 (PDB ID 6da6) **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand manganese (MN) is shown as sphere and coloured yellow. BDT score of 0.648 and MCC score of 0.673, respectively **(B)** The observed ligand binding site for T0873 with binding site residues shown as sticks and coloured blue, the predicted ligand flavin mononucleotide (FMN) is shown as sphere and coloured yellow.

The third predicted CASP12 target with a ligand in the crystal structure is T0873 and the MCC and BDT score was 0.673 and 0.648, respectively. There were five under predictions and two over predictions, with the IntFOLD4 server predicting 18 ligand binding residues and the observed structure containing 24 ligand binding residues. Of these 24 residues, 15 were correct predictions and is shown in Figure 4.7A. Additionally, IntFOLD4 predicted MN and the observed biologically relevant ligand is FMN.

Figure 4.8 below shows the TM-align, superposition of the observed and predicted structure. A TM-score of 0.91925 was achieved showing very good structural homology.

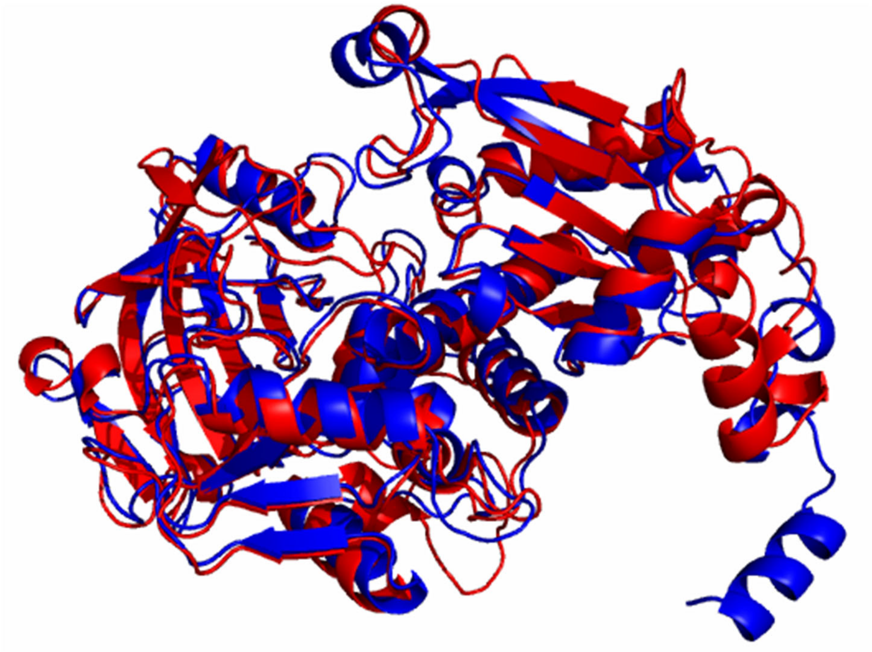


Figure 4.8. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0873 (PDB ID 6da6)
The structure in blue is the observed structure for T0873 and the predicted structure is in red. A TM-score of 0.91925 was achieved for the protein structures. The score was normalised for the observed structure T0873 as it is the reference molecule

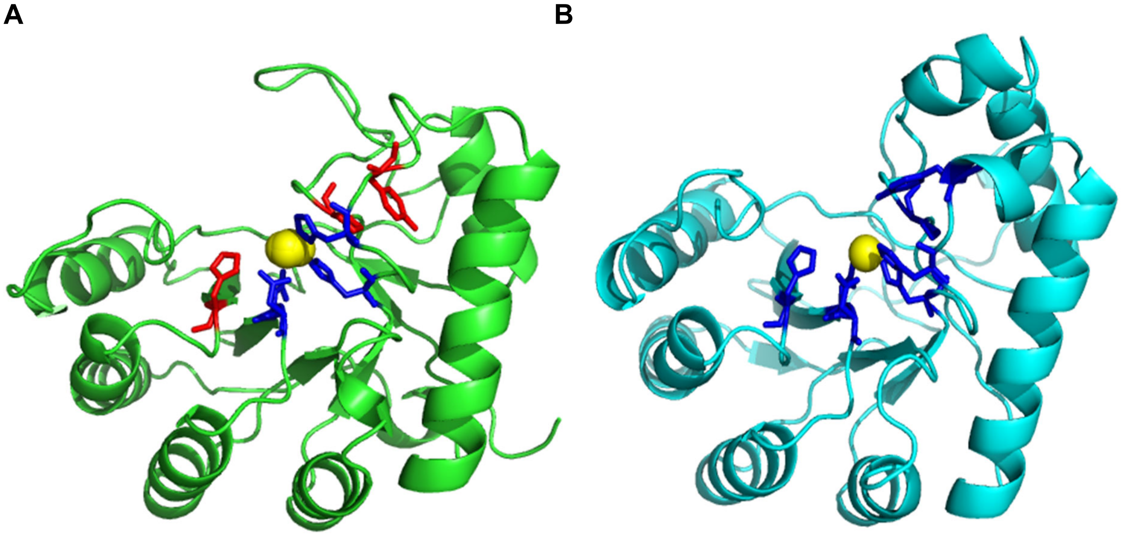


Figure 4.9. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0879 (PDB ID 5jmu)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand zinc (ZN) is shown as sphere and coloured yellow. BDT score of 0.750 and MCC score of 0.571, respectively **(B)** The observed ligand binding site for T0879 with binding site residues shown as sticks and coloured blue, the predicted ligand ZN is shown as sphere and coloured yellow.

The fourth predicted CASP12 target with a ligand in the crystal structure is T0879 and the MCC and BDT score was 0.571 and 0.750, respectively. There were three under predictions and no over predictions, with the IntFOLD4 server predicting four ligand binding residues and the observed structure containing seven ligand binding residues (as shown in Figure 4.9). Of these seven observed residues , four were correct predictions in effect, all the residues predicted by IntFOLD4 were correct predictions. Additionally, IntFOLD4 predicted the observed biologically relevant ligand ZN.

Figure 4.10 below shows the TM-align, superposition of the observed and predicted structure. A TM-score of 0.81721 was achieved showing very good structural homology. The differences between the two structures appears to be modelling of an extra beta sheet in the predicted 3D model (red) and one alpha helix being missed off, which is in the observed structure (blue).

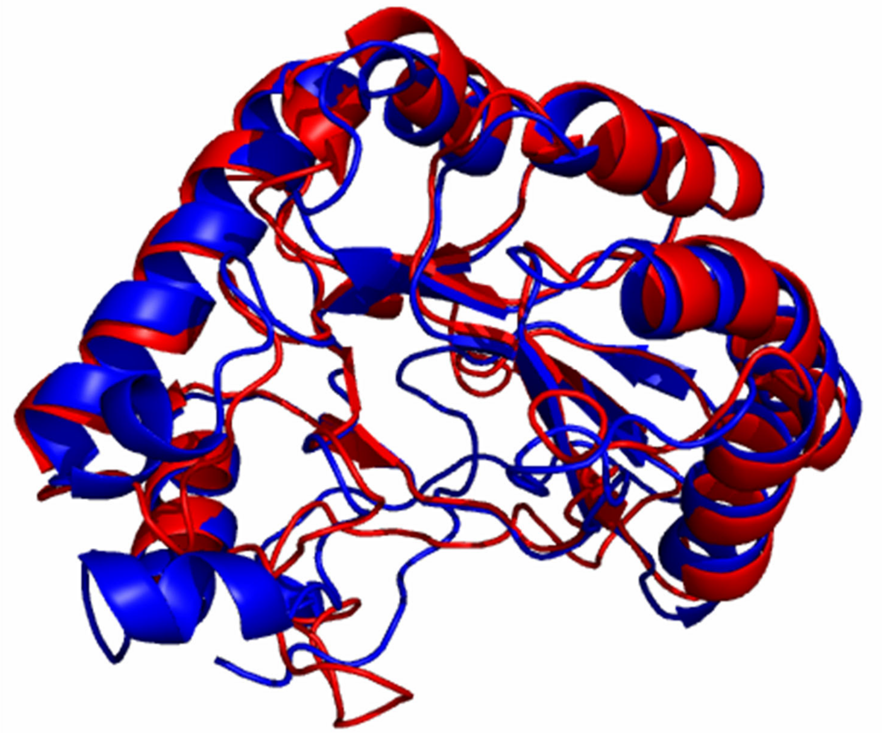


Figure 4.10. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0879 (PDB ID 5jmu)

The structure in blue is the observed structure for T0879 and the predicted structure is in red. A TM-score of 0.81721 was achieved for the protein structures. The score was normalised for the observed structure T0879 as it is the reference molecule

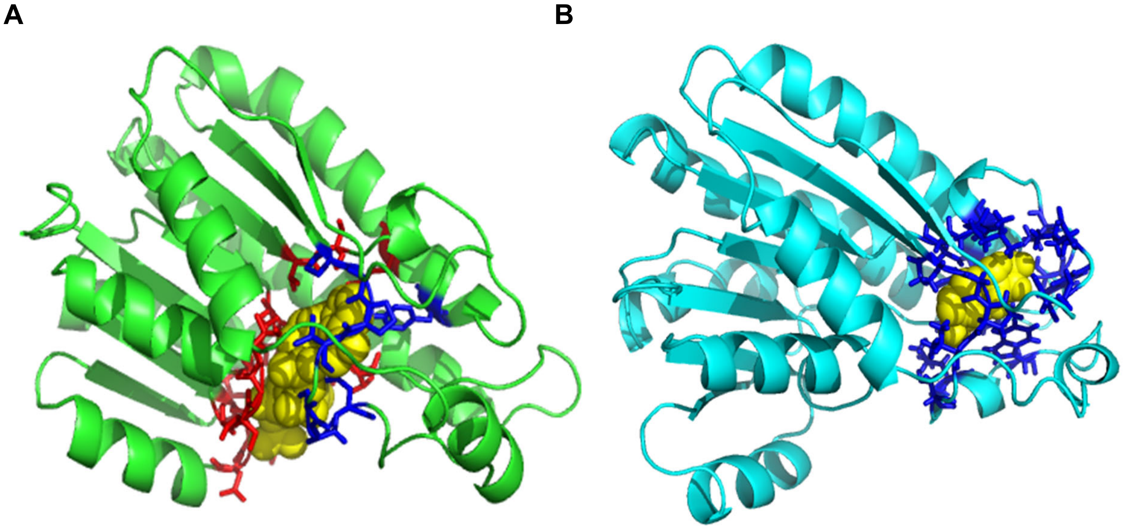


Figure 4.11. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0889 (PDB ID 5jo9)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand NADP (nicotinamide-adenine-dinucleotide phosphate) is shown as sphere and coloured yellow. BDT score of 0.36 and MCC score of 0.29, respectively **(B)** The observed ligand binding site for T0889 with binding site residues shown as sticks and coloured blue, the predicted ligand D-sorbitol (SOR) is shown as sphere and coloured yellow.

The fifth predicted CASP12 target with a ligand in the crystal structure is T0889 and the MCC and BDT score was 0.29 and 0.36, respectively. There were nine under predictions and 20 over predictions, with the IntFOLD4 server predicting 29 ligand binding residues and the observed structure containing 21 ligand binding residues. Of these 29 residues predicted by IntFOLD4, eight were correct predictions and is shown in Figure 4.11A.

Figure 4.12 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.93428 was achieved, showing very good structural homology. The MCC and BDT score is due to the prediction of the ligand-binding pocket space, the observed ligand “sits” within the flexible loops of the observed ligands, whereas with the predicted protein model, the ligand appears to be within the beta sheet and alpha helices of the model. Additionally, the observed ligand SOR has a molecular weight of 182.17 and the predicted NADP ligand has a molecular weight of 743.40 and thus occupies a bigger ligand-binding space.

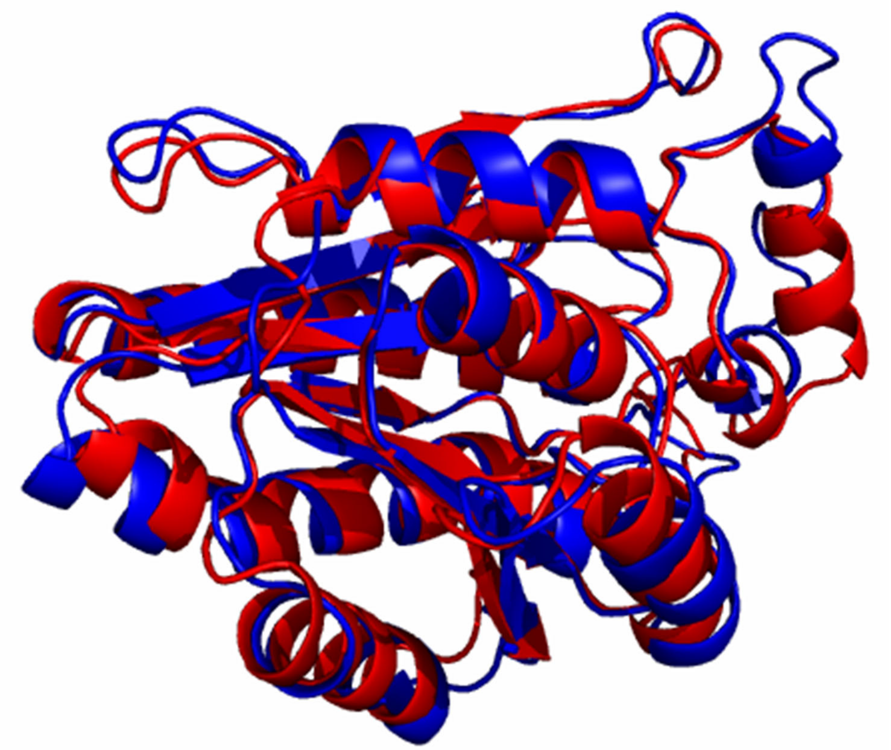


Figure 4.12. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0889 (PDB ID 5jo9)
The structure in blue is the observed structure for T0889 and the predicted structure is in red. A TM-score of 0.93428 was achieved for the protein structures. The score was normalised for the observed structure T0889 as it is the reference molecule

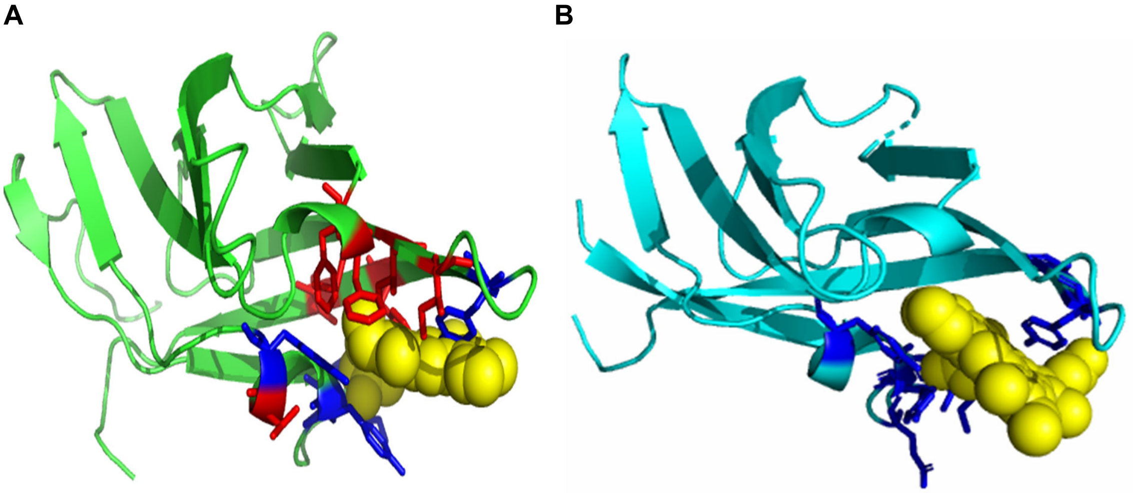


Figure 4.13. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0891 (PDB ID 4ymp) **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand protoporphyrin IX containing Fe (HEM) is shown as sphere and coloured yellow. BDT score of 0.510 and MCC score of 0.448, respectively **(B)** The observed ligand binding site for T0891 with binding site residues shown as sticks and coloured blue, the predicted ligand HEM is shown as sphere and coloured yellow.

The sixth predicted CASP12 target with a ligand in the crystal structure is T0891 and the MCC and BDT score was 0.448 and 0.510, respectively. There were three under predictions and eight over predictions, with the IntFOLD4 server predicting 13 ligand binding residues and the observed structure containing 11 ligand binding residues. Of these 13 residues predicted by IntFOLD4, five were correct predictions from the observed CASP12 target. Additionally, IntFOLD4 predicted the same ligand as present in the observed target structure and is shown in Figure 4.13.

Figure 4.14 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.89491 was achieved, showing very good structural homology. The predicted model had the same number of beta sheets and alpha helices as the observed structure, the main difference between the structures appears to be the flexible loop in the predicted model (red) .

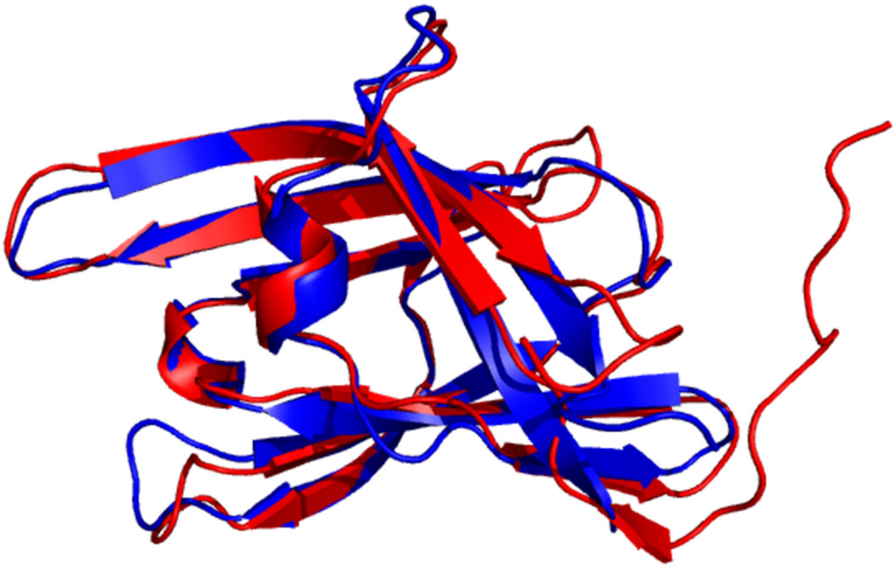


Figure 4.14. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0891 (PDB ID 4ymp)
The structure in blue is the observed structure for T0891 and the predicted structure is in red. A TM-score of 0.89491 was achieved for the protein structures. The score was normalised for the observed structure T0891 as it is the reference molecule

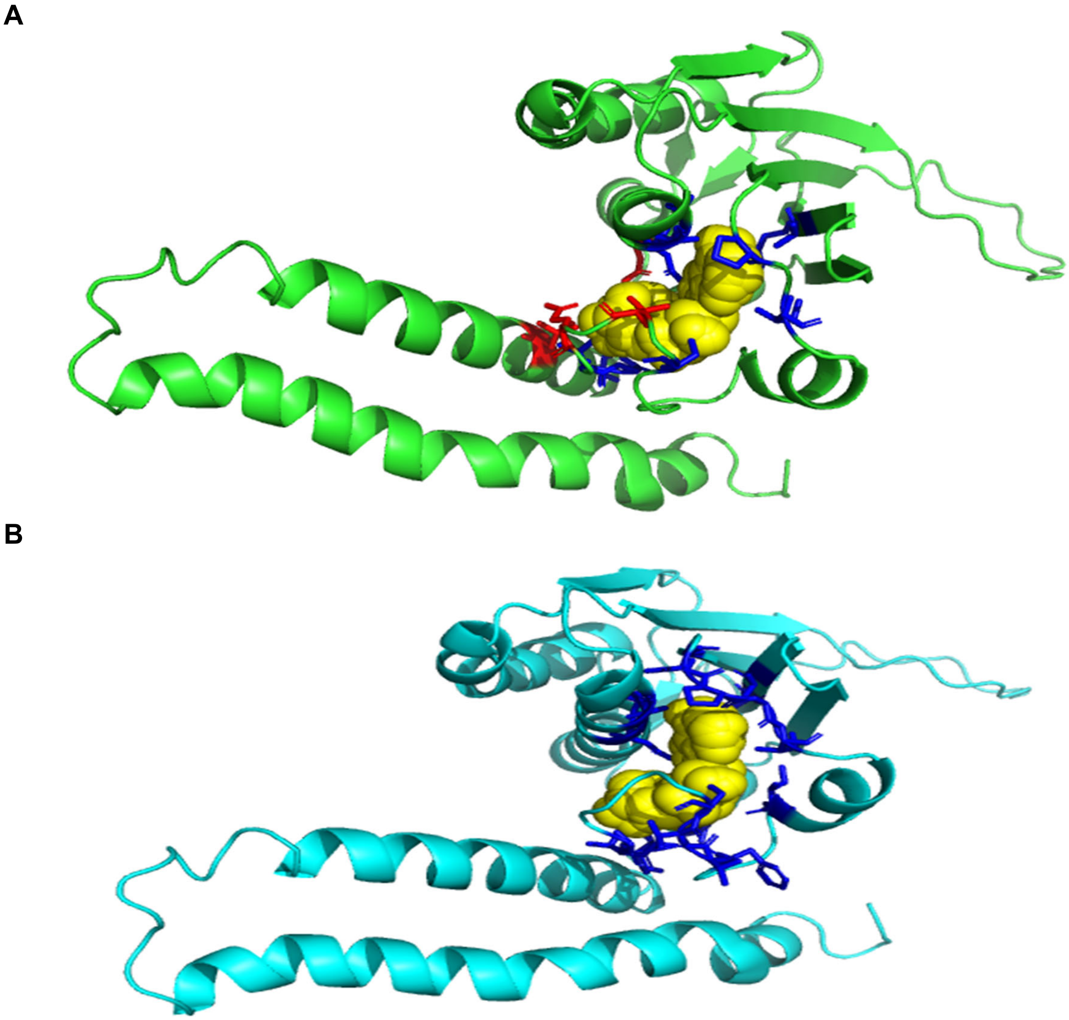


Figure 4.15. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0893 (PDB ID 5ldj)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the predicted ligand adenosine-5'-phosphate (ADP) is shown as sphere and coloured yellow. BDT score of 0.610 and MCC score of 0.584 respectively **(B)** The observed ligand binding site for T0893 with binding site residues shown as sticks and coloured blue, the predicted ligand ADP is shown as sphere and coloured yellow.

The seventh predicted CASP12 target with a ligand in the crystal structure is T0893 and the MCC and BDT score was 0.584 and 0.610, respectively. There were nine under predictions and four over predictions, with the IntFOLD4 server predicting 16 ligand binding residues and the observed structure containing 22 ligand binding residues. Of these 16 residues predicted by IntFOLD4, 12 were correct predictions from the observed CASP12 target. Additionally, IntFOLD4 predicted the same ligand as present in the observed target structure. The predicted structure has been shown for the observed ligand binding site residues (Figure 4.15) as the observed ligand, ADP was not identified in the PDB file for the crystal structure. The observed ligand binding residues were obtained from the supplementary data published by the CASP organisers (Liu *et al.*, 2018).

Figure 4.16 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.64494 was achieved, showing good structural homology. The main difference between the two models can be seen by the rotation of the alpha helices in the bottom half of the model.

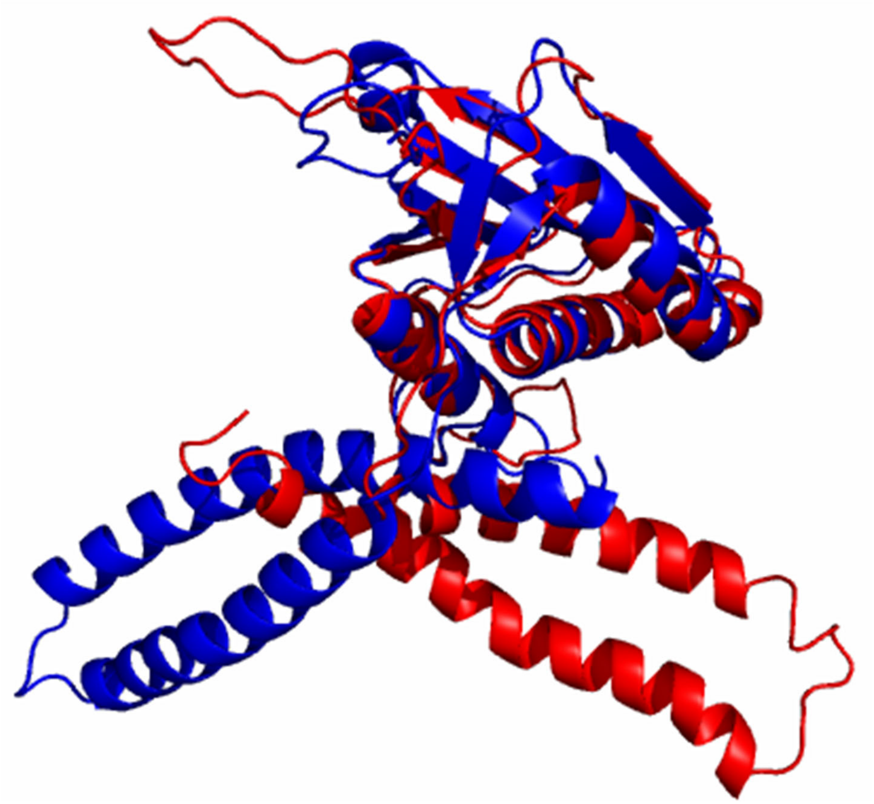


Figure 4.16. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0893 (PDB ID 5ldj)
The structure in blue is the observed structure for T0893 and the predicted structure is in red. A TM-score of 0.64494 was achieved for the protein structures. The score was normalised for the observed structure T0893 as it is the reference molecule

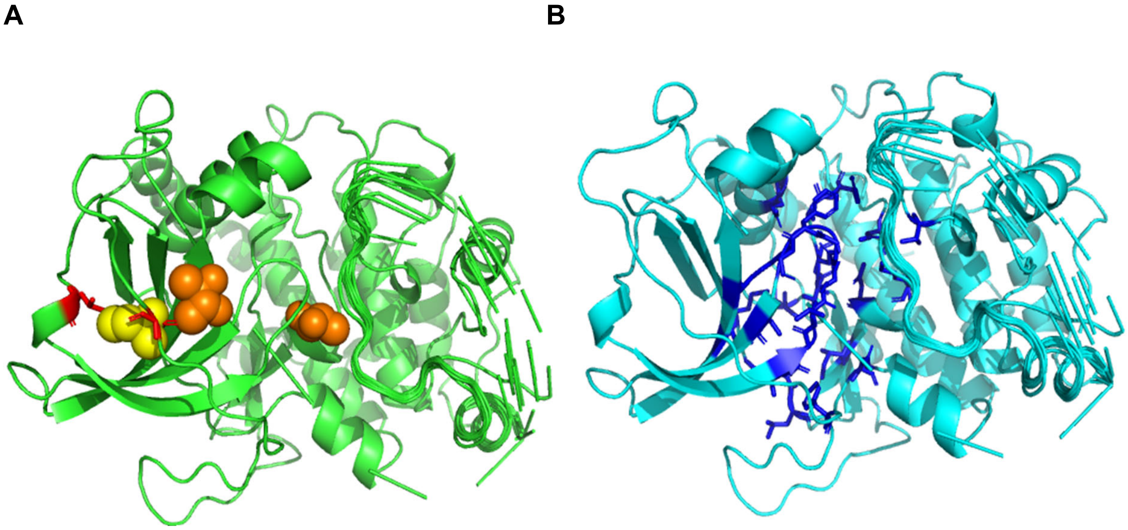


Figure 4.17. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0910
(A) Predicted ligand binding site residues shown as sticks with incorrect -predictions in red, the predicted ligand SBT shown as sphere and coloured yellow. The centroid ligands, BUD and BU3 are shown as sphere and coloured orange. A BDT score of 0.02 and MCC score of 0.004 was obtained **(B)** The observed ligand binding site for T0910 with binding site residues shown as sticks and coloured blue no observed ligands are shown as the observed structure was not released by CASP12 organisers. The protein model is the one predicted by the IntFOLD4 server and is for illustrative purposes only

The eighth predicted CASP12 target with a ligand in the crystal structure is T0910 and the MCC and BDT score was 0.004 and 0.02, respectively. For this particular CASP12 target, there were no correct predictions and IntFOLD4 predicted two residues, whereas there are 27 observed ligand binding site residues. The ligands predicted by IntFOLD4 are 2-butanol (SBT). The centroid ligands are (2S,3S)-butane-2,3-diol (BUD) and (R,R)-2,3-butanediol (BU3). In comparison, the ligand present in the crystal structure is phosphoaminophosphonic acid-adenylate ester (ANP) and is shown above in Figure 4.17. As this CASP12 target has no associated PDB ID nor was an observed structure released by CASP12 organisers, the structure with its ligand is unable to be shown, observed ligand-binding residues were obtained from the supplementary data published by the CASP organisers (Liu *et al.*, 2018).

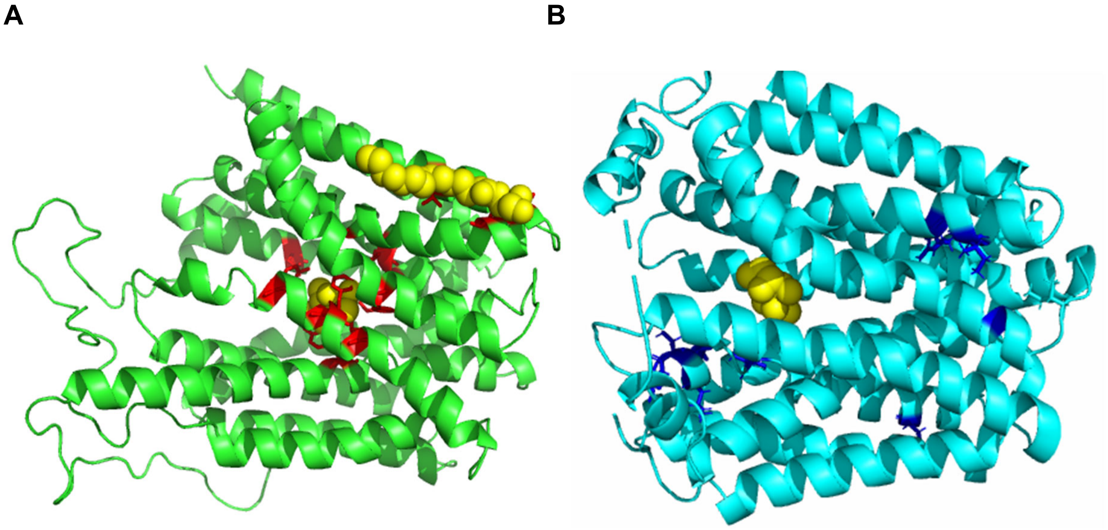


Figure 4.18. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0911 (PDB ID 6e9n)
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted ligand 78M shown as sphere and coloured yellow and located in the middle of the protein. The AFS ligand is shown as sphere and coloured yellow and shown at the top of the protein. A BDT score of 0.05 and MCC score of 0.03 was obtained **(B)** The observed ligand binding site for T0911 with binding site residues shown as sticks and coloured blue the observed ligand GCO is shown as sphere and coloured yellow.

The ninth predicted CASP12 target with a ligand in the crystal structure is T0911 and the MCC and BDT score was 0.03 and 0.05, respectively. As with CASP12 target T0910, there were no correct predictions and IntFOLD4 predicted 13 residues, whereas there are 10 observed ligand binding site residues. The ligands predicted by IntFOLD4 are (2S)-2,3,dihydroxypropyl(7Z)-pentadec-7-enoate (78M) and N-[1R)-1-phosphonoethyl]-L-alaninamide (AFS). In comparison, the ligand present in the crystal structure is gluconic acid (GCO), and the observed ligand binding site residues are predicted in a distal location from the ligand as shown in Figure 4.18.

Figure 4.19 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.82921 was achieved, showing very good structural homology. The main difference between the two models was the disordered flexible loops which is seen in the predicted model (red).

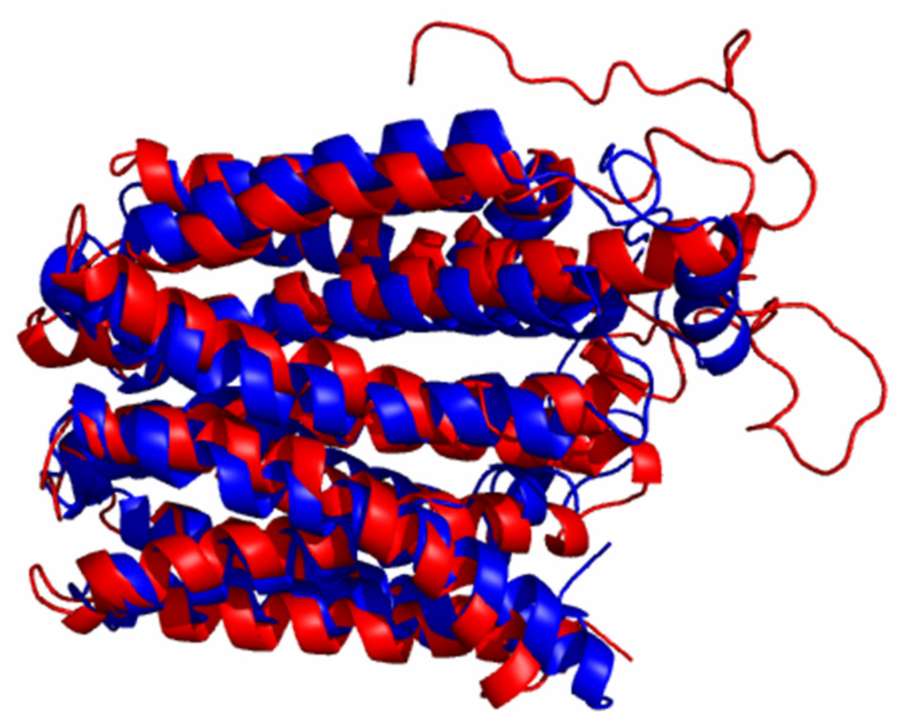


Figure 4.19. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0911 (PDB ID 6e9n)

The structure in blue is the observed structure for T0911 and the predicted structure is in red. A TM-score of 0.82921 was achieved for the protein structures. The score was normalised for the observed structure T0911 as it is the reference molecule

4.3.3 Apo structures: critical residues, known motifs or site-finding residues

CASP12 identified nine targets which came under the apo category, these were T0880-0, T0880-1, T0894 (PDB ID 5hkq), T0895 (PDB ID 5hkq), T0896, T0913, T0917, T0942 (PDB ID 5xi8) and T0947. IntFOLD4 predicted ligand-binding sites for the following targets; T0896, T0913, T0917 and T0942 (PDB ID 5xi8). CASP12 target IDs are given outside parenthesis and PDB IDs, where applicable within parenthesis.

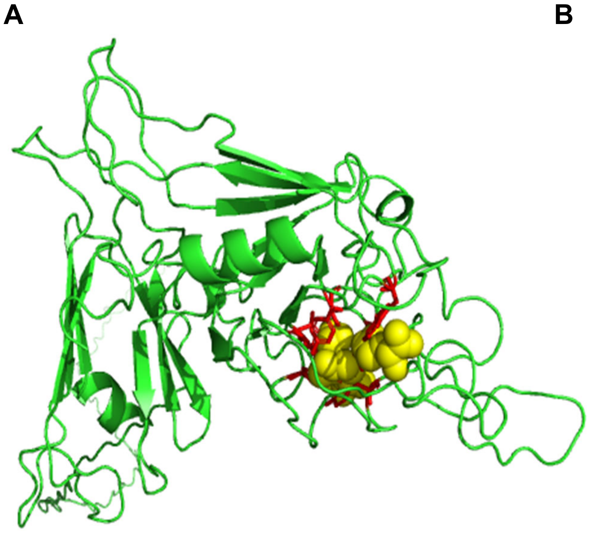


Figure 4.20. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0896
(A) Predicted ligand binding site residues shown as sticks with incorrect -predictions in red, the predicted ligands d-glutamic acid (DGL) and (2S,3R,4S)-4-{[(3S,5R)-5-(dimethylcarbamoyl)pyrrolidin-3-yl]sulfanyl}-2-[(2S,3R)-3-hydroxy-1-oxobutan-2-yl]-3-methyl-3,4-dihydro-2H-pyrrole-5-carboxylic acid (MXR) shown as spheres and coloured yellow. BDT score of 0.00978 and MCC score of 0.0344 **(B)** The observed ligand binding site for T0896 with binding site residues shown as sticks and coloured blue no observed ligands are shown as no PDB ID has been released.

The first predicted IntFOLD4 apo predicted CASP12 target is T0896 and the MCC and BDT score was 0.00978 and 0.0344, respectively. There were no correct predictions and IntFOLD4 predicted 10 residues, whereas there are 23 observed ligand binding site residues as shown in Figure 3.20. One of the predicted ligands; DGL exists as a free ligand in 39 entries on PDB (Liu *et al.*, 2018). Examples of proteins which have DGL has a bound ligand include isomerase and hydrolase (Burley *et al.*, 2017), as there is no PDB ID associated with the target, at the time of writing, classification of the protein and potential identification of ligands is unable to be completed at this stage. In comparison, the other predicted ligand

MXR (meropenem), is only present in one entry on PDB (3vyp), a transferase (Burley *et al.*, 2017). Meropenem is a member of the carbapenem class of β-lactams and contains a bicyclic nucleus, a pyrroline ring and a β-lactam ring. Research has been conducted into the role of meropenem in transpeptidase enzymes to investigate other therapies for multidrug-resistant and extensively drug-resistant strains of *M.tuberculosis* (Burley *et al.*, 2017).

Figure 4.21 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.36086 was achieved, showing poor structural homology. The difference between the two models is a part of the predicted 3D model (red) which was not folded into the structure (partly shown).

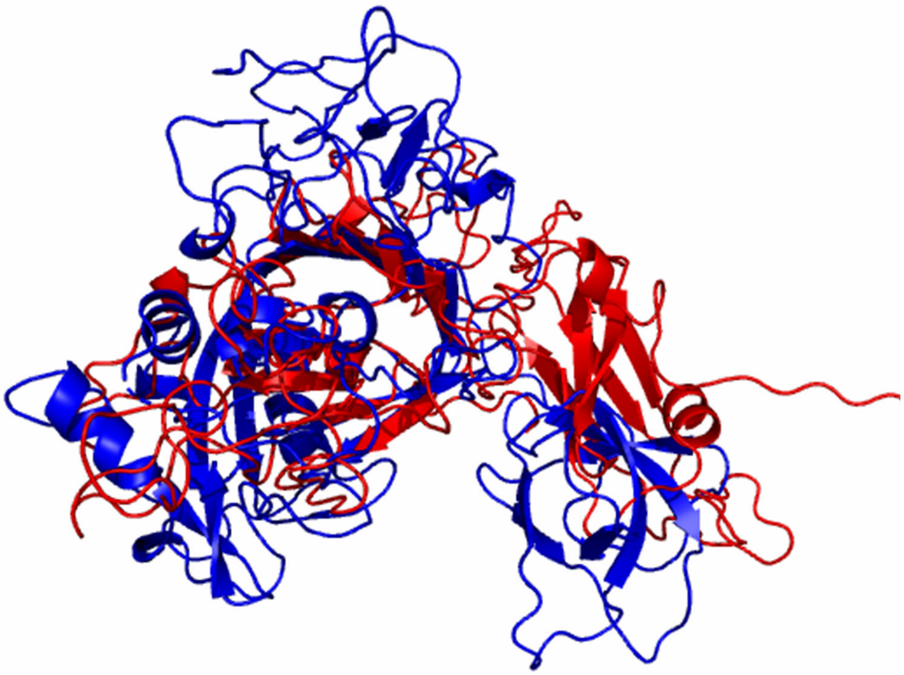


Figure 4.21. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0896
The structure in blue is the observed structure for T0896 and the predicted structure is in red. A TM-score of 0.36086 was achieved for the protein structures. The score was normalised for the observed structure T0896 as it is the reference molecule

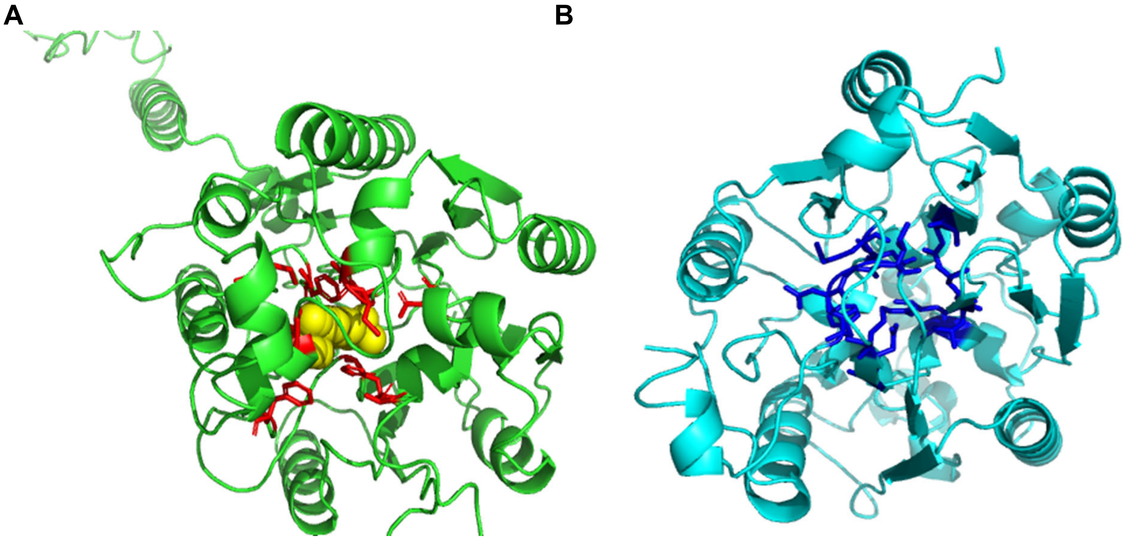


Figure 4.22. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0913
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted ligands arginine (ARG) and citrulline (CIR) shown as spheres and coloured yellow. BDT score of 0.09 and MCC score of 0.03 **(B)** The observed ligand binding site for T0913 with binding site residues shown as sticks and coloured blue no observed ligands are shown due to no PDB ID being released

The second IntFOLD4 apo CASP12 target is T0913 and the MCC and BDT score was 0.03 and 0.09, respectively. There were no correct predictions and IntFOLD4 predicted 12 residues, whereas there are 13 observed ligand-binding sire residues. It can be debated if ARG is a biologically relevant ligand, as it is an essential amino acid (Li *et al.*, 2013). However, the other predicted ligand CIR could potentially be a biologically relevant ligand, as it has been used in nutritional supplementation and for treating dietary shortage or imbalance. As with other CASP12 targets where no PDB ID has been associated, the observed ligand has not been shown in the predicted structure and is shown above in Figure 4.22.

Figure 4.23 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.77492 was achieved, showing good structural homology, despite the poor folding of the predicted 3D model (red).

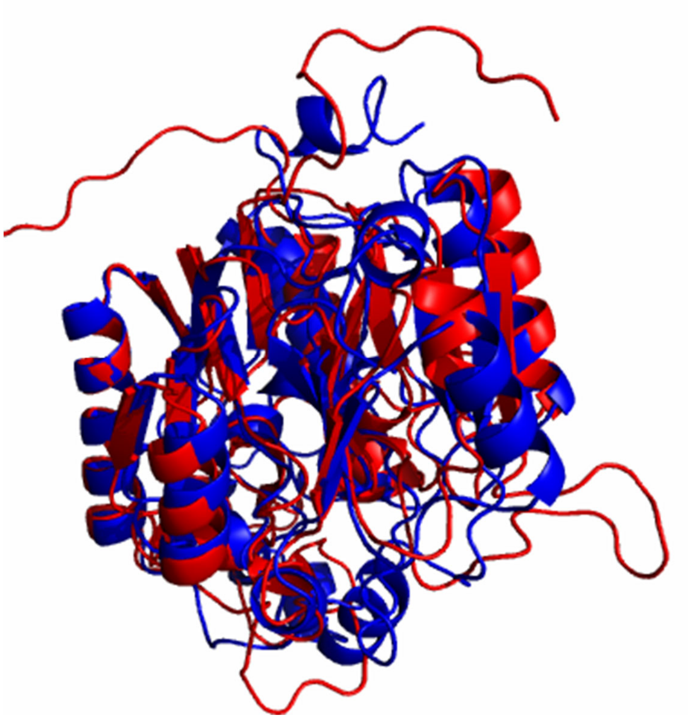


Figure 4.23. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted and observed structure for T0913

The structure in blue is the observed structure for T0913 and the predicted structure is in red. A TM-score of 0.77492 was achieved for the protein structures. The score was normalised for the observed structure T0913 as it is the reference molecule

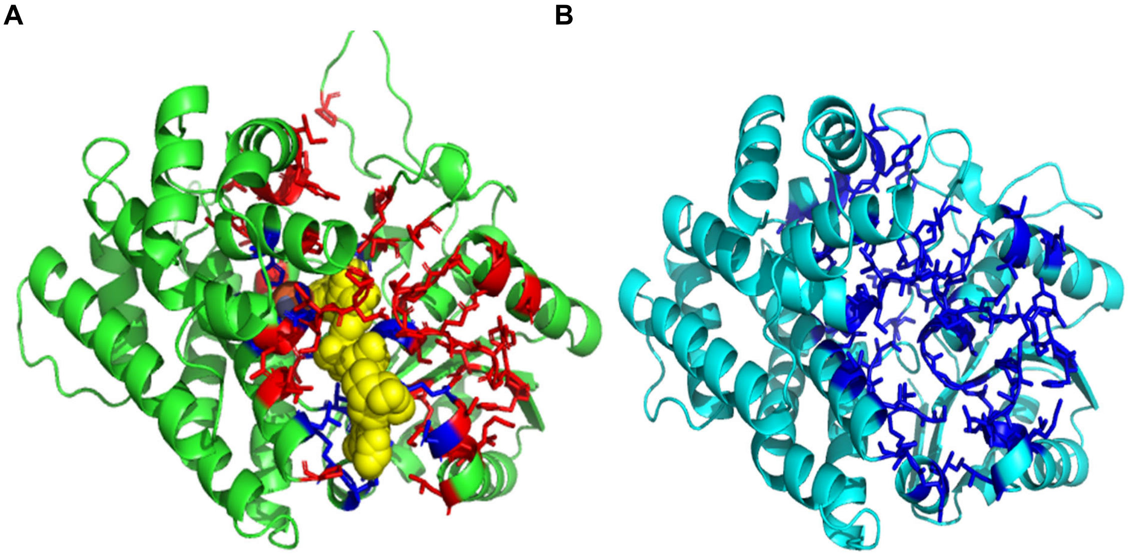


Figure 4.24. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0917
(A) Predicted ligand binding site residues shown as sticks with incorrect -predictions in red and correct predictions in blue the predicted ligands nicotinamide adenine dinucleotide (NAD) and 5,6-dihydroxy-NADP (NZQ) shown as spheres and coloured yellow. BDT score of 0.275 and MCC score of 0.488 **(B)** The observed ligand binding site for T0917 with binding site residues shown as sticks and coloured blue no observed ligands are shown due to no PDB ID being associated with the target

The third predicted IntFOLD4 apo CASP12 target is T0917 and scored the best MCC and BDT score for the apo targets with 0.488 and 0.275, respectively, IntFOLD4 predicted 19 residues whereas there are 73 observed residues. Of the 19 residues which were predicted by IntFOLD4 all were correct predictions. However, IntFOLD4 missed off 54 residue predictions and is shown above in Figure 4.24. The biologically relevant ligands predicted by IntFOLD4 are NAD and NZQ. The ligand NAD, when present in the reduced; NADH is a ubiquitous cellular electron donor and has long been known to control the activity of several of several oxidoreductase enzymes (Burley *et al.*, 2017). The other predicted ligand NZQ exists in two entries on PDB (1oj7 and 5yvm) (Anderson *et al.*, 2017), as there is no PDB ID associated with this target, at this stage without the CASP12 organisers releasing the PDB ID it is impossible to determine if target is one of the entries.

Figure 4.25 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.90785 was achieved, showing very good structural homology with the flexible loops of the predicted protein 3D model (red), not folding exactly as the observed model.

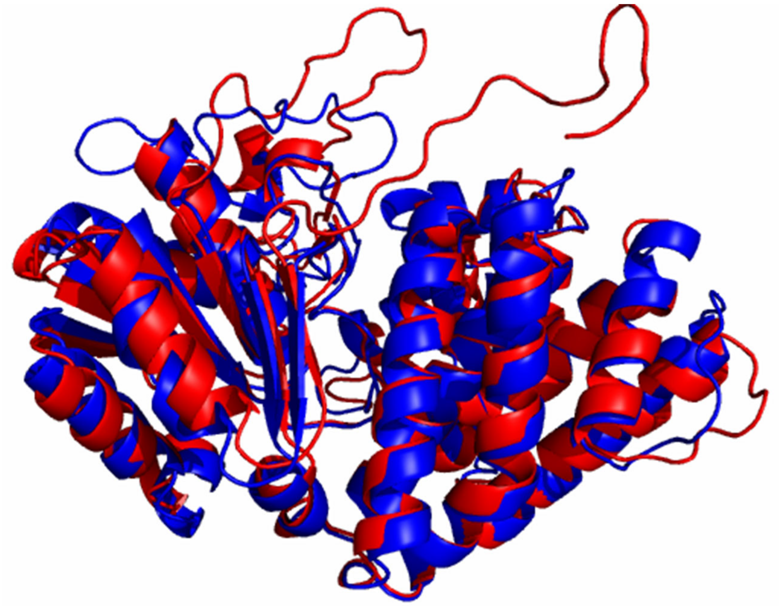


Figure 4.25. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0917

The structure in blue is the observed structure for T0917 and the predicted structure is in red. A TM-score of 0.90785 was achieved for the protein structures. The score was normalised for the observed structure T0917 as it is the reference molecule

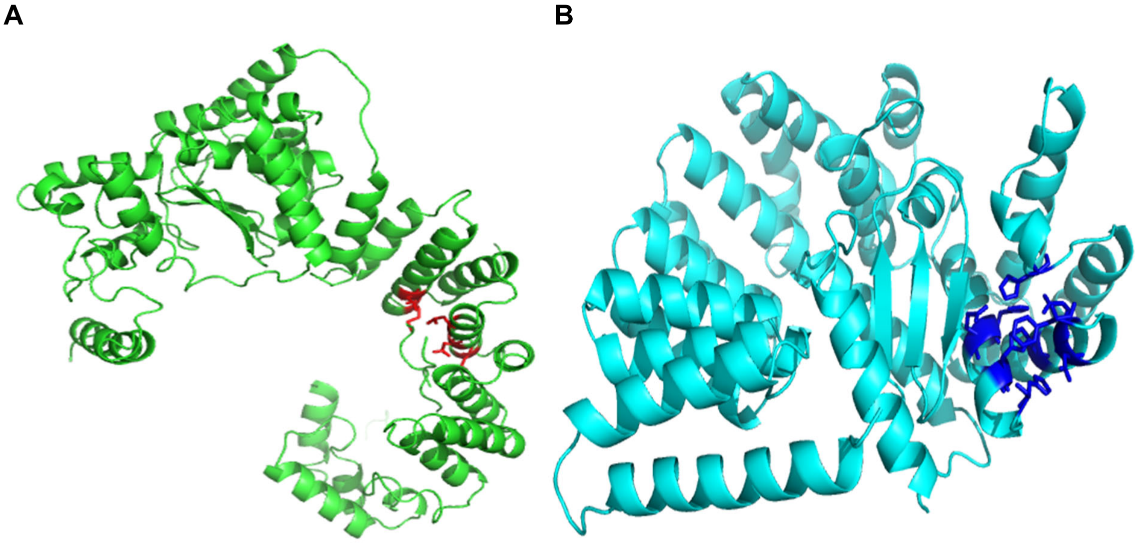


Figure 4.26. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0942 (PDB ID 5xi8)
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted ligands TYR, GLU, GLN and LEU have not been shown. A BDT score of 0.0039 and MCC score of 0.0186 **(B)** The observed ligand binding site for T0942 with binding site residues shown as sticks and coloured blue no observed ligands is shown as it wasn't present in the crystal structure

The final predicted apo CASP12 target is T0942 and scored a MCC and BDT score of 0.0186 and 0.0039, respectively. IntFOLD4 predicted five residues whereas there were 10 observed residues of which none were correct predictions. The ligands predicted by IntFOLD4 were TYR, GLU, GLN and LEU and have not been shown in Figure 4.26 as they might not be biologically relevant. In comparison, the observed ligand was potentially zinc (Liu *et al.*, 2018) and the PDB ID entry for this target identified MSE as a ligand, however previous experience in CASP competitions and available literature information has shown that MSE is not a biologically relevant ligand. Therefore, in this instance IntFOLD4 was unable to predict biologically relevant ligands.

Figure 4.27 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.52751 was achieved suggesting the predicted 3D model has generally the same fold as the observed model. As can be seen from the Figure below, the alpha helices of the predicted model (red) was not aligned with the observed model (blue).

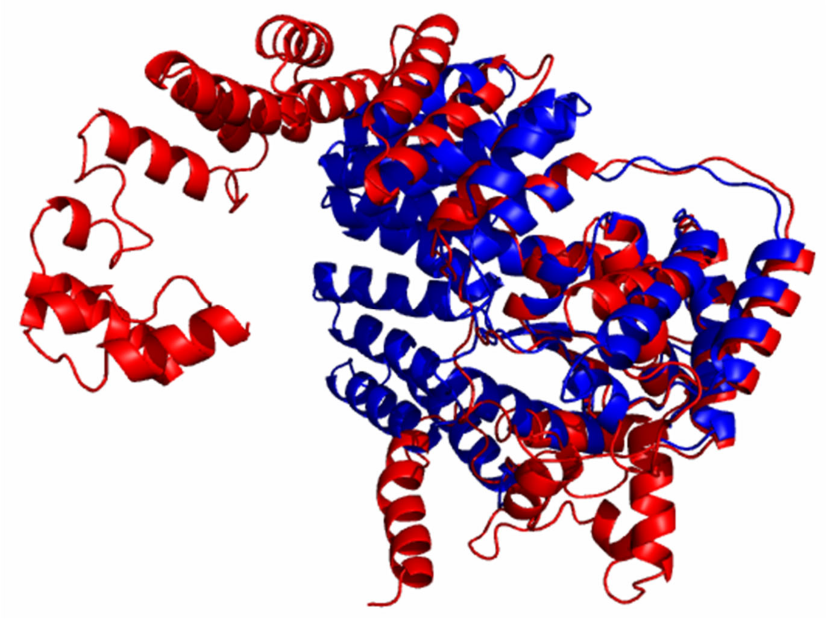


Figure 4.27. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0942 (PDB ID 5xi8)

The structure in blue is the observed structure for T0942 and the predicted structure is in red. A TM-score of 0.52751 was achieved for the protein structures. The score was normalised for the observed structure T0942 as it is the reference molecule

4.3.4 Key, motifs and mutation structures: critical residues, known motifs or site-finding residues

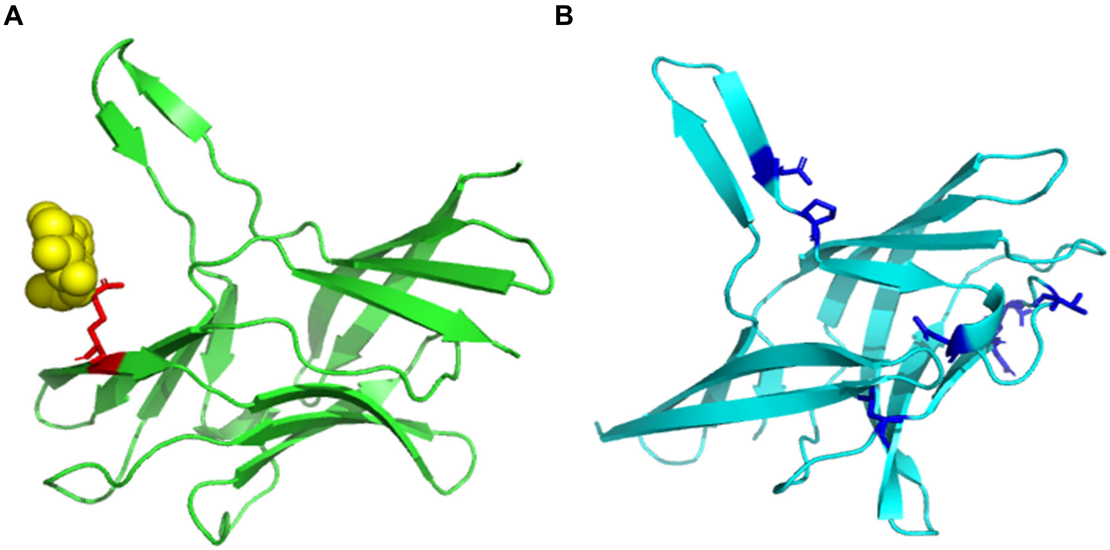


Figure 4.28. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0860 (PDB ID 5fjl)
(A) Predicted ligand binding site residues shown as sticks with incorrect -predictions in red, the predicted ligand beta-d-glucose (BGC) is shown as sphere and coloured yellow. A BDT score of 0.0055 and MCC score of 0.0031 **(B)** The observed ligand binding site for T0860 with binding site residues shown as sticks and coloured blue no observed ligands are shown

The first predicted CASP12 motif target is T0860 and scored a MCC and BDT score of 0.0031 and 0.0055, respectively. IntFOLD4 predicted one residue (Figure 4.28A) whereas, there were 16 observed residues of which none were correct predictions. The ligand predicted by IntFOLD4 was BGC and exists in 821 entries on PDB (Burley *et al.*, 2017) and is not a ligand associated with PDB ID 5fjl. The observed ligand as per the PDB entry and FunFOLD3 was chlorine. The chlorine ligand has not been shown in Figure 4.28B as it does not match the ligand-binding site residues obtained by the CASP12 organisers.

Figure 4.29 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.84792 was achieved showing very good structural homology.

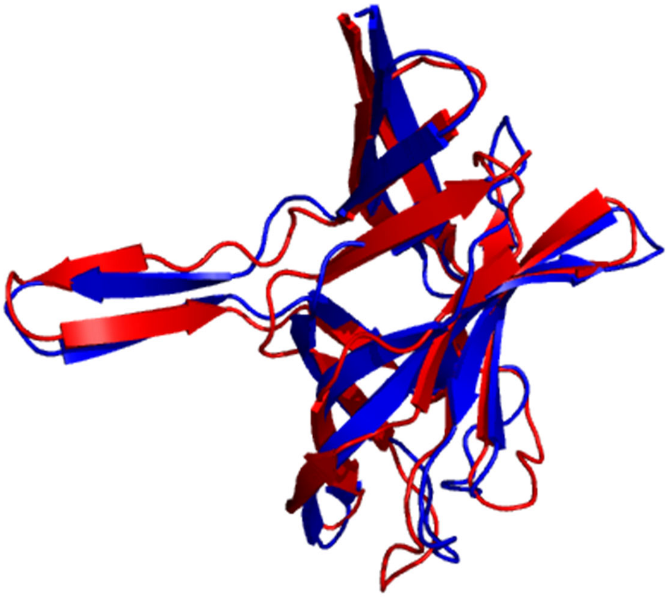


Figure 4.29. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0860 (PDB ID 5fjl)
The structure in blue is the observed structure for T0860 and the predicted structure is in red. A TM-score of 0.84792 was achieved for the protein structures. The score was normalised for the observed structure T0860 as it is the reference molecule

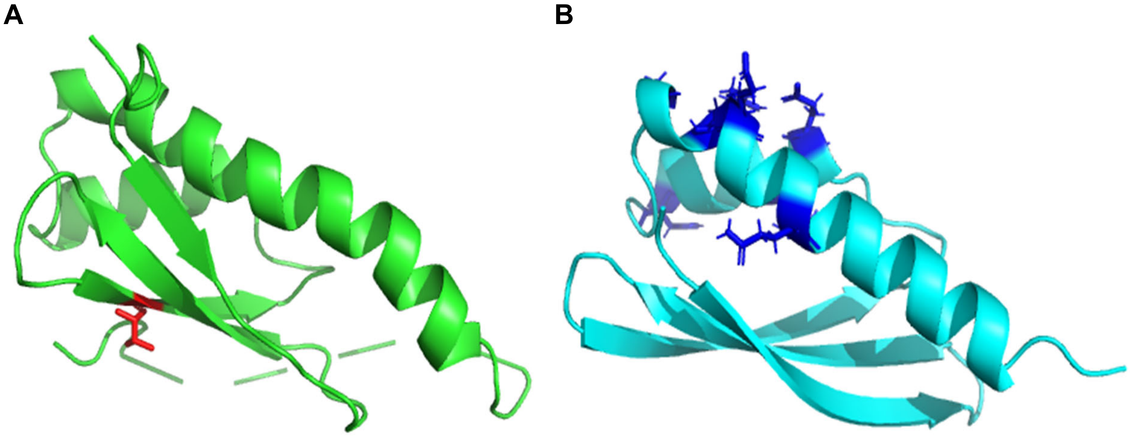


Figure 4.30. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0882 (PDB ID 5g3g) **(A)** Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted ligand VAL has not been shown. A BDT score of 0.0076 and MCC score of 0.0035 **(B)** The observed ligand binding site for T0882 with binding site residues shown as sticks and coloured blue no observed ligands are shown

The first predicted CASP12 target falling into the key residue category and scored a MCC and BDT score of 0.0035 and 0.0076, respectively. IntFOLD4 predicted one residue as shown in Figure 3.30A whereas, there were seven observed residues of which none was a correct prediction as shown in Figure 4.30B. The ligand predicted by IntFOLD4 was VAL. In comparison, the PDB entry for 5g3g identified the following as ligands; sodium ion (Na), 2-methypentane-2,4-diol (MRD), copper ion (Cu) and 2-methyl-2,4-pentanediol (MPD) (Burley *et al.*, 2017). As has been seen in previous PDB entries, not all ligands are biologically relevant. However, a literature search of thermostable multicopper oxidase The catalytic motif in these family of proteins which comprises of laccases, ferroxidases and ascorbate oxidase and ceruloplasmin, includes at least four copper atoms (Burley *et al.*, 2017). Furthermore, data available from the literature has shown that MPD was used as part of a reservoir solution for crystallisation of the protein (Serrano-Posada *et al.*, 2011). No observed structure was released by CASP12 organisers, therefore results are shown for the crystal structure from the PDB.

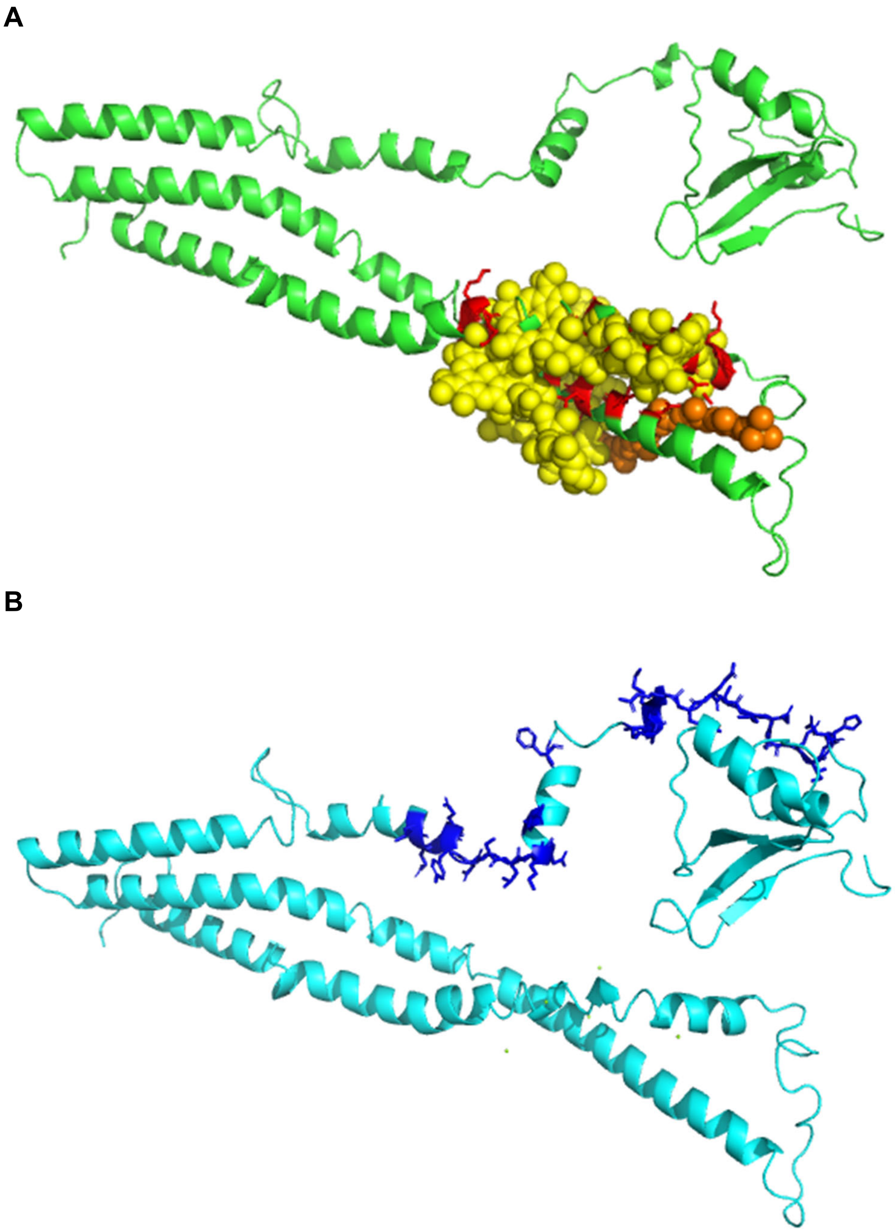


Figure 4.31. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0914 (PDB ID 6cp8)
(A)) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red. The predicted ligand bacteriochlorophyll A (BCL) is shown as sphere and coloured yellow and the other predicted ligand spirilloxanthin (CRT) is shown as sphere and coloured orange. A BDT score of 0.015 and MCC score of 0.08
(B) The observed ligand binding site for T0914 with binding site residues shown as sticks and coloured blue no observed ligands are shown

The second predicted CASP12 target falling into the key residue category is T0914. Currently, the CASP12 organisers have not released the structure. Therefore, the predicted structure from IntFOLD4 has been used to show the observed ligand-binding site residues as shown in Figure 4.31. The MCC and BDT score based on the predicted structure is 0.09 and 0.015, respectively. Furthermore, IntFOLD4 predicted 29 residues and there were 26 residues in the observed structure. As can be seen from Figure 4.31, the predictions appear to be on opposite side of the protein.

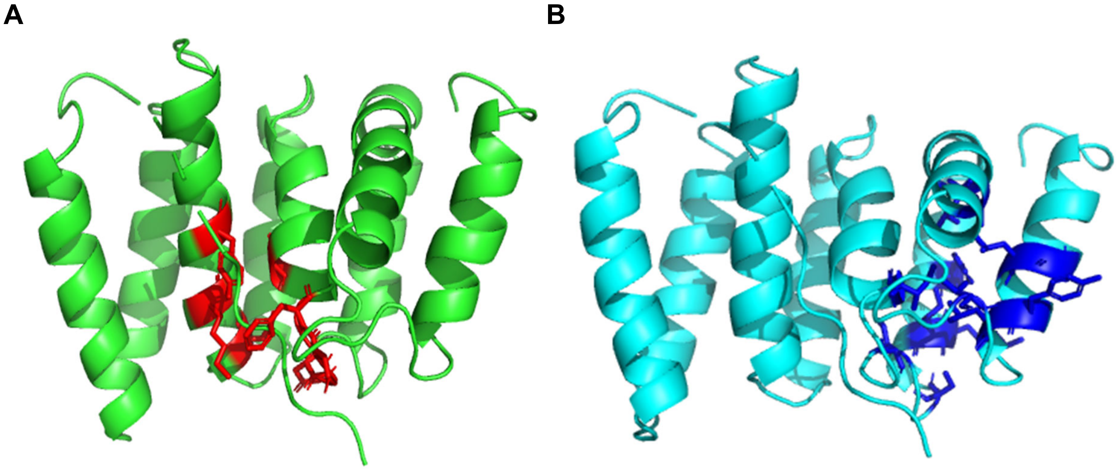


Figure 4.32. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0915 (PDB ID 6cp8)

(A) Predicted ligand binding site residues shown as sticks with incorrect -predictions in red .The predicted ligand LEU has not been illustrated. A BDT score of 0.012 and MCC score of 0.06 **(B)** The observed ligand binding site for T0915 with binding site residues shown as sticks and coloured blue no observed ligands are shown

The third predicted CASP12 target falling into the key residue category is T0915. As with T0914 there is no associated structure. Therefore, the predicted structure from IntFOLD4 has been used to illustrate the observed ligand-binding site residues and is shown above in Figure 4.32. The predicted structure was used to calculate the MCC and BDT score of 0.013 and 0.06, respectively. IntFOLD4 made eight residue predictions and there are 14 observed residues. The ligand predicted by IntFOLD4 was LEU. As with T0914, no structure has been released by CASP12 organisers and the predicted 3D model has been shown for the observed ligand-binding site residues.

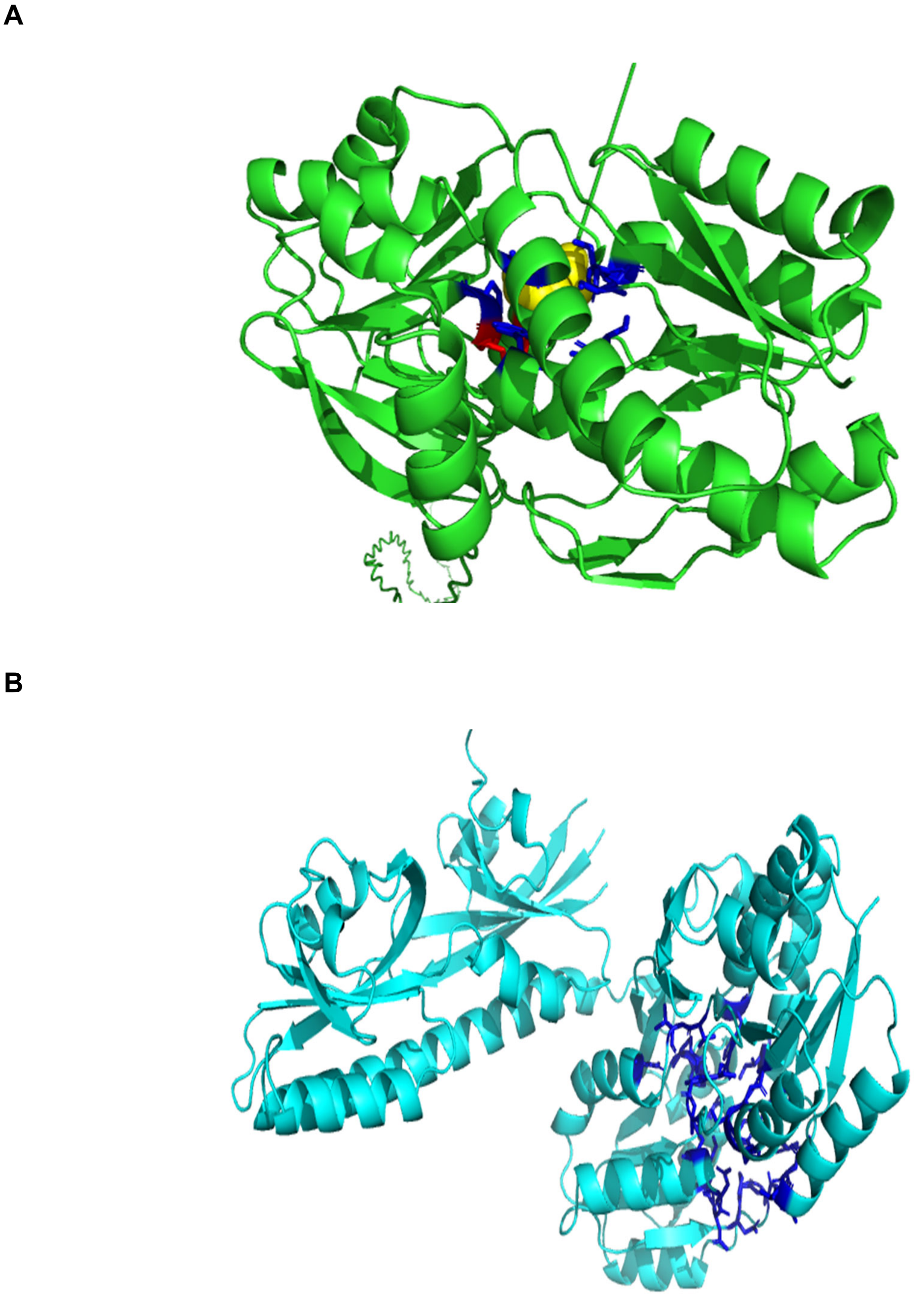


Figure 4.33. Comparison of IntFOLD4 ligand binding site predictions for CASP12 target T0920-0 (PDB ID 5ere)
(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect - predictions in red. The predicted ligand nicotinic acid (NIO) is shown as sphere and coloured yellow. A BDT score of 0.37 and MCC score of 0.57 **(B)** The observed ligand binding site for T0920-0 with binding site residues shown as sticks and coloured blue no observed ligands are shown

The final predicted CASP12 target in the key residue category is T0920-0 a MCC and BDT score of 0.57 and 0.37, respectively. IntFOLD4 predicted 12 residues of which ten of there were correct predictions and is shown above in Figure 4.33. In comparison, there were 31 observed residues. In this category, this prediction scored the highest MCC and BDT score. IntFOLD4 predicted NIO as a biologically relevant ligand. The ligands on the PDB entry are listed as cytosine, glycerol, calcium, ethylene glycol, alpha-ketoisocaproic acid and acetic acid. The PDB entry for this protein has no literature currently published, there is a publication awaiting to be published and is entitled a novel extracellular ligand receptor, suggesting there is more to be determined about the protein.

Figure 4.34 below shows the TM-align superposition of the observed and predicted structures. A TM-score of 0.58477 was achieved, this score appears to be driven by the lack of folding of one of the dimers (not shown completely).

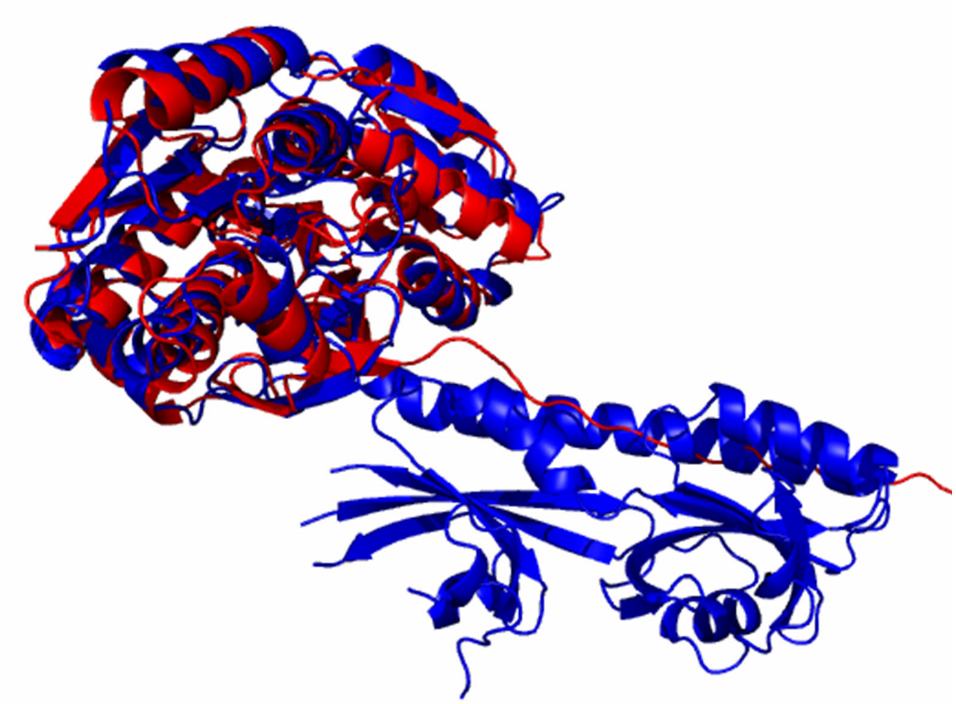


Figure 4.34. Comparison of TMalign (Zhang and Skolnick, 2005) structures for predicted and observed structure for T0920-0 (PDB ID 5erel)
The structure in blue is the observed structure for T0920-0 and the predicted structure is in red. A TM-score of 0.58477 was achieved for the protein structures. The score was normalised for the observed structure T0920-0 as it is the reference molecule

Table 4.9 below is a comparison of the ligands predicted by the FunFOLD3 component of IntFOLD4 in comparison to the ligands associated with the CASP13 targets, as per the PDB entry.

Table 4.9. Comparison of ligands predicted using the FunFOLD3 component of IntFOLD4 and ligands identified in the crystal structure for holo CASP12 targets

CASP 12 target	IntFOLD4 ligand	Observed ligand
T0861 (PDB ID 5j5v)	PLP	LLP
T0863 (PDB ID 5syl)	ASN, THR, LEU	CLR
T0873 (PDB ID 6da6)	MN	FMN
T0879 (PDB ID 5jmu)	ZN	ZN
T0889 (PDB ID 5j09)	NADP	SOR
T0891 (PDB ID 4ymp)	HEM	HEM
T0893	ADP	ADP
T0910	SBT	ANP
T0911	AFS and 78M	GCO

4.4 Retrospective analysis of IntFOLD4 for biological relevance in CASP12

The results from IntFOLD4 were analysed as part of biological relevance which was a topic presented under the umbrella of functional assessment at the CASP 12 meeting (Forli *et al.*, 2016) There were three categories around biological relevance/function assessment and these were holo structures, apo structures and functional patches, which contained one or more key functional residues. The automated FunFOLD3 approach, as part of the IntFOLD4 web server, scored an average MCC and BDT of 0.35 and 0.39, respectively for holo structures and predicted the correct ligand for three of the nine proteins. In this category, IntFOLD4 was ranked second in the strong server performance categories (Altman, 2016). For specific targets, IntFOLD4 was ranked onto of the top servers that predicted the best functional model for T0861 (MCC and BDT 0.497 and 0.543, respectively), the top scoring server for T0873 (MCC and BDT 0.673 and 0.648, respectively), T0889 (MCC and BDT 0.29 and 0.36, respectively) and T0910 (MCC and BDT 0.004 and 0.002, respectively) and the second top scoring server for T0891 (MCC and BDT 0.448 and 0.510, respectively) and T0891 (MCC and BDT 0.448 and 0.510, respectively). IntFOLD4 did not appear in the rankings for T0911 and the rankings for T0863, T0879 and T0893 were not reported by the assessors (Altman, 2016). In comparison, for the apo structure predictions IntFOLD4 did not feature among the top ranked servers. However, IntFOLD4 did rank highly for CASP12 target T0942 (MCC and BDT 0.0186 and 0.0039, respectively) but did not feature in the rankings for the top scoring MCC and BDT target T0917 which achieved a score of 0.488 and 0.275, respectively.

As can be seen from the results, there is a clear difference between the results obtained for holo structures compared to apo structures. FunFOLD3 (forming the ligand binding site prediction aspect of IntFOLD4) was able to make predictions for all nine holo structure targets, whereas predictions could only be made for four of the nine apo structure targets. One of the reasons as to why there could be a difference between predictions is the very nature of the holo and apo structures. Holo structures are in the ligand bound state whereas apo structures are unbound ligand structures. The underlying premise for prediction of functional assessment using ligands by FunFOLD3 is having the ligand already present in the structure, as the predictions are based on proteins which have similar structure will bind similar ligands. Hence, predictions with an unbound ligand will be more difficult as the server is limited in its ability as the apo structure won't have the ligand bound to it. It has been previously reported that there are differences in crystallisation conditions between holo and apo structures and the crystallisation conditions required for apo structures are not necessarily transferable to the protein/ligand (holo) complex (Liu *et al.*, 2018).

A particular problem is the conformational transitions of the receptor associated with ligand binding pose a severe challenge for the structure elucidation of holo complexes (Danley DE, 2008). Hence, why servers which performed well in holo structure prediction did not perform well in apo structure prediction and the converse (Seeliger and de Groot, 2010). In order to obtain successful holo structure prediction from apo structure prediction a degree of flexibility is required within the protein model. One of the top scoring apo structure methods was Rosetta (Altman, 2016) and this differs from the IntFOLD4 server, in that the main focus of Rosetta is structure prediction with no focus on protein-ligand interactions and as apo structures have no ligand bound in its structure this will explain why the server performed highly in this structure category but came eighth in the holo structure category, whereas IntFOLD came second (Ovchinnikov *et al.*, 2018). For the key patches category, IntFOLD4 was not ranked as one of the top scoring servers. However, IntFOLD4 models did score highly for targets T0860 and T0920-1, interestingly there was no functional assessment prediction by FunFOLD3 associated, this may suggest that the best-ranked overall prediction may not have the best functional quality and could sometimes be significantly worse (Altman, 2016).

One of the conclusions from CASP12 for biological relevance, was that models with high structure quality have high functional quality (Altman, 2016) and when looking at the model quality and functional quality as determined by MCC and BDT. For IntFOLD4 in the holo structure category the average global assessment quality score for the top ranked models was 0.68 with the average MCC and BDT score being 0.35 and 0.39, respectively. For the apo structure category the average global assessment quality score was 0.50 and the average MCC and BDT was 0.14 and 0.09, respectively and finally the key motifs category the average global quality assessment score was 0.41 and the average MCC and BDT score was 0.13 and 0.09, respectively. Therefore, the results from IntFOLD4 server certainly support this conclusion and further work would need to be undertaken to determine if model quality is important for functional assessment or if functional assessment is important for model quality and a further consideration will be if the target is holo or apo structure.

FunFOLD3 is residue-centric for the prediction of ligand binding sites, as opposed to pocket-centric. Residue-centric prediction and evaluation will favour spatially precise prediction of one larger binding site over a few smaller ones and this has been demonstrated with targets in CASP11 in particular T0845 (Figure S.10).

As mentioned previously, FunFOLD3 is a template-based modelling (TBM) method for the prediction of ligand-binding site residues. One of the problems with relying on this method, is the over-reliance on the ground truth as defined by known protein-ligand complexes from PDB. It is naïve to assume that in our datasets all possible binding sites are demarked by bound ligands. Locations labelled as negatives might have the particular ligand centered in a different PDB entry which is not something which can be captured, at this stage (Altman, 2016). However, TBM are well known for being the most successful of currently available methods and often produce high confidence predictions supported by examples in template library. On the other hand, since this method and others are based on templates unable to predict novel sites. In comparison, IntFOLD4 has a limited input with only requiring the amino acid sequence. Whereas, the standalone version of FunFOLD3 requires a 3D protein model, a list of templates and the primary sequence. This relatively limited input for IntFOLD4 could be both a positive or a negative and could potentially explain the differing results obtained with standalone FunFOLD3 compared with the FunFOLD3 component of IntFOLD4. The positive of IntFOLD4 requiring just the sequence to predict ligand-binding residues could be that the binding residues of proteins are closely bound with their tertiary structure, so it is possible to predict binding residues from amino acid sequences (Cui *et al.*, 2019) and these residues could be conserved among proteins therefore, it is possible to predict ligand-binding residues from sequence. The FunFOLD3 component of IntFOLD4 predicted ligands for a total of 18 CASP 12 protein targets (refer to Tables 4.6, 4.7 and 4.8) out of a total of 28 protein targets (refer to Table 4.2). Six of the targets were classified as free-modelling (T0863, T0880-1, T0880-0, T0864, T0914, T0915) and the FunFOLD3 component was able to make a prediction for one of the targets (T0863) but not the remaining five. In comparison, standalone FunFOLD3 predicted ligands and ligand-binding predictions for two of these protein targets (T0911 and (T0913). Standalone FunFOLD3 made ligand and ligand-binding site predictions for three additional targets T0912, T0916 and T0919, which were not included in the biologically relevant list from the CASP organisers and the reason for this is currently unknown. On these specific datasets from CASP12, FunFOLD3 component of IntFOLD4 showed a clear strength in comparison to standalone FunFOLD3 and one of the reasons for this could be the use of templates to predict ligands. With TBM methods, ligands will be predicted if a template with a similar fold contains a biologically relevant ligand and the orientation of this ligand in the predicted protein model will depend on the orientation within the template. Additionally, standalone FunFOLD3 has a quality assessment tool so this could lead to different but not necessarily stranger results. Despite the differences between the FunFOLD3 component of IntFOLD4 and standalone FunFOLD3, it is worth mentioning that this is based on only one dataset and not across different CASP protein targets. Furthermore, the highest MCC and BDT score was 0.571 and 0.750, respectively (refer to Table 4.6). In comparison, for FunFOLD3 the highest MCC and BDT was 1.0 albeit not for the same ligand and on a different dataset (refer to Figure 3.27). It is important to note, as with many computational methodologies, extensive testing and validation of these algorithms has been a common topic of literature review (Ghersi & Sanchez, 2011;Fischer, Mayer and Söding, 2008). Therefore, based on the number of datasets available for standalone FunFOLD3 and that early iterations of FunFOLD has been extensively benchmarked in both CASP8 and CASP9 datasets. With FunFOLD among the top 10 methods in CASP9 (Schmidt *et al.*, 2011), it was deemed the best methodology out of the two to go ahead with development and refinement using docking. Further directions of this thesis could be development of IntFOLD4 utilising docking once more data from future CASP experiments is known.

In summary, FunFOLD3 is user friendly it requires a fairly simple input being the amino acid and predictions of ligand-binding site are among the most accurate compared with other available methods, as shown by the results of the CASP competitions .

Chapters 3 and 4 have explored two methodologies for the prediction of function using ligands and ligand-binding site residues. In order to explore a “gold standard” for the prediction of function, benchmarking of GO terms will be explored in the next chapter. Additionally, computational docking is used widely for the study of protein-ligand interactions and/or for drug discovery or development (Punta *et al.*, 2012). Single docking experiments are useful for exploring the function of a target (Forli *et al.*, 2016) and could have great potential for investigating the interaction of T0899 with the magnesium ligand. Servers such as AutoDock already exist for this purpose and AutoLigand is a program for predicting optimal sites of ligand-binding on receptors (Forli *et al.*, 2016). Furthermore, AutoDock might also be useful for a number of targets where there is a general consensus in binding site but the ligand needs to be orientated in order to improve the prediction and will be analysed in Chapter 6.

Chapter 5: CAFA3 Challenge and Prediction of GO terms

5.1 introduction

Whilst determining protein-ligand interactions and protein-binding affinity is significant, the general functionality of a protein is also important. This can be inferred from the ligands which are bound, but more precisely, using GO terms and, specifically for enzymes, this is done using EC numbers.

The EC was launched in 1955, by the International Congress of Biochemistry to create a nomenclature for enzymes. The classification works numerically and each enzymatic function is described by a set of four numbers; referred to as EC numbers. Each of the four numbers represents specific description of the enzyme and its activity (Alborzi, Devignes and Ritchie, 2017). The first digit represents the top-level or branch of the hierarchy and selects one of the six principle enzyme classes which are; oxidoreductase (EC 1), transferase (EC 2), hydrolase (EC 3), lyase (EC 4), isomerase (EC 5) and ligase (EC 6) (Punta & Ofran, 2008). The second digit defines a general enzyme class, specifically the chemical substrate type. The third digit, a more specific enzyme-substrate class (e.g. distinguishing methyl transferase from formyl transferase). The fourth digit, a particular enzyme substrate (Alborzi, Devignes and Ritchie, 2017) An example to demonstrate the classification is, carboxylesterase (3.1.1.1) and isochorismatase (3.3.2.1). Both these enzymes share the same basic activity of hydrolase however, the subsequent numbers show that the enzymes act on different types of bonds; 3.1- act on ester bonds and 3.3- act on an ether bond (Alborzi, Devignes and Ritchie, 2017). As a result of EC numbers being assigned according to the reaction a protein catalysed, it is possible for different proteins to be assigned the same EC number, even if they have no sequence similarity or if they belong to different structural families (Punta & Ofran, 2008).

The GO project is a controlled vocabulary to describe the function of any gene product in any organism. GO terms are organised in a tree-like structure, starting from a more general “root” to specific “leaves”, (Alborzi, Devignes and Ritchie, 2017) a method called semantic similarity, which measures the degree of relatedness between two entities. This method is based on similarity of their annotations and provides a repository of biological annotations of genes and proteins (Rubin & Yarden, 2001). GO is organised as three independent directed acyclic graphs (DAGs) based on three specific aspects of proteins; molecular function (the underlying activity of a gene product at the molecular level, such as binding or catalysis), biological process (operations or sets of molecule events with defined beginning and end, fundamental to functionality of integrated living cells, tissues, organs or organisms) and cellular component (parts of a cell or its extracellular environment) (Dutta, Basu and Kundu, 2017). The nodes represent GO terms and the edge represent different hierarchical relationships. The two most important relations for GO terms are ‘is a’ and ‘part of’ (Dutta, Basu and Kundu, 2017). ‘Is a’ means term A is a subtype of term B, e.g., transcription is a type of nucleic acid metabolic; and ‘part of’ means term A is always part of term B e.g. transcription is always part of gene expression (Dutta, Basu and Kundu, 2017).

GO is loosely hierarchical with ‘child’ terms showing more specificity than their ‘parent’ terms, but this hierarchy is not strict with ‘child’ terms having the potential to possess more than one parent term (*QuickGO*, 2017). Every term has a term name and a unique zero-added seven digit identifier (often referred to as term accession or term accession number), prefixed by GO:, e.g. GO:0006818. The numerical portion of the ID has no inherent meaning or relation to the position of the term in the ontologies (*Gene Ontology Consortium*, 2019). Instead, GO IDs are assigned to individual ontology editors or editing groups and can thus be used to trace who added the term (*Gene Ontology Consortium*, 2019).

Using the GO annotation method, the large transmembrane protein complex, cytochrome c (GO:0004129) would be described by the **molecular function** term oxidoreductase activity, the **biological process** terms oxidative phosphorylation and induction of cell death and the

cellular component terms mitochondrial matrix and mitochondrial inner membrane (*Gene Ontology Consortium*, 2019). Figure 5.1 is the hierarchical mapping for cytochrome c, illustrating the biological process and molecular function.

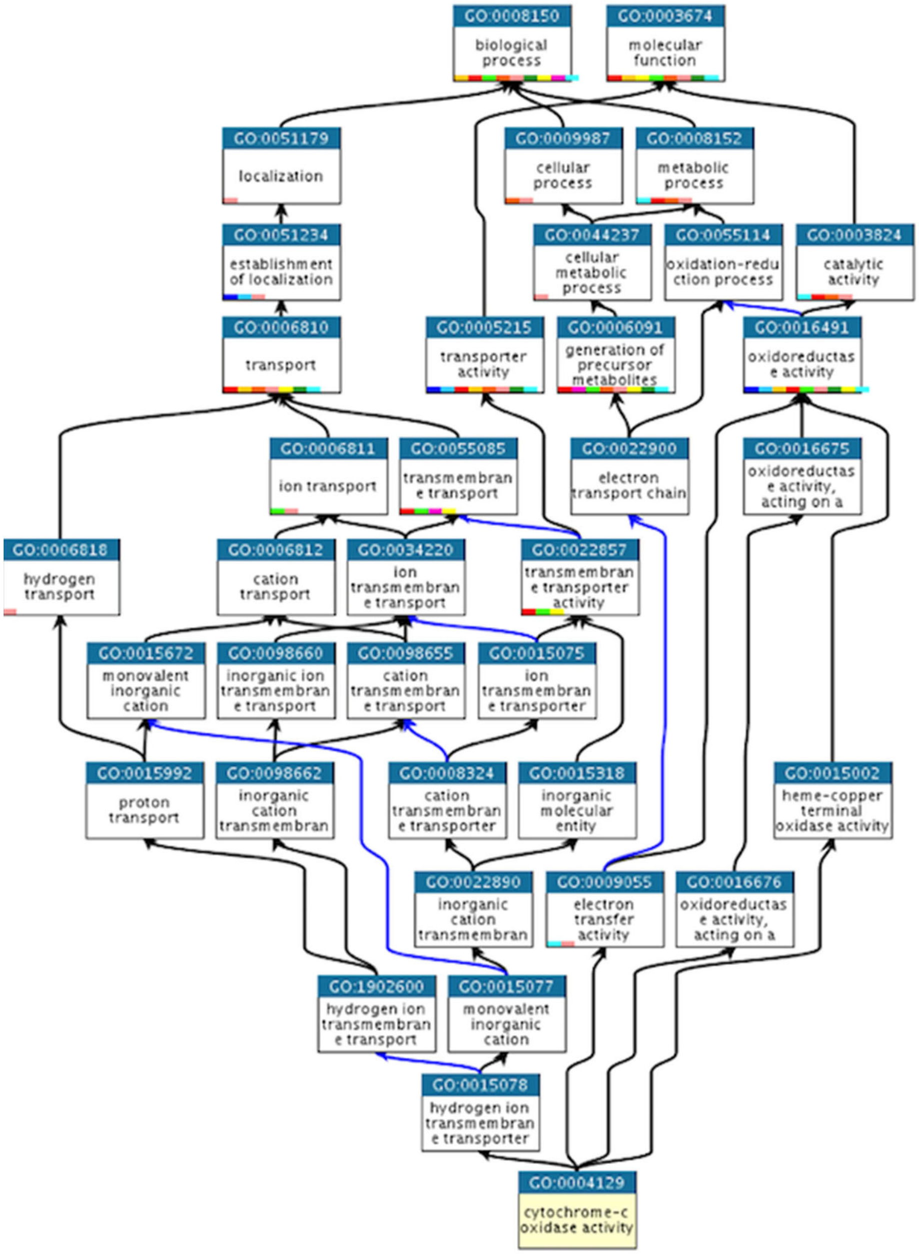

GO:0008150
biological process
GO:0003674
molecular function
GO:0051179
localization
GO:0009987
cellular process
GO:0008152
metabolic process
GO:0051234
establishment of localization
GO:0044237
cellular metabolic process
GO:0055114
oxidation-reduction process
GO:0003824
catalytic activity
GO:0006810
transport
GO:0005215
transporter activity
GO:0006091
generation of precursor metabolites
GO:0016491
oxidoreductase activity
GO:0006811
ion transport
GO:0055085
transmembrane transport
GO:0022900
electron transport chain
GO:0016675
oxidoreductase activity, acting on a
GO:0006818
hydrogen transport
GO:0006812
cation transport
GO:0034220
ion transmembrane transport
GO:0022857
transmembrane transporter activity
GO:0015672
monovalent inorganic cation
GO:0098660
inorganic ion transmembrane transport
GO:0098655
cation transmembrane transport
GO:0015075
ion transmembrane transporter
GO:0015992
proton transport
GO:0098662
inorganic cation transmembran
GO:0008324
cation transmembrane transporter
GO:0015318
inorganic molecular entity
GO:0015002
heme-copper terminal oxidase activity
GO:0022890
inorganic cation transmembran
GO:0009055
electron transfer activity
GO:0016676
oxidoreductase activity, acting on a
GO:1902600
hydrogen ion transmembrane transport
GO:0015077
monovalent inorganic cation
GO:0015078
hydrogen ion transmembrane transporter
GO:0004129
cytochrome-c oxidase activity

Figure 5.1. Hierarchical mapping of Gene Ontology (GO) categories enriched in cytochrome c oxidase activity Biological process is on the left and molecular function is on the right. The black lines illustrate “is a” and the blue lines illustrate “part of”. The top of the chart (biological process and molecular function) are less specific concepts and further down the chart are more specific concepts, finishing with the protein in question; cytochrome c. The colours in the boxes are slim colours and denote which organism the activity was identified in (see Appendix A for further information). Figure taken from *Gene Ontology Consortium*, 2017

The GO term annotations can be used to infer the functional relationship between two proteins. The semantic similarity between two interacting proteins can be estimated by combining the similarity scores of GO terms associated with the proteins (*QuickGO*, 2017). GO has become the standard for assessing the performance of function prediction methods (Dutta, Basu and Kundu, 2017) due to the semantic similarity utilised. Semantic similarities are used in protein-protein interaction predictions, interaction network predictions, biological pathway modelling and clustering of proteins (Punta & Ofran, 2008).

In comparison to EC numbers, GO terms are used more widely in the prediction of proteins. This is supported by the number of methods (Dutta, Basu and Kundu, 2017) to predict GO terms compared with the fewer methods to utilise EC numbers (Gerlt *et al.*, 2015; Gundersen *et al.*, 2015; Koskinen *et al.*, 2015; Piovesan, Giollo, Leonardi, *et al.*, 2015; Sahraeian, Luo and Brenner, 2015; Yu, Zhu and Domeniconi, 2015). However, there are methods for predicting both GO and EC terms, including COACH (Jianyi Yang, Roy and Zhang, 2013) and FunFOLD3 (Roche & McGuffin, 2015).

The Critical Assessment of Functional Annotation (CAFA) is a community challenge that seeks to bridge the gap between the expanding pool of molecular data and the limited resources available to understand protein function (Radivojac *et al.*, 2013). The first two CAFA challenges (CAFA1 and CAFA2) and were carried out in 2010-2011 and 2013-2014, respectively. CAFA1 adopted a time-delay evaluation method, where protein sequences that lacked experimentally verified annotations, or targets were released for prediction (Radivojac *et al.*, 2013). CAFA2 expanded on CAFA1 by the number of ontologies used for predictions, the number of target and benchmark proteins and the introduction of new assessment metrics. CAFA3, continued with all type of evaluation from CAFA1 and CAFA2, with the addition of experimental screens to identify genes associated with specific functions (Radivojac *et al.*, 2013).

Aim: The aim of this chapter was to determine if there is a gold standard when it comes to protein function prediction. Thus, the question to answer is; does the prediction of ligand and ligand-binding residues provide insight into a protein's function or is it the prediction of GO terms? Overall, this chapter aimed to analyse the results of FunFOLDQ in a blinded experiment focusing solely on the prediction of GO terms.

5.2 Materials and Methods

5.2.1 Materials

CAFA organisers provided participants with 24 target files which were split into three categories (i) prokaryotic (ii) eukaryotic and (iii) moonlighting. An example of the target files for each of the three types is given below, with the entire target file for moonlighting proteins provided in the Appendix 3. Targets files are available for downloaded from the CAFA website (http://biofunctionprediction.org/cafa/).

1. Prokaryotic

```
>T833330000001 3MG1_ECOLI
MERCGWVSQDPLYIAYHDNEWGVPETDSKKLFEMICLEGQQAGLSWITVLKKRENYRACF
HQFDPVKVAAMQEEDVERLVQDAGIIRHRGKIQAIIGNARAYLQMEQNGEPFVDFVWSFV
NHQPQVTQATTLSEIPTSTSASDALSKALKKRGFKFVGTTICYSFMQACGLVNDHVVGCC
CYPGNKP
```

2. Eukaryotic

```
>T37020000001 14310_ARATH
MENEREKQVYLAKLSEQTERYDEMVEAMKKVAQLDVELTVEERNLVSVGYKNVIGARRAS
WRILSSIEQKEESKGNDENVKRLKNYRKRVEDELAKVCNDILSVIDKHLIPSSNAVESTV
FFYKMKGDYYRYLAEFSSGAERKEAADQSLEAYKAAVAAAENGLAPTHPVRLGLALNFSV
FYYEILNSPESACQLAKQAFDDAIAELDSLNEESYKDSTLIMQLLRDNLTLWTSDLNEEG
DERTKGADEPQDEN
```

3. Moonlighting

```
>M96060000001 IPPK_HUMAN
MEEGKMDENEWGYHGEGNKSLVVAHAQRCVVLRFLKFPPNRKKTSEEIFQHLQNIVDFGK
NVMKEFLGENYVHYGEVVQLPLEFVKQLCLKIQSERPESRCDKDLDTLSGYAMCLPNLTR
LQTYRFAEHRPILCVEIKPKCGFIPFSSDVTHEMKHKVCRYCMHQHLKVATGKWKQISKY
```

```
CPLDLYSGNKQRMHFALKSLLQEAQNNLKIFKNGELIYGCKDARSPVADWSELAHHLKPF
FFPSNGLASGPHCTRAVIRELVHVITRVLLSGSDKGRAGTLSPGLGPQGPRVCEASPFSR
SLRCQGKNTPERSGLPKGCLLYKTLQVQMLDLLDIEGLYPLYNRVERYLEEFPEERKTLQ
IDGPYDEAFYQKLLDLSTEDDGTVAFALTKVQQYRVAMTAKDCSIMIALSPCLQDASSDQ
RPVVPSSRSRFAFSVSVLDLDLKPYESIPHQYKLDGKIVNYYSKTVRAKDNAVMSTRFKE
SEDCTLVL
```

5.2.2 Methods

The methodology has been described in Chapter 2.

For the Critical Assessment of protein Function Annotation three different algorithms were investigated for gene ontology prediction. FunFOLDQ (Roche, Tetchner and McGuffin, 2011; Roche, Buenavista and McGuffin, 2012, 2013; Roche & McGuffin, 2016), HHblits (Remmert *et al.*, 2012) and a combined approach. During the CAFA prediction season, the combined method was used.

HHsearch and HHblits are two main programs which are part of the HH-suite package. Both programs can be used to help determine function however, HHblits (HMM-HMM-based lightning fast iterative sequence search) is a faster iteration. HHblits has a profile alignment prefilter which reduces the number of full HMM-HMM alignment to a few thousand, making it faster than PSI-BLAST, yet as sensitive as HHsearch (Remmert *et al.*, 2012). Figure 5.2 below outlines the algorithm for HHsearch and HHblits.

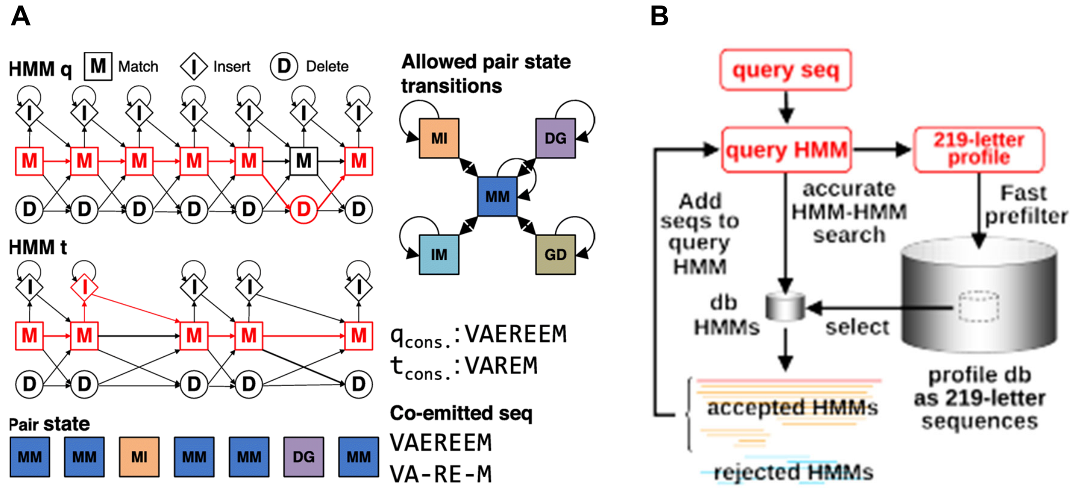


Figure 5.2. Flowchart of the HHsearch and HHblits algorithm
(A) HMM-HMM alignment used by HHsearch. HMM-HMM alignment of query and target. The alignment is represented as red path through both HMMs. M=match, I=insert, D=delete, G=gap. Figure taken from Steinegger et al., 2019. This alignment is more sensitive than profile-profile comparison, profile-sequence comparison and sequence-sequence comparison (Söding, 2005). **(B)** HHblits workflow. HHblits used iterative HMM-HMM alignment to search for homologous sequences in large sequence databases e.g. UniProt. The HHblits databse is a clustered version in which each set of full length alignable sequences is represented by an HMM. represented by an HMM. Sequences from matched HMMs with significant E-value are added to the query MSA, from which a new HMM is calculated for the next search iteration. A prefilter reduces the number of full HMM-HMM alignments approximately 2500-fold Remmert et al., 2012. Figure taken from Remmert et al., 2012.

The combined method, FunFOLDQ and HHblits

The combined method is a combination of the FunFOLDQ algorithm and HHblits. FunFOLDQ is similar to FunFOLD3, but using starting models built from templates identified using HHsearch, which is a rapid homology modelling method. This combination of two orthogonal approaches amalgamates the sequence-based function prediction component from HHblits, along with the structure-based prediction component from FunFOLDQ. FunFOLDQ used a similar approach to FunFOLD3, however it made use of starting models built from templates identified using HHsearch. HHseach aligns a profile HMM against a database of target HMMs. The search first aligns the query HMM with each of the target HMMs using the Viterbi dynamic programming algorithm, which finds the alignment with the maximum score. The E-value for the target is calculated from the Viterbi score (Remmert *et al.*, 2012). Target HMMs that reach sufficient significance to be reported are realigned using the maximum accuracy algorithm (Biegert & Söding, 2008). The algorithm maximises the expected number of correctly aligned pairs of residues and values near 0 produce, long nearly global alignments and values above 0.3 result in shorted, local alignments (Steinegger *et al.*, 2019).

The FunFOLDQ/3 approach follows on from the original FunFOLD method, which was designed upon the basis that; proteins structural templates from the PDB containing biologically relevant ligands and having the same fold as determined by TM-align will likely bind the same ligands. Now, FunFOLD3 makes use of the BioLip database and the prediction of GO terms, as GO term predictions are included in the BioLip template information, in effect the GO term predictions are taken from the closest templates identified (Liu *et al.*, 2018).

HHblits is a HMM-HMM-based lightning fast iterative sequence search extending from HHSearch. HHSearch utilised profile-profile and HMM-HMM search (sequence profiles and profile hidden Markov models) to make it a sensitive class of sequence search methods. HHsearch scores a predicted secondary structure either against a predicted secondary structure or against a known secondary structure (Söding, 2005). HHblits has a profile-profile alignment pre-filter, which reduces the number of full HMM-HMM alignments from many millions to a few thousands, thereby making it faster than PSI-BLAST but still as sensitive as HHSearch (Roche & McGuffin, 2016). HHblits has been part of the HH-suite since 2001 and build high quality multiple sequence starting from a single query sequence or MSA, as mentioned previously it works iteratively, repeatedly constructing new query profiles by adding the results found in the previous round (Remmert *et al.*, 2012).

The combined method pipeline firstly executed HHblits against the UniProt database and the results were used to determine if the sequence was easy or hard to classify. A sequence was determined as easy to classify, if the sequence had 10 or more hits to sequence in UniProt database with an expected value less than 0.001. In comparison, a hard target had less than 10 sequences, with hits having an expected value more than 0.001. The expected (E) value is a parameter that describes the number of hits that can be expected to see by chance when searching database of a particular size. The lower the E-value, or the closer it is to zero, the more significant the match.

The annotation of GO terms was divided into two pipelines for easy and hard targets. The easy target pipeline used HMM-HMM based homology transfer, where the GO terms for the top 15 UniProt hits from HHblits were taken as the predicted GO terms. The probability was determined using the following equation:

Equation 5.1. Easy target pipeline predicted GO terms score

$$\left(1-\frac{1}{no\,of\ predicted\,GO\,terms}\right)*\left(1-\frac{Occurrence\,of\,GO\,within\,UniProt\,DB}{no\,of\,GO\,within\,UniProt\,DB}\right)$$

The hard target (<20% sequence identity) pipeline utilised structural-based homology transfer with FunFOLDQ, i.e., FunFOLD3 was used to predict ligand binding sites based on the models constructed from templates identified using the rapid HHsearch algorithm. The predicted GO terms were scored as below:

Equation 5.2. Hard target pipeline predicted GO terms score

$$\left(1-\frac{1}{no\,of\ predicted\,GO\,terms}\right)*\left(1-\frac{Occurrence\,of\,GO\,within\,BioLip\,DB}{no\,of\,GO\in BioLip\,DB}\right)$$

For each of the targets, GO terms were predicted according to the following ontologies: MFO, BPO and CCO and in the case of human targets; human phenotype ontology (HPO). The evaluation was performed separately for each target. As part of the competition teams could choose to predict function using one or more of the above ontologies and did not have to predict using all of them. A positive prediction was defined as a prediction that was identified by the organisers.

Receiver operator characteristics (ROC) curves were used to objectively measure the accuracy of the predictions compared to the actual function predictions as provided by the

CAFA assessors. The ROC curves provide a graphical representation for each method of the proportion of proteins with correctly identified function prediction against the proportion of proteins incorrectly identified as positive. An ideal curve should have a high sensitivity or a high true positive rate, which is denoted on the y-axis and the false positive rate as low as possible. The x axis is the false positive rate and therefore allows the association between sensitivity and specificity to be explored as the threshold between a positive and negative prediction (Remmert *et al.*, 2012). If a protein had all function annotations correctly predicted and with 100% accuracy, both sensitivity and specificity would be equal to one, and the false positive rate equal to zero. In this instance, the ROC cure would have to pass through the top left hand corner of the plot. The curve would start at the origin; go vertically up the y-axis to a sensitivity of 1.0 and then horizontally across to a false positive rate of 1.0. The closer the curve to the bottom of the curve the worse the method is at predicting protein function accurately (Sedgwick, 2015). Area under the curve can also be calculated from ROC curves and the closer to 1.0 the better. In general an area of 1.0 represents a perfect method or a better predictive power and an area of 0.5 represents a random predicator/method.

5.3 Results

5.3.1 Summary of results

The main findings in this chapter were:

- Prediction of GO terms by FunFOLDQ is not without difficulty. GO terms were incorrect, if the predicted terms did not match the observed terms. The common reason for the difference were the predicted GO terms by FunFOLDQ being a parent term, instead of the actual child terms. Overall, FunFOLDQ did not predict GO terms with enough specificity to be considered a correct prediction.
- The GO terms predicted by FunFOLDQ were often ancestors of the correct GO terms and this could lead to the possibility of a semi-automated process within FunFOLDQ, as opposed to being fully automated. However, fully automated methods are preferred by the McGuffin group.
- Methods which perform well in CAFA are based on sequence (e.g. DeepGoPlus and GOLabeler), whereas FunFOLDQ is based on structure, so could explain the limitations provided in the bullet points above.
- Prediction of GO terms is a complex aspect, so ligand and ligand-binding site residues by FunFOLD3 was determined to be the best method to develop and refine and is explored in Chapter 6.

5.3.2 Overall results

The third CAFA experiment aimed to predict the GO terms of 121,914 protein sequences with domains. Each of the protein sequences had a hierarchical vocabulary split into three categories; biological process, molecular function and cellular component.

FunFOLDQ predicted GO terms across 24 different organisms; a list of the organisms is given below in Table 5.1 and is also separated by eukaryotes and prokaryotes. For ease of reporting, no differentiation is given between different types of organisms, such as bacteria (e.g. gram-positive or gram-negative would still come under bacteria) or fungus would cover fungi and mould. The only difference compared to previous years, was that *E.coli* was separated from the bacteria umbrella as this was done in the most recent CAFA competition, (Serrano-Posada *et al.*, 2011) currently the organisers have provided no rationale for this. Some explanations, could be that there are more sequences for *E.coli* or it might be a model organism so is better understood and potentially easier.

Table 5.1. Table to demonstrate the list of organisms in the eurkaryotes and prokaryotes

Eukaryotes		Prokaryotes
Plants	Fungus	Bacteria
African clawed frog	Yeast	*E.coli*
Dog	Sheep	
Puffer fish	Brown spider	
Mouse	Archaea	
Rat	Roundworm	
Guinea pig	Fruit fly	
Zebra fish	Monkey	
Rabbit	Chimpanzee	
Chicken	Bovine	
Human		

Three methods were used as part of the CAFA3 challenge; the combined method, FunFOLDQ and HHblits. Receiver operator characteristic (ROC) curves were used to assess the sensitivity and specificity of the GO term predictions based on the true positive and false positive rate. The ROC curves for the three methods are given in Appendix 4. Receiver operator characteristic curves were derived using the GO term predictions obtained via each of the individual methods used in the CAFA3 competition. Each predicted GO term assigned as true positive (TP) or false positive was included in the analysis and was ranked from 0.0 to 1.0 and sorted in descending order. The data was then used to produce a ROC curve, using the ROCR plug-in for the R statistical package (Jiang *et al.*, 2016).

The area under the curve (AUC) for FunFOLDQ, HHblits and combined method is 0.47, 0.45 and 0.47, respectively. Predictions were defined as true positive (TP) if the GO term predicted by the server was an exact match for the protein and a false positive (FP) if the server predicted a GO term, which was not associated with the protein. Precision can be interpreted as the error rate and is then defined as the number of TPs over the number of TPs+FPs, recall is synonymous with sensitivity and can also be referred to as true positive rate and is defined as the number of TPs over number of TPs+FNs (Sing *et al.*, 2005). The equation for precision and recall is given below in Equation 5.3. The precision and recall results for each method are shown in Table 5.2.

Equation 5.3. Precision and Recall equation
The equation below illustrates how precision and recall are determined based on TP, FP and FN

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Table 5.2. Precision and recall for each of the three methods used in CAFA3

Server	Precision	Recall
FunFOLDQ	0.29	0.29
Combined	0.29	0.29
HHSearch	0.46	0.74

In order to assess what may have worked well and what specifically requires further improvement with the prediction of GO terms by FunFOLDQ, the protein sequences are looked at in further detail. Table 5.3 is an example of a protein sequence annotation that went right (ACE_RAT) and Table 5.4 is an example of what went wrong (ACH1_CANAL) with the GO term prediction. Further examples of predictions by combined method is provided in Appendix 3.

Correctly predicted GO term annotation example

Table 5.3. Correctly predicted GO terms annotation example ACE_RAT
The table below is an example of GO prediction for CAFA target ACE_RAT, GO term predictions for molecular function and biological process are highlighted in black and bold to denote where exact GO term predictions were made. The GO term prediction for cellular component is bold to denote it is not an exact match but is an ancestor of the CAFA3 target. As per the UniProtKB entry this target had the highest annotation score (5)

Annotation Score: Experimental evidence at protein level	
Existing information in UniProtKB	
Molecular Function	GO:0008237
Molecular Function	GO:0008241
Biological Process	G0:0006508
Cellular Component	**GO:0005886**
FunFOLDQ predicted GO terms	
Molecular Function	G0:0008237 (Exact match) Conf. =0.75
Molecular Function	G0:0008241 (Exact match) Conf. =0.75
Biological Process	G0:0006508 (Exact match) Conf. =0.75
Cellular Component	GO:0016020 (Ancestor) Conf. =0

The hierarchy chart on the next page in Figure 5.3, illustrates for the relationship between the cellular component GO terms 0005886 and 0016020, as can be seen from the Figure there is a relationship between both GO terms, where 0005886 could be deemed a child of 0016020, as 0016020 denotes a location but 005886 provides the specific location. Despite the clear relationship between the GO terms, the prediction for 0016020 was deemed a false positive, as it was not an exact match. For comparison, in Table 5.4 the HHblits method and combined method predicted the following GO terms for ACE_RAT.

Table 5.4. GO terms predicted by HHblits and Combined method for ACE_RAT
Predicted terms with their associated GO terms are given below. GO terms in red and bold had no match in UniProtKB, blue and bold are exact matches in UniProtKB, black and bold are exact matches as per CAFA3 and green and bold terms are ancestor terms

HHblits predictions			
Molecular Function	**GO:0003677**	No annotation in UniProtKB	Conf.=0.93
Molecular Function	**GO:0003899**	No annotation in UniProtKB	Conf.=0.93
Molecular Function	**GO:0004180**	Exact match in UniProtKB	Conf.=0.93
Molecular Function	**GO:0008237**	Exact match in CAFA3	Conf.=0.93
Molecular Function	**GO:0008241**	Exact match in CAFA3	Conf.=0.93
Molecular Function	**GO:0008270**	No annotation in UniProtKB	Conf.=0.93
Molecular Function	**GO:0016740**	No annotation in UniProtKB	Conf.=0.93
Molecular Function	**GO:0016779**	No annotation in UniProtKB	Conf.=0.93
Molecular Function	**GO:0016787**	Exact match in UniProtKB	Conf.=0.93
Molecular Function	**GO:0032549**	No annotation in UniProtKB	Conf.=0.93
Biological Process	**GO:0006351**	No annotation in UniProtKB	Conf.=0.93
Biological Process	**GO:0006508**	Exact match in CAFA3	Conf.=0.93
Cellular Component	**GO:0005634**	Exact match in CAFA3	Conf.=0.93
Cellular Component	**GO:0016020**	Ancestor	Conf.=0.93
Combined predictions			
Molecular Function	**GO: 0008237**	Exact match in CAFA3	Conf.=0.75
Molecular Function	**GO: 0008241**	Exact match in CAFA3	Conf.=0.72
Biological Process	**GO: 0006508**	Exact match in CAFA3	Conf.=0.75
Cellular Component	**GO: 0016020**	Ancestor	Conf.=0.75

As can be seen from the predictions, HHblits predicted more annotations; 14 compared to four predictions by FunFOLDQ and combined methods and half of these predictions were not part of the 140 annotations associated with the protein sequence on UniProtKB. The combined method predicted the exact same annotations as FunFOLDQ.

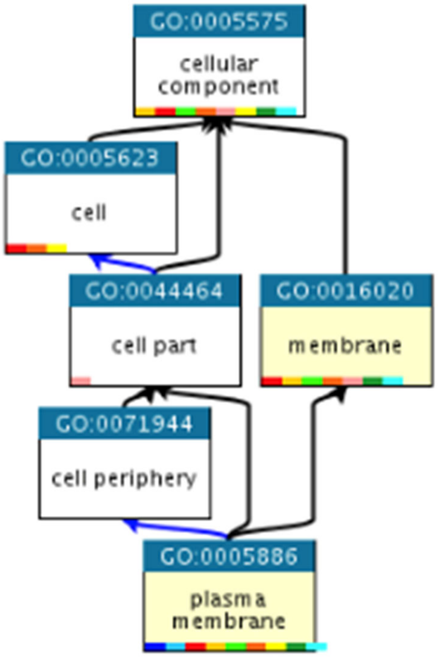


Figure 5.3. A GO hierarchical chart demonstrating the relationship between GO:0016020 and GO:0005886 as predicted by the three methods

The chart clearly demonstrates that GO:0005886 is associated with a membrane (GO:0016020) but more importantly, is a plasma membrane as shown by the specific GO term related to this (GO:0005886). Figure created using QuickGO(Fischer, Mayer and Söding, 2008). An explanation of the GO slim colours is given in Appendix 3.

Incorrectly predicted GO term annotation example

Table 5.5. Incorrectly predicted GO terms annotation example ACH1_CANAL
The table below is an example of GO prediction for CAFA target ACH1_CANAL , there were no correctly predicted GO terms for molecular function, biological process or cellular component. GO terms which were part of the UniProtKB entry but not annotated by CAFA are in blue and bold. GO term predictions which were not annotated in UniProtKB are in red and bold. GO term predictions which are an ancestor term are in green and bold. As per the UniProtKB entry this target had annotation score of 4 out of 5 which is still the highest annotation score with experimental evidence at protein level for the protein

Annotation Score: Experimental evidence at protein level	
Existing information in UniProtKB	
Biological Process	G0:0071469
Cellular Component	GO:0005739
FunFOLDQ predicted GO terms	
Molecular Function	**G0:0003824** Exact match in UniProtKB Conf. =0.88
Molecular Function	**GO:0008814** No annotation in UniProtKB Conf.=0.80
Molecular Function	**GO: 0008815** No annotation in UniPotoKB Conf.=0.65
Molecular Function	**GO: 0016740** No annotation in UniProtKB Conf.=0.88
Molecular Function	**GO: 0016829** No annotation in UniProtKB Conf.=0.88
Biological Process	**GO: 0006084** Exact match in UniProtKB Conf.=0.61
Cellular Component	**GO: 0005737** Ancestor Conf-=0.88
Cellular Component	**GO: 0009346** No annotation in UniProtKB Conf.=0.64

Table 5.5 above, shows the incorrect predicted GO terms for CAFA3 target ACH1_CANAL and the hierarchy chart below in Figure 5.4, illustrates for the relationship between the cellular component GO terms 0005739 and 0005737 as can be seen from the Figure there is a relationship between both GO terms, where 0005739 could be deemed a child of 0005737, as 0005737 denotes a location but 0005739 provides the specific organelle. As with the previous example, despite the clear relationship between the GO terms, the prediction for GO:005737 was deemed a false positive, as it was not an exact match. In comparison HHsearch and Combined predicted the following GO terms:

Table 5.6. GO terms predicted by HHblits and Combined method for ACH1_CANAL
Predicted terms with their associated GO terms are given below. GO terms in red and bold had no match in UniProtKB, blue and bold are exact matches in UniProtKB, and green and bold terms are ancestor terms

HHblits predictions			
Molecular Function	**GO:0003676**	No annotation in UniProtKB	Conf.=0.83
Molecular Function	**GO:0003824**	Exact match in UniProtKB	Conf.=0.83
Molecular Function	**GO:0003986**	Exact match in UniProtKB	Conf.=0.82
Molecular Function	**GO:0016740**	No annotation in UniProtKB	Conf.=0.83
Molecular Function	**GO:0016787**	Exact match in UniProtKB	Conf.=0.83
Biological Process	**GO:0006084**	Exact match in UniProtKB	Conf.=0.58
Combined predictions			
Molecular Function	**GO:0003824**	Exact match in CAFA3	Conf.=0.88
Molecular Function	**GO:0008814**	No annotation in UniProtKB	Conf.=0.80
Molecular Function	**GO:0008815**	No annotation in UniProtKB	Conf.=0.65
Molecular Function	**GO:0016740**	No annotation in UniProtKB	Conf.=0.88
Molecular Function	**GO:0016829**	No annotation in UniProtKB	Conf.=0.88
Biological Process	**GO:0006084**	Exact match in UniProtKB	Conf.=0.61
Cellular Component	**GO: 0005737**	Ancestor	Conf.=0.88
Cellular Component	**GO:0009346**	No annotation in UniProtKB	Conf.=0.64

In this example, the Combined method predicted more annotations than HHblits as shown in Table 5.6 above. However, as with the previous example, despite HHblits predicting terms, which were correctly associated with the protein sequence as per UniProtKB, these were not part of the CAFA3 prediction annotation and were thus deemed negative. The Combined method predicted the exact same terms as the FunFOLDQ method and therefore the predictions made by FunFOLDQ have not been shown. The relationship between the predicted GO term 0005737 and the CAFA3 annotation GO:0005739 is shown below in Figure 5.3.

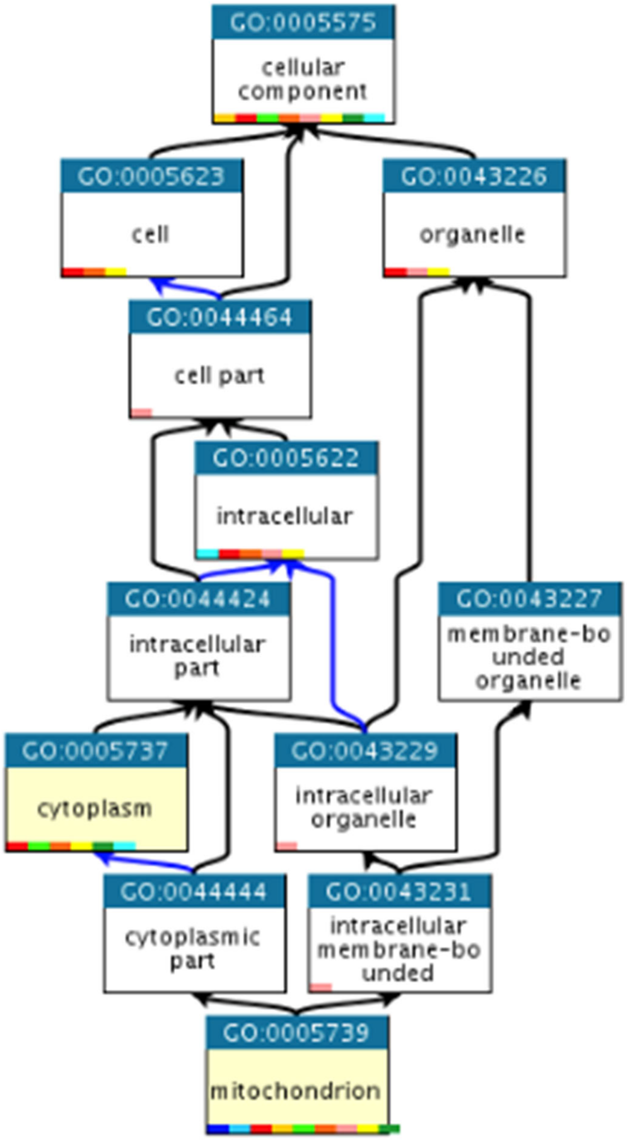


Figure 5.4. A GO hierarchical chart demonstrating the relationship between GO:0005737 and GO:0005739
The chart clearly demonstrates that GO:0005737 is cytoplasm but more importantly, is a mitochondrion in the cytoplasm as shown by the specific GO term related to this (GO:0005739). Figure created using QuickGO (*QuickGO*, 2017)

The following GO predictions from FunFOLDQ were not included in the UniProtKB; GO:0008814, GO:0008815, GO:0016740, GO:0016829 and GO:0009346 and would therefore count as a negative prediction. Two FunFOLDQ predictions were part of UniProtKB but were not included in the CAFA3 annotation; GO:0003824 and GO: 0006084. Hierarchical charts for the GO terms not included in the CAFA3 prediction are given below in Figure 5.5**.**

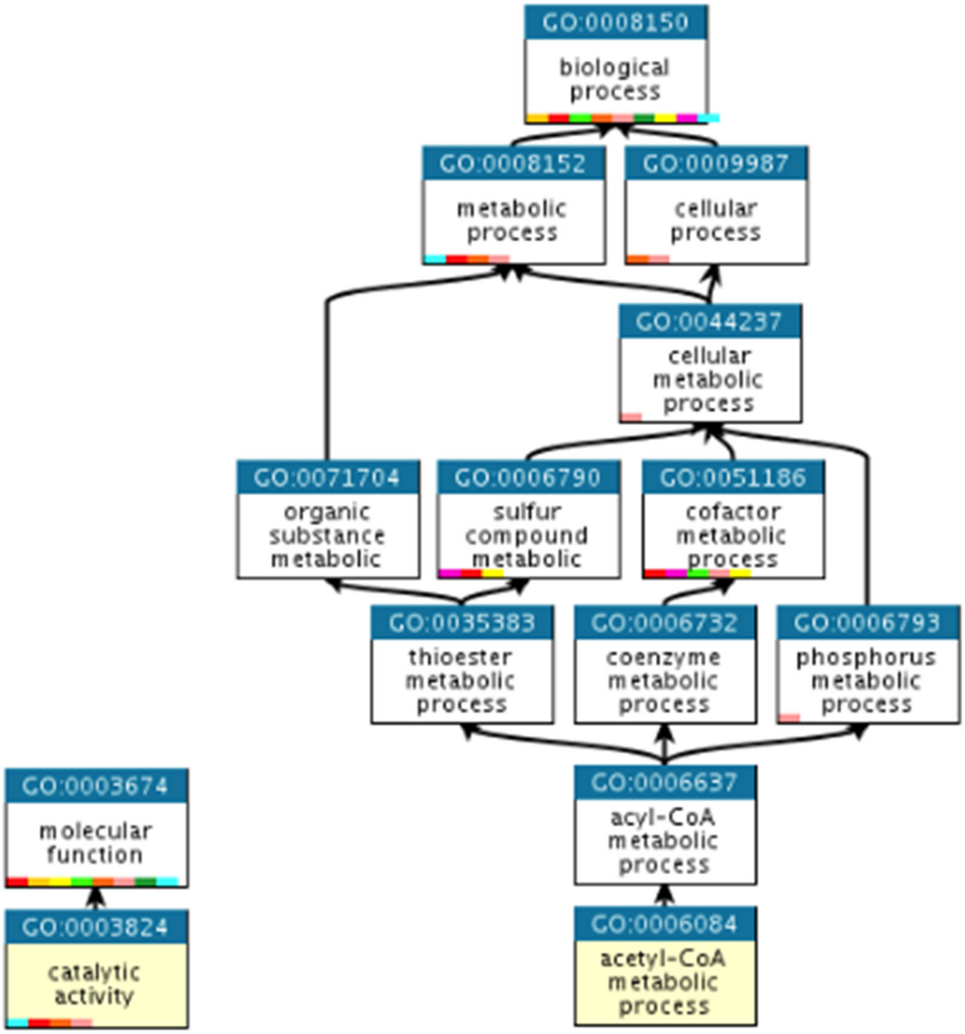

GO:0008150
biological process
GO:0008152
metabolic process
GO:0009987
cellular process
GO:0044237
cellular metabolic process
GO:0071704
organic substance metabolic
GO:0006790
sulfur compound metabolic
GO:0051186
cofactor metabolic process
GO:0035383
thioester metabolic process
GO:0006732
coenzyme metabolic process
GO:0006793
phosphorus metabolic process
GO:0003674
molecular function
GO:0006637
acyl-CoA metabolic process
GO:0003824
catalytic activity
GO:0006084
acetyl-CoA metabolic process

Figure 5.5. A GO hierarchical chart for GO terms predicted by FUNFOLDQ and annotated in UniProtKB
Figure created using QuickGO (*QuickGO*, 2017)

Two protein sequences from CAFA3 have been presented to provide an example of how annotation worked for CAFA3. The first example, ACE_RAT has 140 associated annotations (*QuickGO*, 2017) and CAFA3 benchmarked 22 annotations of which FunFOLDQ correctly predicted three of these. The “incorrect” predictions are not incorrect, in as much that it is completely unrelated to the function of the protein sequence, as illustrated in Figure 5.4 and is annotated on UniProtKB. As is demonstrated in this example, more than one function is associated with a protein and is termed a multiple sub-directed acyclic graph (DAG). It is worth considering whether these annotations should be considered false, when in fact they are true, but not specific enough to determine firm conclusions around function.

The next example, ACH1_CANAL has 15 associated annotations (*QuickGO*, 2017) and CAFA3 benchmarked two annotations and FunFOLDQ did not correctly predict either of these. As with the previous example, FunFOLDQ predicted GO terms, which were annotated by UniProtKB but not benchmarked by CAFA3. The reason why these GO terms were not picked as annotations by CAFA3, could be due to the specificity, as already demonstrated in the previous example. GO:0003824 relates to catalytic activity but the specific activity performed by this protein is acetyl-CoA hydrolase activity, which would have GO:0003986 as the associated molecular function. However, it is worth noting that CAFA3 did not provide a benchmark for this target pertaining to molecular function. Once again, there was a DAG annotation from GO:0005737.

As previously mentioned, there were five incorrect predictions for this protein sequence. GO:000815 is molecular function but specifically citrate (pro-3S)-lyase activity. The parent of this GO term is catalytic activity and is associated with GO:0003824, which is a correct annotation (Figure 5.6A). In this instance it appears FunFOLDQ has picked an incorrect homology whilst maintaining the parent function of the protein. Indeed, a similar result has occurred for GO:0008814, which has GO:0003824 as a parent but is specifically a citrate CoA-transferase activity (Figure 5.6B). This pattern is repeated for GO:0016740 (Figure 5.6C**)**, GO:0016829 (Figure 5.6D)

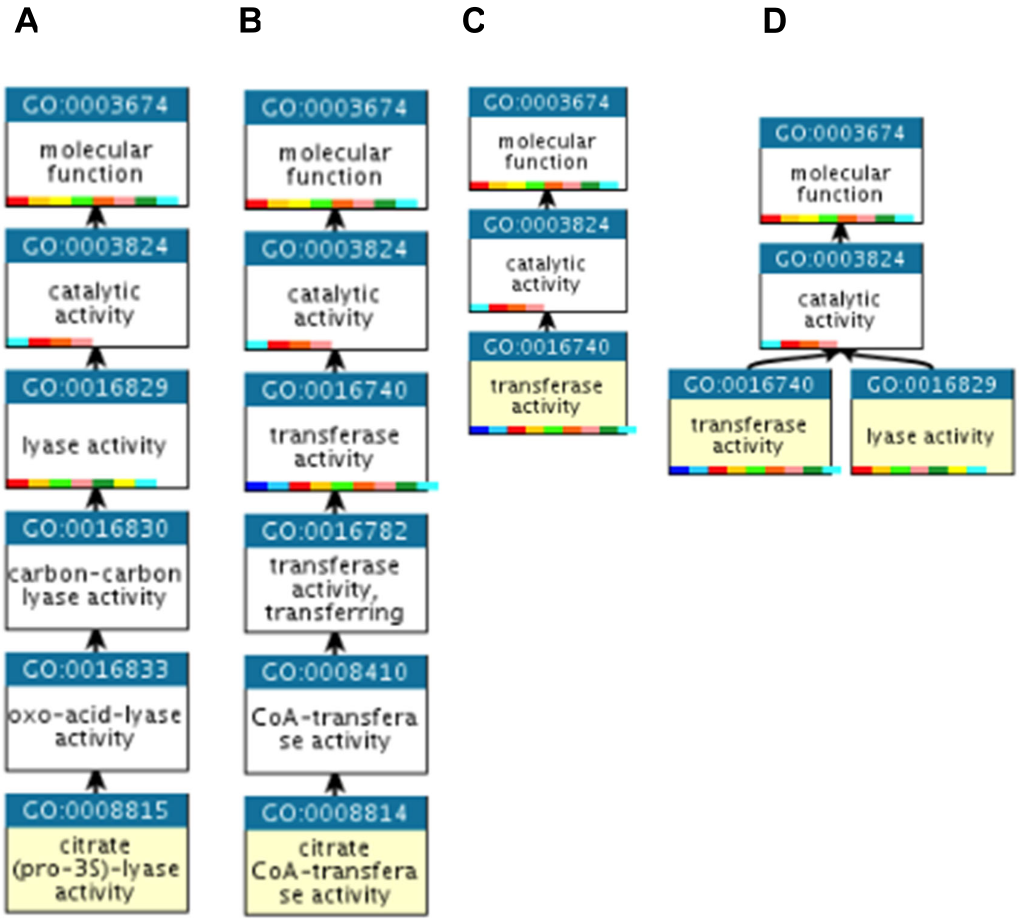


Figure 5.6. GO hieracrchial charts for false positive terms predicted by FunFOLDQ
The GO tems predicted by FunFOLD are presented in a yellow box, despite being false positive predictions all share a correctly annotated GO term as per UniProtKB. Figure created using QuickGO (*QuickGO*, 2017)

5.4 Discussion

FunFOLDQ was entered into the CAFA3 competition to determine performance in the prediction of GO terms, which could potentially provide insights into a protein's function. Based on the ROC curves obtained for FunFOLDQ and the AUC show no better than random prediction and, on this basis, one may presume that the server performed poorly in the CAFA3 challenge.

However, when the examples of annotations are looked at in detail, the GO terms selected by FunFOLDQ are not necessarily incorrect, as these annotations are included in the UniProtKB. As can be appreciated, the level of understanding of a protein can evolve from a basic understanding to very detailed with further evidence from experimental data. This does not make previous basic understanding of a protein necessarily incorrect, just incomplete. However, in the CAFA competition and predicted GO terms, which were not specific, right down to the exact molecular function, biological process or cellular component, were not included as part of the GO term benchmarking. Therefore, if we want to participate in future CAFA challenges, there needs to be more understanding of how CAFA includes annotations and thus further development of FunFOLDQ is needed to determine the relevance of these annotations. This could be explored by better determination between ancestor and child annotations, with more emphasis on child annotations, are these are more likely to be specific and could they match the benchmarking terms identified by CAFA. Furthermore, our results show that a GO term prediction is very black and white, so results are either right or wrong and there is no other way to objectively measure the results. This is very different to participation in CASP competitions where the utilisation of MCC and BDT scores enables credit to be given for predictions, which are close but not perfect. On this basis, GO term prediction as objectively measured in CAFA competitions is not necessarily the best method for elucidation of a protein's function in all cases, hence our main focus on ligand-binding residues where credit can be given for residues in close proximity, accounting for structural flexibility. This focus will better enable refinements to be made to FunFOLD3.

The results produced by FunFOLDQ and the assessment by CAFA show there are different levels of benchmarks which can be applied. Initially, new protein sequences have no annotations and later GO terms can be associated within the different categories. Additionally, there could be limited knowledge related to a protein with just annotations related to one of three categories and further data around other categories adding to the knowledge, this is shown in Figure 5.7. By way of example, FunFOLDQ is predicting terms indicated by red circles, whereas CAFA3 is predicting terms in blue. The results from the participation in CAFA3 show that there is no leeway, results are either right or wrong and currently there is no other way to objectively measure GO terms. Due to this, it was decided to refine FunFOLD3 around the predicted ligand and this will be discussed in Chapter 6. Therefore, the results of this chapter will not be used in Chapter 6, as Chapter 6 will focus on the development of FunFOLD3, in terms of ligands and ligand-binding site residues, as opposed to the prediction of GO terms as presented in this chapter.

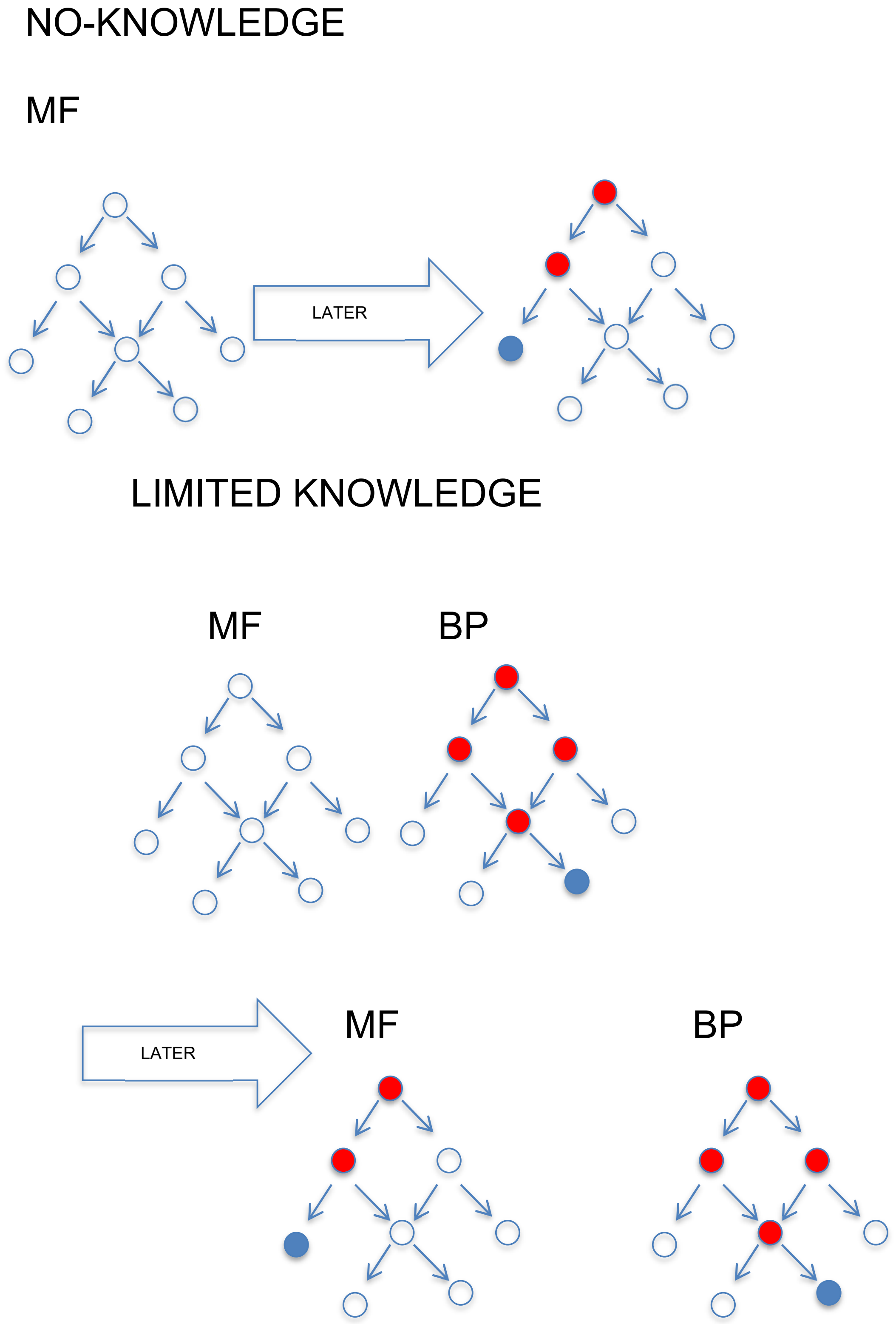


Figure 5.7. Types of benchmarking
A schematic diagram presenting the evolution of knowledge in the understanding of proteins. The red circes represent the level of GO term prediction, predicted by FunFOLDQ, whereas the blue circles present the GO term prediction which would have been benchmarked by CAFA3. MF is molecular fucntion and BP is biological process

GoFDR is a method which scored highly in the CAFA2 experiment and is one of the few methods where a source code has been published (Kriválk & Hoksza, 2018). The authors state the two key steps in GoFDR is the identification of GO term-specific FDRs from the query sequence and another is the raw score adjustment (Gong, Ning and Tian, 2016).

GoFDR is a sequence alignment-based algorithm, so differs from FunFOLDQ in this instance and adopts the functionally discriminating residues (FDRs) that has been used previously by EFICAz for predicting protein function. The FDRs defined in GoFDR are determined through comparing sequence conservation within sequences with the GO term to those sequences without the GO term and are therefore specific to the target GO term (Gong, Ning and Tian, 2016). However, what makes GoFDR different is avoidance of construction of multiple sequence alignments (MSA) for the protein function annotations defined by GO consortium, as it was deemed not practical to prepare high quality MSA for each GO term. Instead, the researchers use query sequence-based MSA directly from PSI-BLAST output, (Gong, Ning and Tian, 2016) GoFDR then identifies all GO terms associated with the sequences in the MSA and determines the FDRs for each GO term from which a specific a position specific scoring matrix (PSSM) is constructed.

The GoFDR algorithm consists of four steps:

1) Preparation of a query sequence-based MSA by database search
2) Identification of FDRs for a given GO term and construction of the FDR-PSSM and involves mapping GO annotations to the sequences in the MSA and identifies all relevant GO terms to be predicted
3) Scoring the query protein using the FDR-PSSM
4) Raw score adjustment

In brief, the workflow of GoFDR takes the input of a query sequence-based MSA produced directly from BLAST or PSI-BLAST search. After mapping GO annotations to all homologous sequences in the MSA, GoFDR identifies all relevant GO terms to be predicted. For each GO term, GoFDR compares the sequence conservation in aligned sequences with the GO term (homo-functional) and the aligned sequences without the GO term (hetero-functional) to identify a number of FDRs for the for distinguishing the homo-functional sequences from the hetero-functional sequences (Gong, Ning and Tian, 2016). An exact match is required to infer function. The workflow of GoFDR is outlined in Figure 5.8 below:

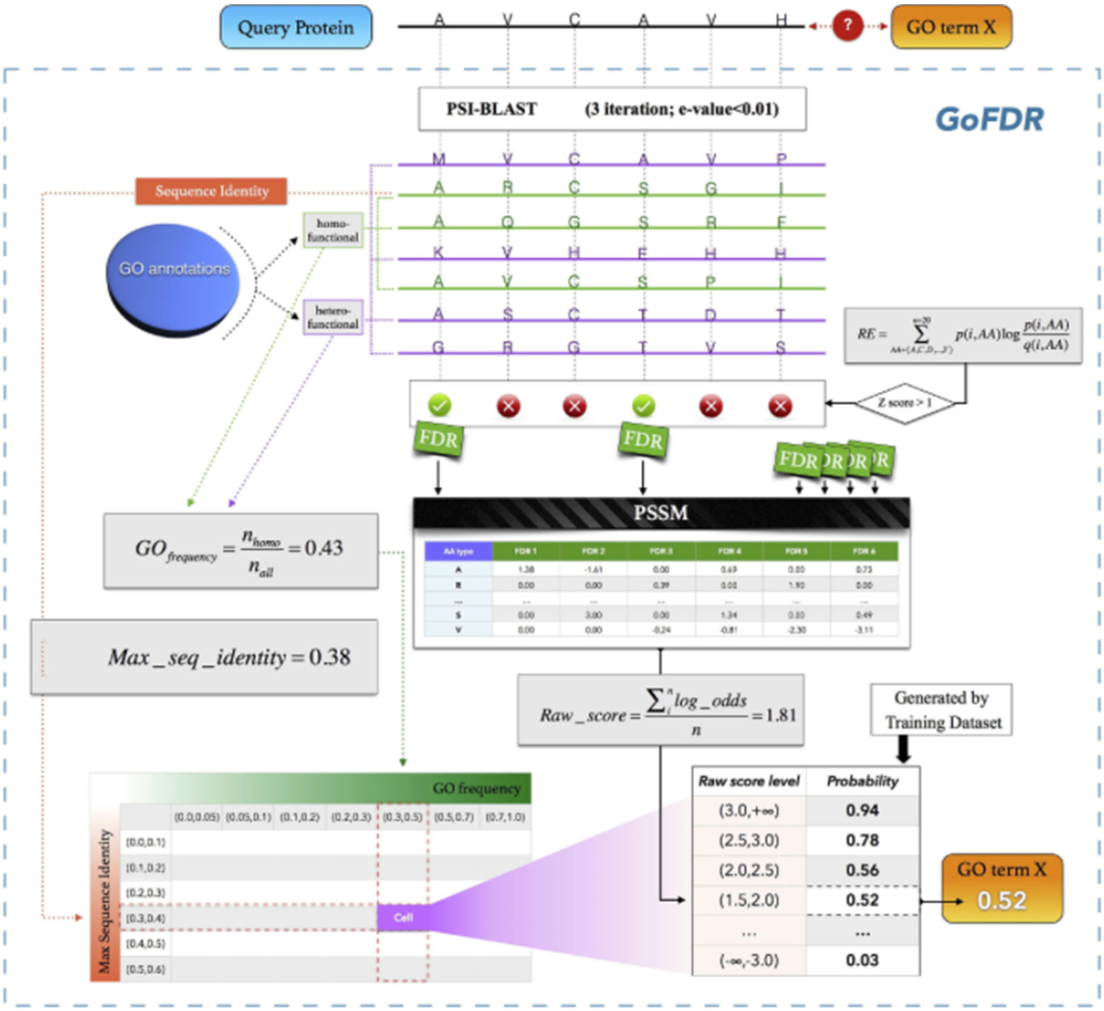


Figure 5.8. Workflow of GoFDR
GoFDR takes the input of a query sequence-based MSA produced directly from BLAST or PSI-BLAST search. After mapping GO annotations to all homologous sequences in the MSA, GoFDR identifies all relevant GO terms to be predicted. GoFDR builds a PSSM for the FDRs and then applies the PSSM to score the query sequence for its association with the target GO term. Finally, GoFDR converts the raw score of a prediction into a probability according to a pre-constructed score-to-probability table from training sequences. Figure taken from Gong, Ning and Tian, 2016

There is a crucial stage in GoFDR, which could potentially explain why GoFDR are among one of the top-scoring groups and more pressing could explain the lack of specificity of GO terms that were predicted by FunFOLDQ. This is the conversion of all the prediction scores of a query sequence into probabilities. This is the probability for each GO term by considering the parent-child GO term relationship i.e. the probability for a given GO term shouldn't be less than that for any of its child terms and if such case was found, then the probability of that GO term would be replaced by the probability of its child term. By way of comparison, FunFOLDQ, utilises HHblits, (HMM-HMM-based lightning-fast iterative sequence search) (Gong, Ning and Tian, 2016) and at its core builds multiple sequence alignments, which is ideal for building the 3D structure of a query protein and the foundation for HHblits; HHsearch is used by many of the best protein structure prediction servers for template-based protein structure prediction (Remmert *et al.*, 2012). It is worth pointing out, that the prediction of GO terms is based upon building an accurate a 3D structure as possible but to perform well in CAFA competitions more emphasis is needed on the prediction of GO terms. A future refinement of FunFOLDQ could be adding this stage, as it was a finding presented earlier in this chapter, that the GO terms were not specific enough and parent terms were being predicted, when there were associated child terms which were the correct predictions and more specific.

In comparison to GoFDR, FunFOLDQ relies on the identification of distant templates, so a strong sequence signal is not needed, however GoFDR requires sequence identity and therefore may not work as well when few PSI-BLAST hits are available. Despite not being obvious with the analysis of CAFA3 results, FunFOLDQ is structure based and this is a strength as if no sequence match the GoFDR matches would fail.

Overall, this chapter has shown that function prediction is difficult and no one method cannot be relied upon for solely evaluating function. Due to the mixture of results obtained with

FunFOLDQ and the lack of manoeuvring for correct results in the CAFA competition, it was deemed FunFOLD3 would be a suitable method to refine further and try and improve the prediction of function utilising docking and this will be explored further in Chapter 6.

Chapter 6: Refinement of FunFOLD3 ligand-binding predictions using AutoDock Vina

6.1 Introduction

Experimental techniques to investigate thermodynamic profiles for a ligand-protein complex can be laborious, time consuming and expensive. Conversely, computational protein-ligand docking methods can quickly predict the most favourable structure of the complex and assess the binding affinity (Du *et al.*, 2016).

Molecular docking is a computational method that attempts to predict non-covalent binding in macromolecular complexes, which most frequently are receptor macromolecules and a smaller molecule, such as a ligand, and usually starts with the unbound macromolecule structure (QuickGO, 2017). Molecular docking is widely used, relatively fast and an economical computational tool for predicting *in silico* the bindings modes and affinities of molecular recognition events. Protein-ligand docking is a branch of this field and is of particular importance due to its utilisation in drug discovery processes. AutoDock (Morris *et al.*, 2009), GOLD (Jones *et al.*, 1997), DOCK (Ewing et al., 2001, Allen et al., 2015), FlexX (Matthias et al., 1996) and Glide (Friesner *et al.*, 2004) are examples of well-established methods that exist and each method implements different algorithms to solve the docking problem (Pagadala, Syed and Tuszynski, 2017).

Protein-ligand docking plays an important role in predicting the orientation of a ligand when it is bound to a receptor using shape and electrostatic interactions to quantify it (Trott & Olson, 2010). Therefore, the goal of protein-ligand docking is to predict the bound conformations and the binding affinity for a protein to its ligand(s) (Pagadala, Syed and Tuszynski, 2017). In general protein-ligand docking methods contain two essential components; the search algorithm and the scoring function. Search algorithms are responsible for searching through different ligand conformations and orientations, sometimes referred to as poses within a given target protein and scoring functions are responsible for estimating the binding affinities of the generated poses. Some methods will rank these poses and identify the most favourable binding mode(s) of the ligand to the given target (Lee and Im, 2013). Theoretically, the search space for protein-ligand binding should consist of all possible conformations of the protein and the ligand in their unbound forms, all possible orientations and conformations of the ligand within a given protein conformational state and all possible conformations of the protein paired with all the possible conformations of the ligand. As one can imagine, it is impossible to exhaustively explore the search space given the limitations of current computational power and search algorithms (Pagadala, Syed and Tuszynski, 2017).

The earliest docking methods focused on the classical protein-ligand relationship, of a "lock-and-key" assumption. The ligand and the receptor are treated as rigid bodies and their affinity is directly proportional to a geometric fit between their shapes (Trott & Olson, 2010). Rigid-body algorithms are the simplest approach when it comes to sampling the conformational space. The methods used are ZDOCK, MDock and older versions of DOCK. On the other hand, flexible-ligand algorithms consider only the ligand flexibility; and neglect the protein flexibility, DOCK is an example of this (Ewing *et al.*, 2001).

As understanding about protein-ligand interactions increased and that conformational changes occur in the receptor to bind the ligand, the "lock-and-key" theory moved onto the induced-fit theory, in which the ligand and receptor should be treated as flexible during docking and was proposed by Koshland (Mezei, 2003). Over the past twenty years, more than 60 different docking tools have been developed (Hammes, 2002,Trott & Olson, 2010), and can be broadly categorised into the following algorithms; incremental construction approaches (Rarey et al., 1996), shape-based, genetic, systematic search techniques and Monte Carlo simulations (Rarey et al., 1996).

The binding affinity of a ligand and protein can be affected by the entropy loss of a flexible ligand in a rigid anisotropic environment of the receptor and the change in internal energy upon binding (Venkatachalam *et al.*, 2003). As a result of this, information regarding binding site location before the docking processes becomes very important in increasing docking efficiency (Pagadala, Syed and Tuszynski, 2017). However, the caveat to this is that, it depends on the accuracy of the binding site predictions as well as the method which has been used to predict protein function and more importantly if the protein has no sequence homology. One of the common issues with FunFOLD3 predictions has been the variety in quality of the ligand-binding site predictions, as has been seen in the analysis of CASP11-13. The MCC and BDT score can vary quite markedly from 0.877 and 0.853, respectively (e.g. for CASP11 target T0819) being the highest score and -0.05 and 0.0375, respectively (for CASP11 target T0849) being the lowest score.

AutoDock Vina uses flexible docking and in standard virtual docking studies, ligands are freely docked into a rigid receptor with the number of active rotatable bonds ranging from 0-32 (Pagadala, Syed and Tuszynski, 2017). However, it is becoming increasingly clear that side chain flexibility plays a crucial role in ligand-protein complexes. These changes allow the receptor to alter its binding site according to the orientation of the ligand. The ligand orients in a space within the binding site which is translational, rotational and conformational variable in the anisotropic environment of the receptor (Trott & Olson, 2010). This will compliment FunFOLD3, as FunFOLD3 will predict the ligand-binding site and then AutoDock Vina will orientate the ligand into different conformations and pick the “best one”. Subsequently we can determine how similar this new orientation is to the original predicted ligand location. The next step would be to confirm the MCC and BDT score to determine if there has been any improvement based on the orientation of the ligand.

The utilisation of molecular docking using AutoDock Vina to refine ligand-binding poses has been studied previously (Wu *et al.*, 2018). However, COACH-D failed to demonstrate a statistically significant difference in the improvement of ligand-binding sites between COACH and COACH-D, and the major improvement was in resolving steric clashes. As a result, of AutoDock Vina being used to refine ligand-binding poses it has been chosen to do the same with FunFOLD3. Futhermore, AutoDock Vina was chosen as the protein-ligand method of choice, due to the ability to specify a search box for the simulation which can manually be adjusted by the user, thus enabling the search algorithm to search only within a specific area. Thereby, reducing two critical elements; speed and effectiveness in covering the relevant conformational space (Du *et al.*, 2016).

Aim: The aim of this chapter was to determine if ligand-binding site predictions by FunFOLD3 could be improved following docking with AutoDock Vina. Improvements in ligand-binding site predictions will be reported as a change in MCC and BDT score following docking and this will act as an objective measure to determine if there has been any improvements Therefore, the overall aim of this chapter is to “fix” the problem, i.e. the ligand-binding site residues, in order to obtain ligand-binding site predictions which are closer to the observed ligand-binding site residues. This chapter is based upon the results provided mainly in Chapter 3 and to a lesser extent Chapter 4. The CASP11, 12 and 13 experiments provided a basis to identify the “problem” by providing unsolved protein targets, to enable prediction of ligand and ligand-binding site residues. As the protein targets would eventually have solved structures and where applicable solved ligands, this provided an “ideal” to work towards with improving ligand-binding site predictions. As a result of developing FunFOLD3 into FunFOLD3-D, this chapter is the main chapter of thesis.

6.2 Materials & Methods

6.2.1 Materials

As mentioned previously in Chapter 3, the CASP organisers provided amino acid sequences and prediction of biologically relevant ligands is performed by FunFOLD3. In this Chapter, refinements will be made to predictions in CASP11, CASP12 and CASP13. The amino acid sequences for these targets have been provided previously in Chapter 3 and is shown in Table 3.1.

6.2.2 Methods

AutoDock Vina has been chosen to refine the ligand-binding site for predictions as it has been used previously for similar purposes with some success (Wu *et al.*, 2018). One of the grid sizes for the ligand-binding pockets search space will be 22.5Å in total as this grid space has been explored previously in the literature (Feinstein and Brylinski, 2015). This also ensures consistency with the earlier tests of AutoDock, where the 22.5Å grid size was chosen and the developers of AutoDock recommend making sure that the search space is large enough for the ligand to be able to rotate inside. In order to further expand refinement with docking, three different percentages will be used as a maximum ligand distance of 10%, 20% and 50% of the volume of the box around the ligand space to help determine if a specific distance around the predicted ligand provides better ligand-binding site predictions. Figure 6.1 below is the grid box space for CASP11 target T0849, two examples are given; 10% of the volume of the box around the ligand and 22.5Å.

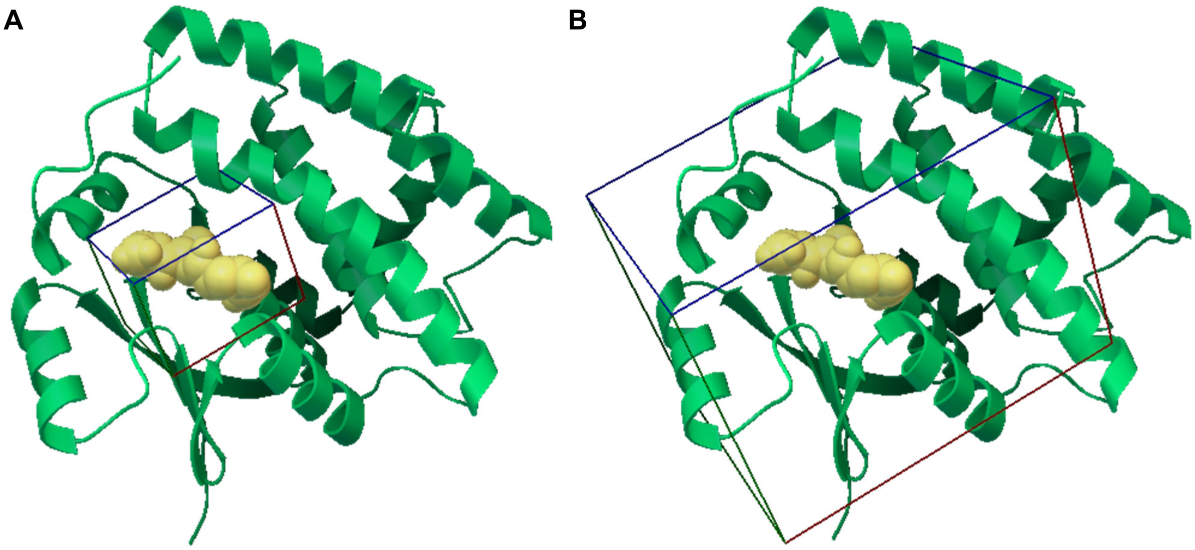


Figure 6.1. 10% and 22.5Å grid box for CASP11 target T0849
(A) Predicted 3D model for CASP11 T0849 coloured green and shown as cartoon and the predicted ligand GSH is shown as sphere and coloured yellow. The grid box volume is 10% around the ligand space. **(B)** Predicted 3D mode for CASP11 target T0849 with a 22.5Å grid box volume around the ligand space**.**

The methodology for docking using AutoDock Vina is given below and the input for AutoDock Vina is the output ligand and receptor file (referred to as the lig2.pdb file. See Chapter 2, point 13C) from FunFOLD3. Therefore, in order to complete the methodology below, the method outlined in Chapter 2 needs to be completed first, or alternatively, a ligand and protein file can be obtained elsewhere.

The purpose of docking was to determine if the ligand-binding site residue prediction, as assessed using the MCC and BDT scores can be improved with AutoDock Vina by finding the best rotation of the ligand within the ligand-binding space.

The steps for using AutoDock Vina are described below:

1. AutoDock Vina can be freely downloaded from: http://vina.scripps.edu/download.html
2. Separate the ligand and receptor into two separate files

3. Prepare ligand and receptor .pdbqt files using the ./prepare_ligand4.py -l and ./prepare_recetopr4.py commands, respectively. This will generate a ligand.pdbqt and recetopr.pdbqt file which will be used to generate the new ligand rotations.
4. Set docking parameters to obtain the required grid box calculation space e.g. grid box of 10% bigger than current ligand-binding space. Utilising the command line ./Centre_v3.py. The user can use any grid box they wish. It is recommended to use 10, 20 or 50%, as it has been explored in this thesis. The source code is provided in Appendix 4
5. The grid box calculation will appear in terminal and the centroid box sizes in x, y and z plane and size for x,y,z plane should be used to determine the new ligand-binding space.
6. The grid box calculation will need to be added to the conf.txt file and also the maximum number of ligand models to generate. The recommended number is nine which has been investigated as part of this thesis
6. Run the vina shell script (./vina_screen_local.sh) and output files will be all the ligand models.
7. Convert the ligand and receptor files back to PDB files and merge the different ligand models with the receptor.
8. Analyse the ligand-binding residues with the preferred functional prediction method and use an objective scoring method to determine if an improvement has been obtained.

These newly adjusted ligand-binding site residue predictions were then compared against the actual ligand binding site residues and a new MCC and BDT scores were calculated. These adjusted binding site scores were then compared against the MCC and BDT scores for the original FunFOLD3 ligand-binding residue predictions, to determine if there has been an objective improvement in ligand-binding site predictions following the docking procedure. The new version of FunFOLD3 which included refinement of the protein-ligand complexes with docking was called FunFOLD3-D.

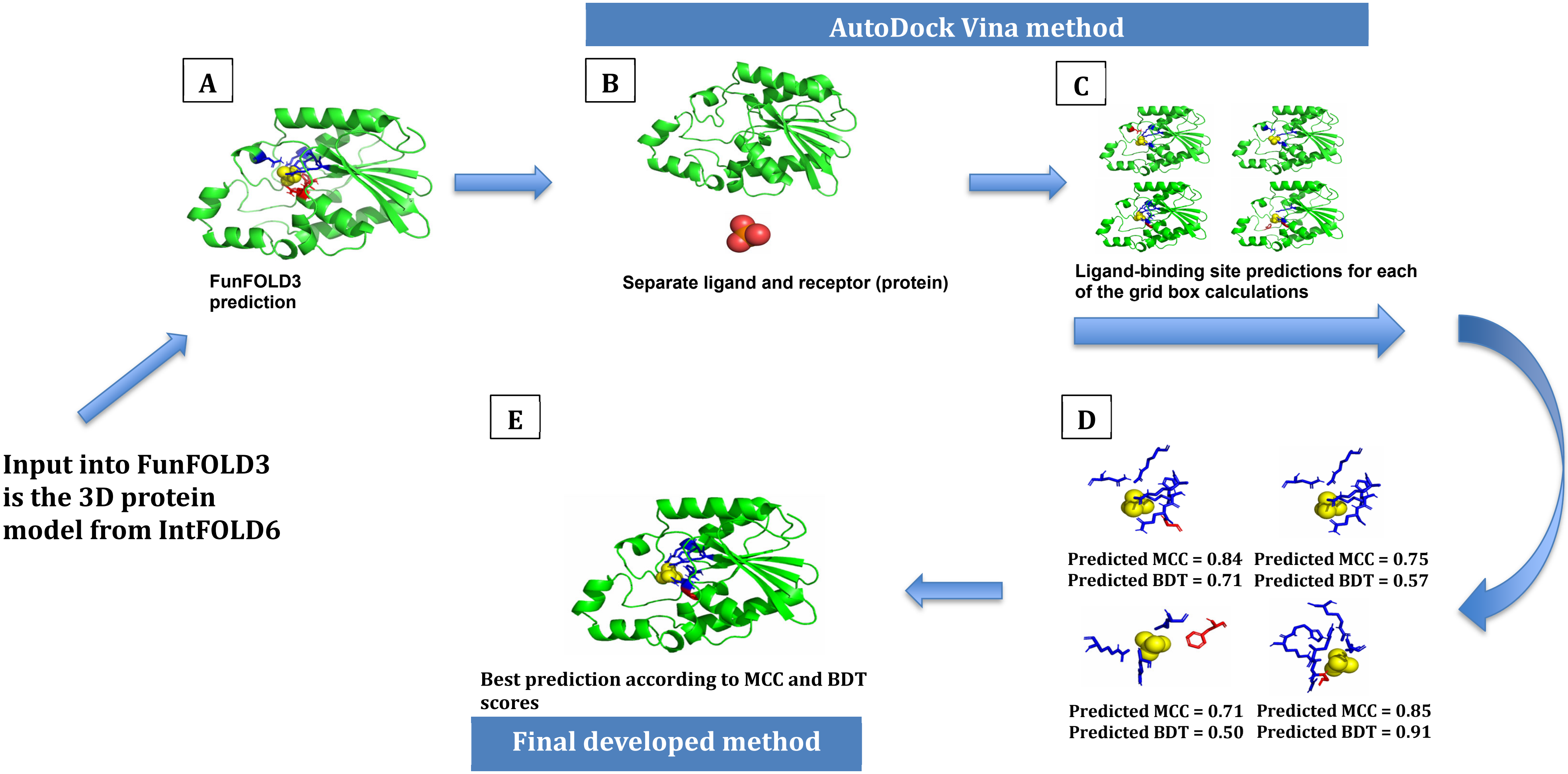


Figure 6.2. Graphical abstract to demonstrate the flow of IntFOLD, FunFOLD3 and AutoDock Vina
(A) Input into AutoDock Vina is the FunFOLD3 ligand and receptor (protein) file with the 3D protein model being an output from the IntFOLD server **(B)** Separate ligand and receptor (protein) PDB files are required for input into AutoDock Vina **(C)** Nine different ligand-binding site predictions are produce for each grid box calculation, for simplicity the top scoring models have been shown **(D)** The top scoring models are determined by the MCC and BDT scores **(E)** The best prediction will be the model which has the highest MCC and BDT scores and is the final developed method. The putative plans are to incorporate AutoDock Vina into future workflows of FunFOLD3

6.3 Results and discussion

6.3.1 Summary of results

The main findings in this chapter were

- The developed method, FunFOLD3-D utilised four different grid boxes (10%, 20%, 50% and 22.5Å) in order to determine how to improve the ligand-binding site residues by FunFOLD3. The main finding was there is no “one size fits all” with regard to a grid box. It very much depends on the ligand and where the ligand has been predicted in relation to the observed ligand and the observed ligand-binding space. For example, if the predicted ligand is distal to the observed ligand, a larger grid box is required and converse (refer to Figure 6.3)
- Based on the above, an adaptable grid box is required and this challenges literature, somewhat, which identified a grid box of 22.5Å being the most suitable (Feinstein & Brylinski, 2015). However, in certain predictions a smaller grid box was required
- FunFOLD3 predictions with an MCC and BDT score of around >0.70 were unable to be improved further with docking. Therefore, this suggests there is a threshold when it comes to docking. In essence, docking can’t improve good predictions even further

6.3.2 Analysis of CASP11, CASP12 and CASP13 targets by AutoDock Vina

Out of a total of 55 targets from CASP11, CASP12 and CASP13 that had ligand-binding site predictions by FunFOLD3, 18 targets were docked using AutoDock Vina. Of the nine targets from CASP11, six were successfully docked. Successfully docked was defined as the conversion of both the receptor and ligand to PDBQT files in order for docking to occur. Thus meaning three targets were unable to be docked; T0845 (PDB ID 4r5o) due to multiple metal ions being predicted, so there is no specific ligand space to focus on. With respect to T0854 (PDB ID 4rn3), the MG metal ion and the ZN metal ion for T0786 (PDB ID 4qvu) had be excluded as the current Babel-based method used by AutoDock does not handle metal charges (Trott & Olson, 2010). Therefore, for metal ligands a different method will need to be explored.

For CASP12, of the 12 targets which FunFOLD3 predicted ligands, two were excluded as there were no biologically relevant ligands were found in the observed structure, this was T0868 (PDB ID 5j4a) and T0872 (PDB ID 5jmb). Four targets; T0899, T0901, T0905, T0907 did not have PDB IDs associated and another target; T0919 had the structure cancelled by CASP organisers. This left five targets which could be docked by AutoDock Vina, however two targets T0909 (PDB ID 5g5n) and T0911 (PDB ID 6e9n) were unable to be docked by AutoDock Vina due to Gasteiger charges not being able to be added to the ligands and thus convert to PDBQT files and polar hydrogens are necessary for docking by AutoDock Vina.

With respect to CASP13, of the 34 targets which FunFOLD3 predicted ligands, eleven had to be excluded as no observed structure was released to do observed ligand-binding site residues by CASP organisers; T0949, T0973 (PDB ID 6yfn), T0975, T0985, T0995, T0997, T1001, T1013, T1017s1 and T1023s3. Nine targets had no observed biologically relevant ligands; T0955 (PDB ID 5w9f), T0957s2 (PDB ID 6cp8), T0958 (PDB ID 6btc), T0970 (PDB

ID 6g57), T0980s1 (PDB ID 6gnx), T0980s2 (PDB ID 6gnx), T0986s2 (PDB ID 6d2y), T0993s1 (PDB ID 6xbd) and T1008 (PDB ID 6msp). Three targets had the observed structure cancelled by CASP organisers; T0972, T0994 and T1012. One target had a perfect prediction despite a different between the predicted an observed ligand so was excluded from docking; T0974s1 (PDB ID 6tri). This left a total of ten targets for docking of which two (T0961 (PDB ID 6sd8) and T1018) didn't work. The reason for T1018 not working could be to the ligand being a metal ion and AutoDock Vina is unable to assign coordinates for metal ions, as mentioned previously. With regards to T0961, it could be due to the location of the ligand, FunFOLD3 predicted the ligand in one ligand-binding space but there are two locations for the ligand. Steps taken to rectify this were to separate out the ligands in PyMOL, however when adding partial charges in AutoDock Vina, Gasteiger parameters could not be added.

6.3.3 Summary of docking results using AutoDock Vina

Tables 6.1, 6.2 and 6.3 below show the best MCC and BDT scores obtained for CASP11, CASP12 and CASP13 targets which were docked using AutoDock Vina. The MCC and BDT scores obtained by FunFOLD3 are given, along with the new MCC and BDT scores using the docking method. Improvements in MCC and BDT scores are highlighted in black and bold. The top five predictions will be shown and for remaining results refer to Appendix 4.

Table 6.1. Summary of the comparison between the results obtained with FunFOLD3 and FunFOLD3-D

CASP11 target	FunFOLD3 prediction		FunFOLD3-D prediction		Grid box parameters
	MCC	BDT	MCC	BDT	
T0783	Mg = 0.17 Cl =0.015	Mg = 0.21 Cl=0.1	**0.63**	**0.45**	22.5Å
T0798	0.753	0.797	0.75	0.65	10%
T0807	0.771	0.849	0.49	0.43	10%
T0813 (NAD)	0.086	0.19	0.46	0.43	50%
T0813 (NAI)	-0.029	0.11	**0.34**	**0.28**	22.5Å
T0813 (NAP)	0.079	0.20	**0.34**	**0.27**	10%
T0819	0.877	0.853	0.87	0.80	10%
T0849	-0.05	0.0375	**0.0086**	**0.24**	22.5Å

Table 6.2. Summary of the comparison between the results obtained with FunFOLD3 and FunFOLD3-D

CASP12 target	FunFOLD3 prediction		FunFOLD3-D prediction		Grid box parameters
	MCC	BDT	MCC	BDT	
T0912 - FRU	-0.00672	0.0295	-0.006	0.076	22.5Å
T0912 – MAV	-0.00892	0.0213	-0.008	0.026	22.5Å
T0913	-0.0367	0.091	-0.04	0.13	20%
T0916	GLC(1) = 0.162 GLC(2) = 0.263	GLC(1)=0.276 GLC(2)=0.37	GLC(2)=0.16	GLC(2)=0.32	50%

Table 6.3. Summary of the comparison between the results obtained with FunFOLD3 and FunFOLD3-D

CASP13 target	FunFOLD3 prediction		FunFOLD3-D prediction		Grid box parameters
	MCC	BDT	MCC	BDT	
T0953s2	0.12	0.11	-0.010	0.21	20%
T0954	-0.015	0.028	-0.015	0.028	10%
T0965	0.12	0.35	**0.22**	**0.33**	22.5Å
T0983	0.715	0.715	**0.83**	**0.71**	50%
T1003	-0.04	0.06	**0,06**	**0.15**	22.5 Å
T1009	0.91	0.94	0.61	0.49	20%
T1014	-0.05	0.05	**0.05**	**0.13**	10%
T1016	0.556	0.646	**0.85**	**0.91**	22.5Å

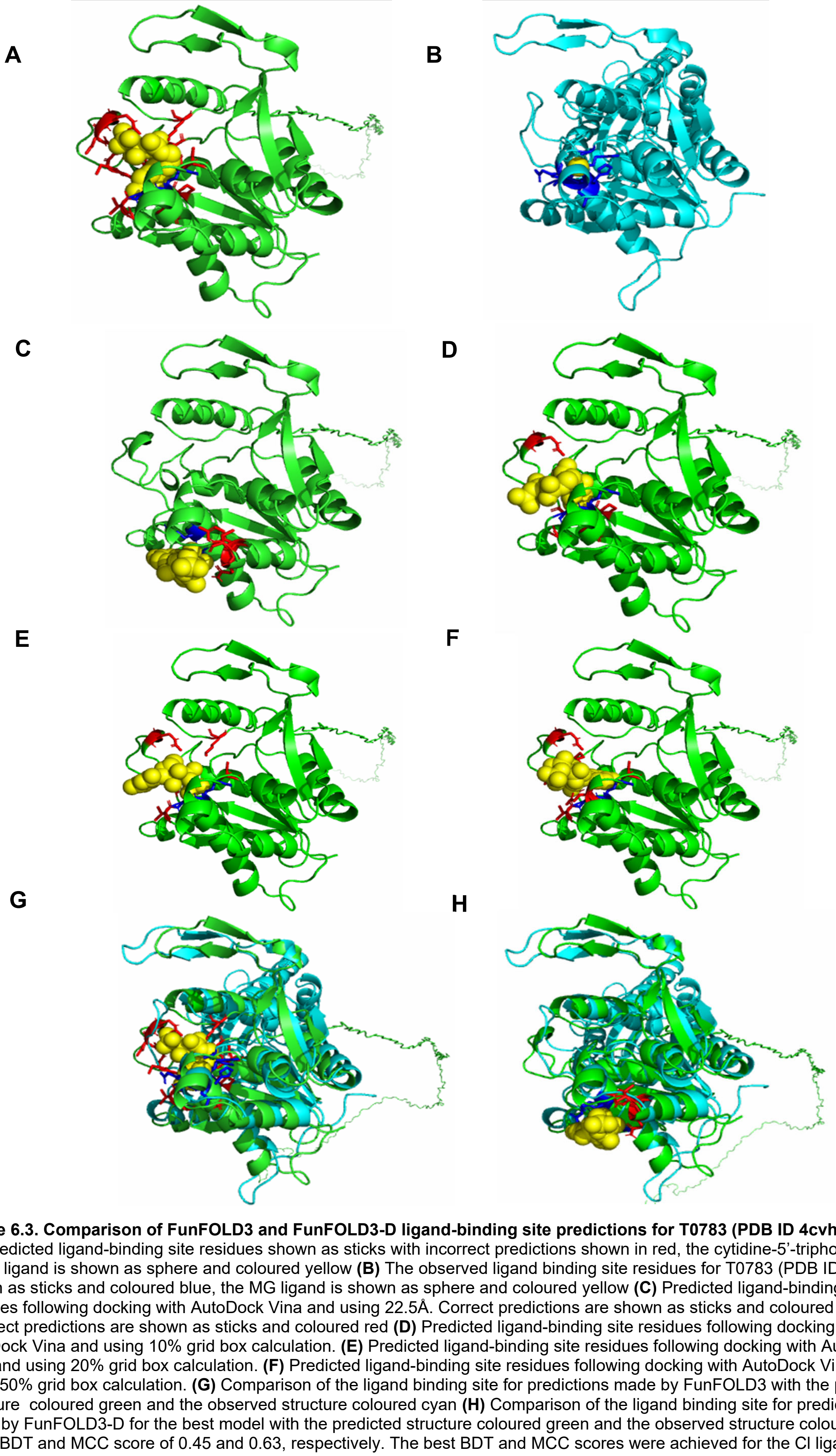


Figure 6.3. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0783 (PDB ID 4cvh)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the cytidine-5'-triphosphate (CTP) ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0783 (PDB ID 4cvh) shown as sticks and coloured blue, the MG ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the predicted structure coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model with the predicted structure coloured green and the observed structure coloured cyan. BDT and MCC score of 0.45 and 0.63, respectively. The best BDT and MCC scores were achieved for the Cl ligand using 22.5Å

Results from docking with this target is given below in the Tables 6.4-6.7. Despite Mg being the observed ligand and an incorrect ligand being predicted, the procedure could still improve the prediction of binding site residues. The Cl ligand can be found below in Figure 6.4 with a comparison the FunFOLD3 method. The predicted ligand-binding residues by FunFOLD3 were 12,13,14,15,16,17,18,19,26,27,83,84,**85,86**,89,116,117,118,223 (correct residues in red) and the observed ligand-binding residues were 85,86,87,194,195,197 for Mg ligand and 88,93 for Cl ligand. The MCC and BDT scores were 0.17 and 0.21, respectively when compared to the Mg ligand.

Table 6.4. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for CASP11 T0783 (PDB ID 4vch)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å. MCC and BDT scores are given against both the Mg and Cl predicted ligands

Model number	Pose	Predicted ligand-binding site residues	MCC		Score change		BDT		Score change	
			Mg	Cl	Mg	Cl	Mg	Cl	Mg	Cl
1	1	16,17,18,26,84,85,86,117,118	0.26	-0.01	0.09	0.005	0.30	0.093	0.09	-0.007
	2	12,16,18,26,84,86,116,117,118	0.12	-0.01	-0.05	0.005	0.21	0.078	0.00	-0.022
2	1	12,13,14,15,26,84,85,86,195,198	0.38	-0.01	0.21	0.005	0.44	0.11	0.23	0.010
	2	13,14,15,26,85,86,118,195,198,223	0.38	-0.01	0.21	0.005	0.39	0.092	0.18	-0.008
3	1	15,16,17,19,85,86,173,198,199,223	0.24	-0.01	0.07	0.005	0.30	0.084	0.09	-0.016
	2	16,84,85,86,119,171,173,223	0.28	-0.01	0.11	0.005	0.33	0.10	0.12	0.020
4	1	19,144,170,171,172,225	-0.015	-0.01	-0.185	0.005	0.036	0.02	-0.17	-0.274
	2	19,144,170,171,172,222,223,225	-0.017	-0.01	-0.187	0.005	0.036	0.02	-0.17	-0.27
5	1	86,119,144,171,172,199,223	0.14	-0.01	-0.030	0.005	0.20	0.059	-0.01	-0.11
	2	119,144,171,172,198,199,221,222	-0.017	-0.01	-0.187	0.005	0.10	0.034	-0.11	-0.21
6	1	144,171,222,223,226,229	-0.015	-0.009	-0.185	0.006	0.030	0.018	-0.18	-0.082
	2	119,170,171,223,226,228	-0.014	-0.009	-0.184	0.006	0.033	0.019	-0.18	-0.081
7	1	88,92,186,187,189,190,191	-0.016	0.53	-0.186	0.545	0.20	0.33	-0.01	0.230
	2	**88,92,186,187,189**	-0.014	**0.63**	-0.184	**0.645**	0.15	**0.45**	-0.06	**0.350**
8	1	119,171,222,223	-0.012	-0.007	-0.182	0.008	-0.007	0.022	-0.22	-0.078
	2	119,144,147,170,199,221,222,223	-0.017	-0.01	-0.187	0.005	0.052	0.023	-0.16	-0.077
9	1	15,16,19,26,85,86,116,117,118,195,198	0.36	-0.012	0.19	0.003	0.35	0.082	0.14	0.040
	2	12,15,16,26,85,116,117,118,195,198	0.24	0.30	0.07	0.315	-0.01	0.073	-0.22	-0.320

Table 6.5. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for CASP11 T0783 (PDB ID 4vch)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation is based on 10%. MCC and BDT scores are given against both the Mg and Cl predicted ligands

Model number	Pose	Predicted ligand-binding site residues	MCC		Score change		BDT		Score change	
			Mg	Cl	Mg	Cl	Mg	Cl	Mg	Cl
1	1	15,16,26,84,85,86,116,118,223	0.25	-0.097	0.08	-0.082	0.31	0.097	0.10	0.00
	2	12,13,14,15,16,18,84,86,116,117,118	0.11	-0.011	-0.06	0.004	0.21	0.081	0.00	-0.02
2	1	12,15,16,19,26,86,116	0.14	-0.01	-0.03	0.005	0.22	0.075	0.010	-0.93
	2	12,15,16,26,84,116,117,118	-0.017	-0.01	-0.19	0.005	0.13	0.063	-0.080	-0.04
3	1	13,14,15,16,17,18,19,26,86,223	0.11	-0.01	-0.060	0.005	0.16	0.062	-0.05	-0.04
	2	15,16,18,19,26,84,85,86,116,223	0.24	-0.01	0.070	0.005	0.29	0.09	0.08	-0.01
4	1	12,15,16,18,27,84,85,86,117,118,223	0.23	-0.01	0.060	0.005	0.27	0.09	0.06	-0.01
	2	12,15,16,19,26,27,84,85,86,116,117,118,223	0.21	-0.01	0.040	0.005	0.24	0.08	0.03	-0.02
5	1	14,15,16,19,26,85,86	0.29	-0.01	0.120	0.005	0.35	0.10	0.14	0.00
	2	12,14,15,16,19,84,85,86	0.28	-0.01	0.110	0.005	0.37	0.11	0.16	0.01
6	1	12,15,16,19,84,85,86,116	0.28	-0.01	0.110	0.005	0.36	0.11	0.15	0.01
	2	12,14,15,19,26,84,85,86,116,117,118,223	0.22	-0.01	0.05	0.005	0.26	0.09	0.05	-0.01
7	1	12,13,15,18,19,26,27,84,85,86,116,117,118,223	0.20	-0.013	0.030	0.002	0.23	0.08	0.02	-0.02
	2	12,13,15,16,18,19,26,84,85,86,116,117,118,223	0.20	-0.013	0.030	0.002	0.23	0.08	0.02	-0.02
8	1	12,13,15,19,84,85,86,116,117,223	0.24	-0.01	0.07	0.005	0.30	0.10	0.09	0.00
	2	12,13,15,16,18,19,26,27,84,85,116,117,118,223	0.1	-0.01	-0.07	0.005	0.2	0.06	-0.01	-0.04
9	**1**	**12,13,14,19,84,85,86,89**	**0.28**	-0.01	**0.11**	0.005	**0.39**	0.16	**0.18**	0.06
	2	12,13,14,15,16,19,26,84,85,89	0.11	-0.01	-0.06	0.005	0.24	0.12	0.03	0.02

Table 6.6. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for CASP11 T0783 (PDB ID 4vch)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation is based on 20%. MCC and BDT scores are given against both the Mg and Cl predicted ligands

Model number	Pose	Predicted ligand-binding site residues	MCC		Score change		BDT		Score change	
			Mg	Cl	Mg	Cl	Mg	Cl	Mg	Cl
1	1	12,15,16,18,84,85,86,117,118	0.26	-0.01	0.090	0.005	0.32	0.10	0.11	0.09
	2	12,15,16,18,84,86,117,118,223	0.11	-0.01	-0.060	0.005	0.21	0.08	0	-0.02
2	**1**	**15,16,84,85,86,116,223**	**0.30**	-0.01	**0.130**	0.005	**0.39**	0.12	**0.18**	0.020
	2	12,14,15,16,18,84,85,86,116	0.26	-0.01	0.090	0.005	0.34	0.11	0.13	0.010
3	1	12,15,16,18,84,85,86,117,118,223	0.24	-0.01	0.070	0.005	0.30	0.10	0.090	0.000
	2	12,16,19,26,27,84,85,86,116,117,118,223	0.22	-0.01	0.120	0.005	0.25	0.08	0.040	-0.020
4	1	12,15,16,19,84,85,86,116,223	0.26	-0.01	0.090	0.005	0.33	0.1	0.120	0.000
	2	12,15,19,26,84,86,116,118,223	0.12	-0.01	-0.050	0.005	0.22	0.08	0.010	-0.020
5	1	13,14,15,16,17,19,85,86,223	0.26	-0.01	0.090	0.005	0.29	0.09	0.080	-0.010
	2	15,19,84,85,86,116,223	0.30	-0.009	0.130	0.006	0.38	0.11	0.170	0.010
6	1	12,15,19,84,86,116	0.15	-0.009	-0.020	0.006	0.30	0.11	0.090	0.010
	2	12,15,26,84,86,116,118,223	0.13	-0.01	-0.040	0.005	0.24	0.09	0.030	-0.010
7	1	15,16,26,84,85,86,116,118,223	0.26	-0.01	0.090	0.005	0.31	0.1	0.10	0
	2	12,13,14,15,16,18,84,86,116,117,118	0.11	-0.01	-0.060	0.005	0.21	0.08	0	-0.02

Table 6.7. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for CASP11 T0783 (PDB ID 4vch)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation is based on 50%. MCC and BDT scores are given against both the Mg and Cl predicted ligands

Model	Pose	Predicted ligand-binding site residues	MCC		Score change		BDT		Score change	
			Mg	Cl	Mg	Cl	Mg	Cl	Mg	Cl
1	1	15,16,18,84,85,86,117,118	0.28	-0.01	0.11	0.01	0.35	0.11	0.14	0.01
	2	12,16,26,84,86,116,117,118	0.13	-0.01	-0.04	0.01	0.24	0.085	0.03	-0.015
2	1	13,14,15,16,18,19,85,86,223	0.26	-0.01	0.09	0.01	0.29	0.09	0.08	-0.01
	2	12,14,15,18,19,84,85, 86,116,223	0.24	-0.01	0.07	0.01	0.30	0.10	0.09	0
3	**1**	**15,16,19,26,84,85,86, 116,195**	**0.40**	-0.01	**0.23**	0.01	**0.42**	0.10	**0.21**	0
	2	12,15,16,84,86,116	0.15	-0.01	-0.02	0.01	0.31	0.11	0.1	0.01
4	1	15,16,84,85,86,116,223	0.30	-0.01	0.13	0.01	0.39	0.12	0.13	0.02
	2	12,14,15,16,18,84,85,86,116	0.26	-0.01	0.09	0.01	0.34	0.11	0.13	0.01
5	1	14,15,16,19,84,85,116, 117,118,223	0.11	-0.01	-0.06	0.01	0.20	0.08	-0.01	-0.02
	2	12,13,14,15,18,19,84,85,86,116, 117,118,223	0.21	-0.01	0.04	0.01	0.25	0.09	0.04	-0.01
6	1	15,16,19,26,84,85,86, 116,118,223	0.24	-0.01	0.07	0.01	0.29	0.09	0.08	-0.01
	2	12,13,14,15,16,17,18,27,86,116, 117,118	0.10	-0.01	-0.07	0.01	0.16	0.06	-0.05	-0.04
7	1	12,14,15,19,84,85,86	0.30	-0.01	0.13	0.01	0.41	0.13	0.2	0.03
	2	12,13,14,15,16,84,86	0.14	-0.01	-0.03	0.01	0.29	0.11	0.08	0.01
8	1	12,13,14,15,26,84,85,86	0.28	-0.01	0.11	0.01	0.37	0.12	0.16	0.02
	2	12,13,14,15,26,84,85,86,118,195,198,223	0.34	-0.01	0.17	0.01	0.37	0.10	0.16	0

The first CASP11 target to be docked was T0783 (PDB ID 4cvh). FunFOLD3 had predicted four biologically relevant ligands, with no predicted ligands matching the observed Mg and Cl ligands. However, both cytidine-5'-triphosphate (CTP) and cytidine-5'-monophosphate (C) had two correctly predicted ligand-binding residues (THR85 and ARG86). CTP was successfully docked but C was unable to be docked due to AutoDock being unable to add Gasteiger charges to the PDB file. Despite no correct ligand-binding site residues being predicted for the observed Cl ligand initially, MCC and BDT scores were calculated following docking to determine if there had been an improvement towards either of the observed ligands.

The predicted protein models from the four different grid box calculations along with the predicted ligand from FunFOLD3 and observed ligand are shown in Figure 6.3C-F. The best MCC and BDT models are depicted. All of the predicted ligand-binding site residues from the models with the respective MCC and BDT scores given in Tables 6.4-6.7.

Figure 6.3A shows the prediction using FunFOLD3 for the CTP ligand, a MCC and BDT score of 0.17 and 0.21 was achieved, respectively compared to the Mg ligand and -0.015 and 0.10, respectively compared to the Cl ligand (comparison to the Cl ligand is shown below in Figure 6.4). Overall, the best docked result was complex 7, pose two (Figure 6.3H) for grid box calculation 22.5A where an MCC and BDT score of 0.63 and 0.45, respectively was achieved. These scores were obtained for the Cl ligand, of which poor scores were achieved with FunFOLD3 (MCC and BDT -0.015 and 0.1, respectively). The increase in MCC and BDT was 4300% and 350%, respectively with the MCC and BDT scores increasing to 0.63 and 0.45, respectively. This was due to correct predictions for both of the ligand binding site residues (ARG88 and ASN92). As can be seen in Figure 6.3H, the improvement in ligand-binding came from the rotation of the ligand from a more distal location in relation to the observed ligand-binding residues to more proximal, which is also demonstrated by the correct ligand-binding residues. An important consideration with this target is the difference between molecular weight between the predicted and observed ligands, the CTP ligand has a molecular weight of 483.16, whereas Cl is 35.45 and Mg is 24.30. As a result of the difference in molecular weight, the predicted ligand will occupy a larger binding pocket and it is therefore reasonable to expect incorrect ligand-binding residues due to this. However, despite this difference the target was docked successfully with improvements against both the observed ligands. A more consistent improvement in MCC and BDT scores was observed with the Mg ligand, in comparison to the Cl ligand.

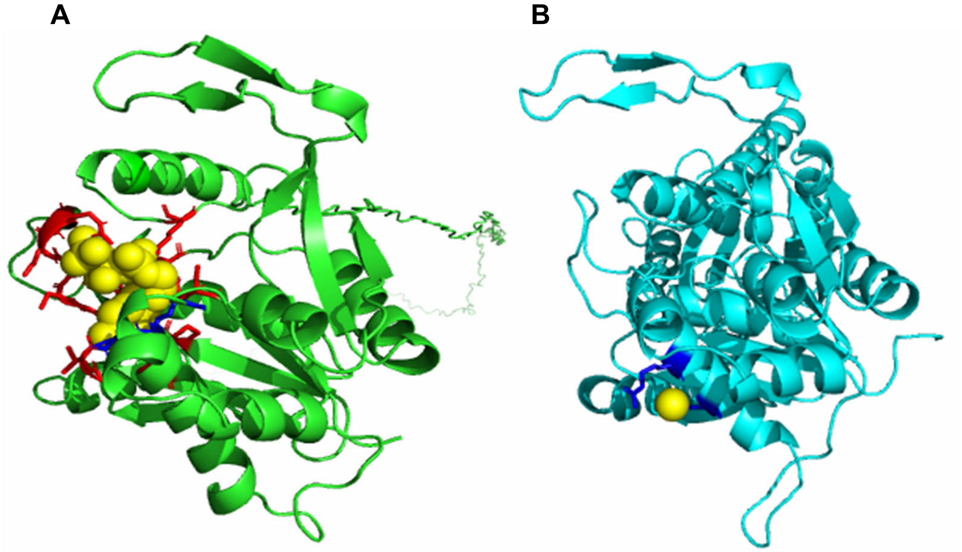


Figure 6.4. Comparison of FunFOLD3 and observedligand-binding site predictions for T0783 (PDB ID 4cvh)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the cytidine-5'-triphosphate (CTP) ligand is shown as sphere and coloured yellow. The best BDT and MCC scores were achieved using 22.5Å **(B)** The observed ligand binding site residues for T0783 (PDB ID 4cvh) shown as sticks and coloured blue, the CL ligand is shown as sphere and coloured yellow

The second CASP11 target was T0798 and the predicted ligand-binding residues by FunFOLD3 were 13,**14,15,16,17,18,19,29,30,31**,33,35,36,61,62,**117,118,120,121,147,148,149** (correct residues in red) and the observed ligand-binding residues were 12,14,15,16,17,18,19,29,30,31,32,34,117,118,120,121,147,148,149. The MCC and BDT scores were 0.753 and 0.797, respectively.

For T0798, FunFOLD3 obtained MCC and BDT scores of 0.753 and 0.797, respectively. There were six over-predictions and three under-predictions. In comparison, the best docked model had an MCC and BDT score of 0.74 and 0.65, respectively and this was achieved for complex 2 using a 10% grid box calculation (refer to Table S.33 in Appendix 4). Although, this was not an improvement in the overall scores, there was an improvement in the number of over-predictions which had reduced to one (SER 35) from the six in FunFOLD3. Additionally, the number of correct predictions had reduced from the 16 predictions by FunFOLD3 to 12 with FunFOLD3-D. This trend was seen across all the different grid box calculations; a reduction in the number of incorrect over-predictions, but also a reduction in the number of correct predictions. This could have been caused by a rotation of the ligand within the ligand-binding space, which is illustrated in Figures S.75C-F. As can be seen in Figure S.76C-F the ligand has rotated from a more "ideal" position seen with FunFOLD3, to a position which has reduced the number of both correct and incorrect predictions. The results from docking with this target highlights the importance of balance when it comes to ligand-binding residue prediction; scores are not necessarily proportionally improved by reducing the number of incorrect residue predictions and decreasing the number of correct residue predictions.

Furthermore, this target also addresses whether FunFOLD3 predictions with good MCC and BDT scores can be improved further, i.e. is there a cut-off when it comes to refinement with docking and further improvements can’t be made?

The third CASP11 target was T0807 and the predicted ligand-binding residues by FunFOLD3 were **20,21,22**,23,**50**,54,**55**,113,**143**,**165,193,194,195,196,**197,**198,199**,200, **201**,207,**224,240,241,242,244,248,251** (correct residues in red) and the observed ligand-binding residues were 20,21,22,50,55,80,142,143,165,193,194,195,196,198,199,201,224, 240,241,242,243,244,245,248,251,252. The MCC and BDT scores were 0.771 and 0.849, respectively.

T0807 was similar to T0798 in terms of good FunFOLD3 prediction, which was unable to be improved by docking. FunFOLD3 obtained scores of 0.771 and 0.849 for MCC and BDT, respectively. However, the best prediction by FunFOLD3-D was complex 5 for 10% box calculation and yielded an MCC and BDT of 0.49 and 0.43, respectively (Figure S.77D). This was driven by the number of correct ligand-binding residues which was greater than for the other complexes of different grid box sizes totalling seven correct ligand-binding residues. As with T0798, the number of incorrect ligand-binding residues reduced from six with FunFOLD3 to five with FunFOLD3-D and two further correct ligand-binding residues were predicted (SER245 and ASN252).

A similar trend was seen with this target as with T0798 where the ligand had rotated out of the ligand-binding space and hence reduced the quality of the prediction. This target adds further weight to the observation of good MCC and BDT scores cannot be improved further with docking, but in fact might be more detrimental to the prediction.

The fourth docked CASP11 target was T0813 and the predicted ligand-binding residues by FunFOLD3 for NAD were 11,12,**13**,14,15,16,37,38,72,73,74,75,98,100,123,127,128,131, 132,235 (correct residues in red) and the observed ligand-binding residues were 13,42,46, 133 for the Mg ligand. The MCC and BDT scores were 0.086 and 0.19, respectively.

T0813 was similar to CASP11 target T0783 where the predicted ligands did not match the observed ligand and a clear size difference between the ligands with NAD having a molecular weight of 663.42 and the observed Mg ligand having a weight of 24.30. As with T0783, despite the size in molecular weight of the ligands, there was an improvement in the MCC and BDT scores following docking. The greatest improvement was seen with 50% grid box calculation where the MCC and BDT score was 0.46 and 0.43, respectively (Figure S.80F). In comparison to FunFOLD3 where the MCC and BDT scores were 0.086 and 0.19, respectively. This was an increase of 435% and 126%, for MCC and BDT, respectively. The increase in ligand-binding site residue prediction was driven by an increase in the number of correct predictions from one with FunFOLD3 (ILE13) to three with FunFOLD3-D (ILE13, THR42 and PRO133). Given the size difference between the predicted ligand and observed ligand, further improvement to the ligand-binding site residues is unlikely.

The predicted ligand-binding residues for T0813 for the NAI ligand by FunFOLD3 were 11,15,16,38,72,73,74,75,77,98,100,123,127,128,131,132,235 and the observed ligand-binding residues were 13,42,46,133 for the Mg ligand. The MCC and BDT scores were 0.029 and 0.11, respectively.

As with the predicted NAD ligand seen in the previous example, AutoDock Vina was able to improve the MCC and BDT scores from -0.029 and 0.11 for MCC and BDT scores, respectively to 0.34 and 0.28, respectively for complex 2 using 22.5A (Figure S.78C). This gave a score increase of 0.23 and 0.17 (1307% and 155%), respectively which is much pronounced compared with the NAD ligand, despite the MCC and BDT scores being lower, however the FunFOLD3 scores were low initially. As seen with the NAD ligand, the improvement in scores is driven by inclusion of correctly observed ligand-binding residues, of which there were none with FunFOLD3, as shown by the poor MCC score, but there were closely aligned residues as shown by the better BDT scores. For complex 2 using 22.5A box calculation, three (ILE13, THR42, PRO133) of the four correct predictions were included in the ligand binding residues, thus greatly increasing the MCC and BDT score.

The predicted ligand-binding residues by FunFOLD3 for NAP ligand for CASP11 target T0813 were 11,12,14,15,16,36,37 38,39,**42**, 55,72,73,74,75,98,100,127,128,131,132,235 (correct residues in red and bold) and the observed ligand-binding residues were 13,42,46,133 for the Mg ligand. The MCC and BDT scores were 0.079 and 0.2, respectively

The predicted NAP ligand for T0813 follows the same pattern as with NAD and NAI ligands. The MCC and BDT scores as obtained with FunFOLD3 were 0.079 and 0.19, respectively. The best MCC and BDT score was for complex 1, pose 1 from 10% box calculation (Figure S.80D) and was 0.34 and 0.27, respectively. This was a score increase of 0.261 and 0.27 (330% and 35%), respectively. As with NAI, this improvement was driven by the prediction of three of the four correct ligand-binding site residues (ILE13, THR42, PRO133). In the FunFOLD3 prediction, only one correct ligand-binding residue had been predicted (THR42). The more modest increase in BDT score between FunFOLD3 and FunFOLD3-D was due to less of a spread of the ligand-binding site residues, in particular incorrect residues with FunFOLD3-D predictions more aligned with the correct ligand-binding site residues.

The fifth docked CASP11 target was T0819 and the predicted ligand-binding residues by FunFOLD3 were **93,94,95,119,167,194,197,223,225,226,234**,347 (correct residues in red) and the observed ligand-binding residues were 93,94,95,119,161,167,194,196,197,223, 225,226,234. The MCC and BDT scores were 0.877 and 0.853, respectively

T0819 was similar to CASP11 targets T0798 and T0807, in relation to all of these had good MCC and BDT scores. Additionally, FunFOLD3-D was unable to improve the MCC and BDT scores with docking. Across all the different grid box calculations, there was a decrease in the MCC and BDT scores. The lowest decrease was seen with a 10% grid box calculation (complex 8 Figure S.81D), which is unsurprising as this grid box is closest to the original ligand-binding space. An MCC and BDT score of 0.87 and 0.80 was achieved, respectively. This gave a modest decrease of 1% for the MCC score and 6% for the BDT score which was a score decrease of -0.007 and -0.053 for MCC and BDT, respectively.

FunFOLD3 had one incorrect over-prediction (ARG347) and missed off two correct predictions (TYR161 and ALA196). FunFOLD3-D, in comparison had no incorrect predictions, however the correct predictions had reduced from 11 to 10 (GLY93 was missed from FunFOLD3-D prediction), and thus decreasing the MCC score. This finding aligns with T0798 regarding improvement of MCC and BDT scores in relation to the number of correct and incorrect predictions.

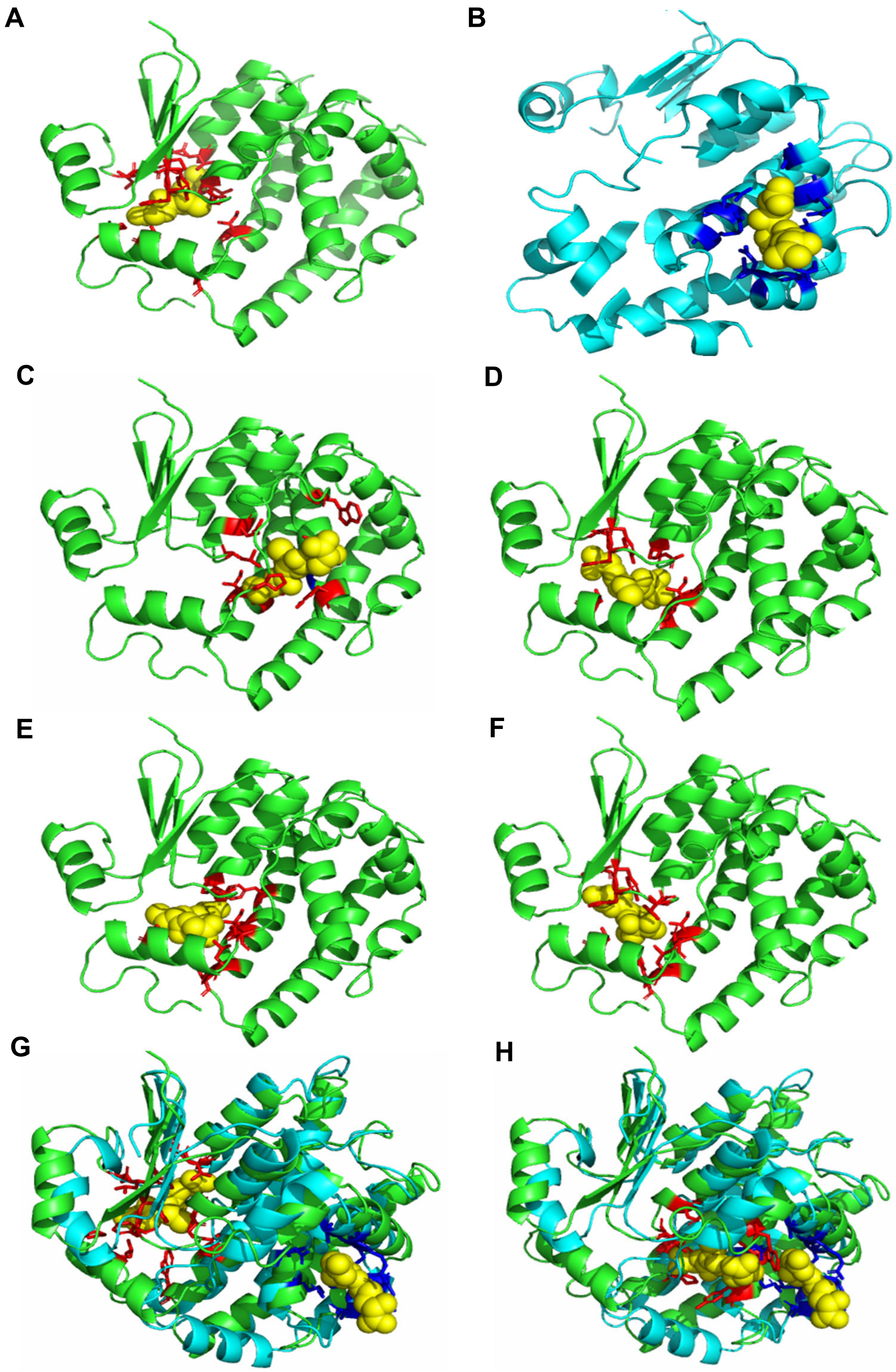


Figure 6.5. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0849 (PDB ID 4w66)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the GSH ligand is shown as sphere and coloured yellow. BDT score of 0.0375 and MCC score of -0.05 **(B)** The observed ligand binding site residues for T0849 (PDB ID 4w66) shown as sticks and coloured blue, the GSH ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct prediction is shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D with 22.5Å the predicted structure coloured green and the observed structure coloured cyan. BDT and MCC score of 0.24 and 0.0086, respectively

The predicted ligand-binding residues for T0849 by FunFOLD3 were 9,10,14,15,54,55,56, 67,68,108,113,226,230 and the observed ligand-binding residues were 168,171,179,182, 183,190,194,197. The MCC and BDT scores were -0.05 and 0.0375, respectively. The results from the different grid box calculations are shown below in Table 6.8-6.11.

Table 6.8. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0849 (PDB ID 4w66)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**13,17,107,108,168,169,172,206,217**	**0.086**	**0.136**	**0.24**	**0.2025**
2	9,10,12,14,15,54,55,108,112,113,218,226	-0.043	0.007	0.028	-0.0095
3	9,10,12,15,55,56,112,113,222,226	-0.039	0.011	0.023	-0.0145
4	9,10,12,15,108,112,113,222,226	-0.037	0.013	0.027	-0.0105
5	13,17,21,104,107,108,111,165,168,169, 172,173,217	0.058	0.108	0.21	0.1725
6	13,14,17,104,107,108,111,112,165,168, 169,172,173,215,217,218	0.043	0.093	0.18	0.1425
7	68,69,72,94,97,98,101,162	-0.034	0.016	0.035	-0.0025
8	9,10,15,55,56,108,112,113,226	-0.037	0.013	0.025	-0.0125
9	10,12,14,55,112,113,222,226	-0.034	0.016	0.025	0.2025

Table 6.9. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0849 (PDB ID 4w66)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	9,10,15,54 55,105,108,109,113,226	-0.039	0.011	0.026	-0.0115
2	9,10,14,15 55,56,67,68,113,226	-0.039	0.011	0.022	-0.0155
3	9,10,15,54,55,56,105,108,109	-0.037	0.013	0.027	-0.0105
4	9,14,15,55,56,68,108,226	-0.034	0.016	0.027	-0.0105
5	9,10,15,54,55,108,109,113,226,230	-0.039	0.011	0.023	-0.0145
6	9,10,14,15,54 55,56,67,68,108,226	-0.040	0.01	0.024	-0.0135
7	9,10,15,52,55,67,68,226	-0.034	0.016	0.019	-0.0185
8	9,10,15,54,55,56,67,68,230	-0.037	0.013	0.019	-0.0185
9	**9,10,15,55,105,108,109,226**	**-0.034**	**0.016**	**0.029**	**-0.0085**

Table 6.10. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0849 (PDB ID 4w66)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	9,10,14,15,55,108,112,113,226	-0.037	0.013	0.027	-0.0105
2	9,10,14,54,55,108,112,113,218,226,230	-0.041	0.009	0.025	-0.0125
3	9,10,15,54,55,105,108,109,113,230	-0.039	0.011	0.026	-0.0115
4	9,14,15,55,56,67,68,108,113,218	-0.039	0.011	0.028	-0.0095
5	9,10,15,55,56,67,68,108,113,226	-0.039	0.011	0.023	-0.0145
6	9,10,14,54,55,108,113,226	-0.034	0.016	0.024	-0.0135
7	9,10,14,15,54,55,108,226,230	-0.037	0.013	0.024	-0.0135
8	9,10,14,55,108,112,113,226,230	-0.037	0.013	0.024	-0.0135
9	**14,15,101,105,108,112,113,218,226**	**-0.037**	**0.013**	**0.036**	**-0.0015**

Table 6.11. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0849 (PDB ID 4w66)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	9,10,14,15,55,108,112,113,226	-0.037	0.013	0.027	-0.0105
2	9,10,12,14,15,55,108,112,113,226	-0.039	0.011	0.028	-0.0095
3	9,10,12,14,15,55,108,112,218,226,230	-0.041	0.009	0.028	-0.0095
4	9,10,14,54,55,108,112,113,226	-0.037	0.013	0.025	-0.0125
5	**9,10,12,14,55,108,112,113,218,226**	-0.039	0.011	**0.029**	**-0.0085**
6	9,10,14,55,112,113 226,229,230	-0.037	0.013	0.020	-0.0175
7	9,10,12,14,15,41,53,54,108,112,218,226, 229,230	-0.046	0.004	0.024	-0.0135
8	**9,10,12,15,112,226,229,230**	**-0.034**	0.016	0.022	-0.0155
9	9,10,12,14,15,55,56,67,68,112,226	-0.041	0.009	0.024	-0.0135

For T0849, FunFOLD3 achieved a BDT and MCC score of 0.0375 and -0.05, respectively. As can be seen from the results in the tables, model one from 22.5Å provided the best MCC and BDT scores across all the four different grid box calculations. The MCC and BDT scores increased to 0.086 and 0.24, respectively aided by one correct ligand residue prediction (GLU 168). This gave a score increase of 0.136 and 0.2025, for MCC and BDT, respectively.

The increase in the MCC and BDT was driven by the rotation of the ligand from a distal place within the protein in relation to the observed ligand to a more proximal location (see Figure 6.5). Unsurprisingly, the grid box calculations which were close to the predicted ligand-binding space for FunFOLD3, did not provide the best MCC and BDT scores due to the ligand needing to rotate to a location further away and therefore a wider ligand-binding space was required.

The first CASP12 target to be docked was T0912 and the predicted ligand-binding site residues for the FRU ligand were 207,208,209,490 and the observed ligand-binding site residues were 155,262,264,268 for the calcium ligand. The MCC and BDT scores were - 0.00672 and 0.0295, respectively. The predicted ligand-binding site residues for the MAV ligand were 334,340,423,426,468,469,470 and the MCC and BDT scores were -0.00892 and 0.0213, respectively.

FunFOLD3 predicted two biologically relevant ligands and neither of the ligands matched the observed calcium ligand. FunFOLD3 obtained MCC and BDT scores of -0.00672 and 0.030, respectively. The BDT which was the most improved was for complex 3, pose 2 which achieved a BDT score of 0.076 which was a 158% increase for the 22.5Å box calculation. This increase was most likely driven by one ligand-binding residue (ARG153) which was close in proximity to ARG155 from the observed structure. It is likely, that a larger ligand-binding rotational space could have improved the score further.

The second ligand to be docked was MAV and as with the FRU ligand, there was no real definite improvement on the ligand-binding site residues and ultimately the MCC and BDT scores. As with the FRU ligand, no figures were made due to the poor MCC and BDT scores.

The second CASP12 target to be docked was T0913 and the predicted ligand-binding site residues by FunFOLD3 were 100,149,153,156,171,206,266,267,318,358,359,364 and the observed ligand-binding site residues were 64,65,66,67,209,210,273,274,320,321,363,368, 371. The MCC and BDT scores were -0.0367 and 0.091, respectively.

T0913 was an interesting target for docking as the observed structure as released by the CASP12 organisers did not have a ligand present in the observed structure, however, ligand-binding site residues were released in a publication pertaining to CASP12 (Liu *et al.*, 2018). As previous docking experiments had resulted in some improvement of MCC and BDT scores if there were differences between predicted and observed ligands, it was reasonable to consider this target for docking. However, there was no improvement in the MCC and BDT scores and hence no figures were created. BDT scores did increased for 10% and 20% grid box calculation. With complex 1, pose 2 20% grid box calculation providing the best increase in BDT scores. This could have been due to ligand-binding site residues predicted by FunFOLD3-D being closer to the observed ligand-binding residues. For example, predicted residues ALA315 and ASN318 are close to the observed residue LEU320, predicted residues GLY360 and GLY361 is close to observed residue LYS363.

The third CASP12 target to be docked was T0916 and the predicted ligand-binding site residues by FunFOLD3 were **14**,**15**,16,17,18,51,59,60,62,**63**,64,68,107 (correct predictions in red and bold) and the observed ligand-binding residues were 62,63,65,66,153,154,155, 340,344 for the GLC(1) ligand and 12,14,15,63,111,153,155,156,230 for the GLC(2) ligand. The MCC and BDT scores were 0.162 and 0.263 when compared against GLC(1), respectively and 0.263 and 0.37, when compared against GLC(2), respectively

T0916 involved docking a proportion of the predicted structure, due to CASP organisers only releasing part of the structure for prediction and was the only target docked in this way. The predicted ligand by FunFOLD3 was NAD for CASP12 and the observed ligands GLC were predicted in two different areas of the protein structure and hence denoted as GLC(1) and GLC(2). BDT and MCC scores were calculated for both ligands, in case there was a preference following docking for one ligand over the other. FunFOLD3 achieved a BDT and MCC score of 0.370 and 0.263, respectively for GLC(2) ligand as there were more correct predictions for this ligand, in comparison to GLC(1), were there was no correct ligand-binding site residues.

As can be seen from Table S.71 model seven from 50% grid box calculation provided the best MCC and BDT scores across all the four different grid box calculations. However, the MCC and BDT scores decreased to 0.16 and 0.32, respectively and was a 39% and 14% reduction. The decrease was due to a decline in the number of correct prediction by FunFOLD3 which was three (VAL14, LEU15 and LYS63) to only two correct predictions (LEU15 and LYS63) by FunFOLD3-D. Additionally, LYS63 seems to be a ligand-binding residue which is conserved across all the higher scoring models.

Figure S.82G is the alignment of the FunFOLD3 predicted ligand-binding site with the observed structure and Figure S.82H is the alignment of the top scored complex following docking and the observed structure and a clear difference between the location of the predicted ligand binding site by FunFOLD3 and FunFOLD3-D can be seen.

The first CASP13 target was T0953s2 and the predicted ligand-binding site residues by FunFOLD3 were 119,120,121,122,124,125,126,127,155,156,157,158,159,**164**,165,166, 167,168,169,170,174,195,198 and the observed ligand-binding site residues were 54,164. The MCC and BDT scores were 0.11 and 0.12, respectively

FunFOLD3 predicted the DJB ligand, whereas the observed ligand was IMD. An MCC and BDT score of 0.12 and 0.11 was achieved, respectively, mainly due to the prediction of one correct ligand-binding site residue (THR164) by FunFOLD3. There was no increase in MCC scores across all four of the grid box calculations, however two grid box calculations increased the BDT score to 0.21 which was a 91% increase. The increase was for complex 4 from 10% box calculation and complex 8 from 20% grid box calculation both box calculations had the same residue predictions and the increase in BDT score was due to the residues closely aligning with the correct observed residue THR164. The closely aligned residues were GLU165, ALA166,GLY167. Hence the increase in BDT scores, as the correctly predicted residue from FunFOLD3 was not predicted this lead to the increase in the MCC scores.

It is worth noting with this structure (see Figure 3.23 in Chapter 3), that due to the overlap in the flexible loops in the predicted structure, there was unlikely to be an improvement due to the predicted and observed structures not aligning.

The second CASP13 target to be docked was T0954 and the predicted ligand-binding site residues by FunFOLD3 for T0954 were 77,119,273,274,275 for the LYS ligand and the observed ligand-binding site residues were 123,124,129,130,131. The MCC and BDT scores were -0.015 and 0.028, respectively

FunFOLD3 predicted LYS and DNA as ligands for T0954 and the observed ligand was Mg. The LYS ligand was docked using AutoDock Vina, it was felt the DNA ligand was not suitable for docking due to the double-helix structure of DNA and that DNA was not a correctly predicted ligand.. Following docking, there was no improvement in either the MCC or BDT scores. For all the four grid box calculations, the MCC and BDT scores either stayed the same or decreased. As a result of no improvement in ligand-binding, no figures were created.

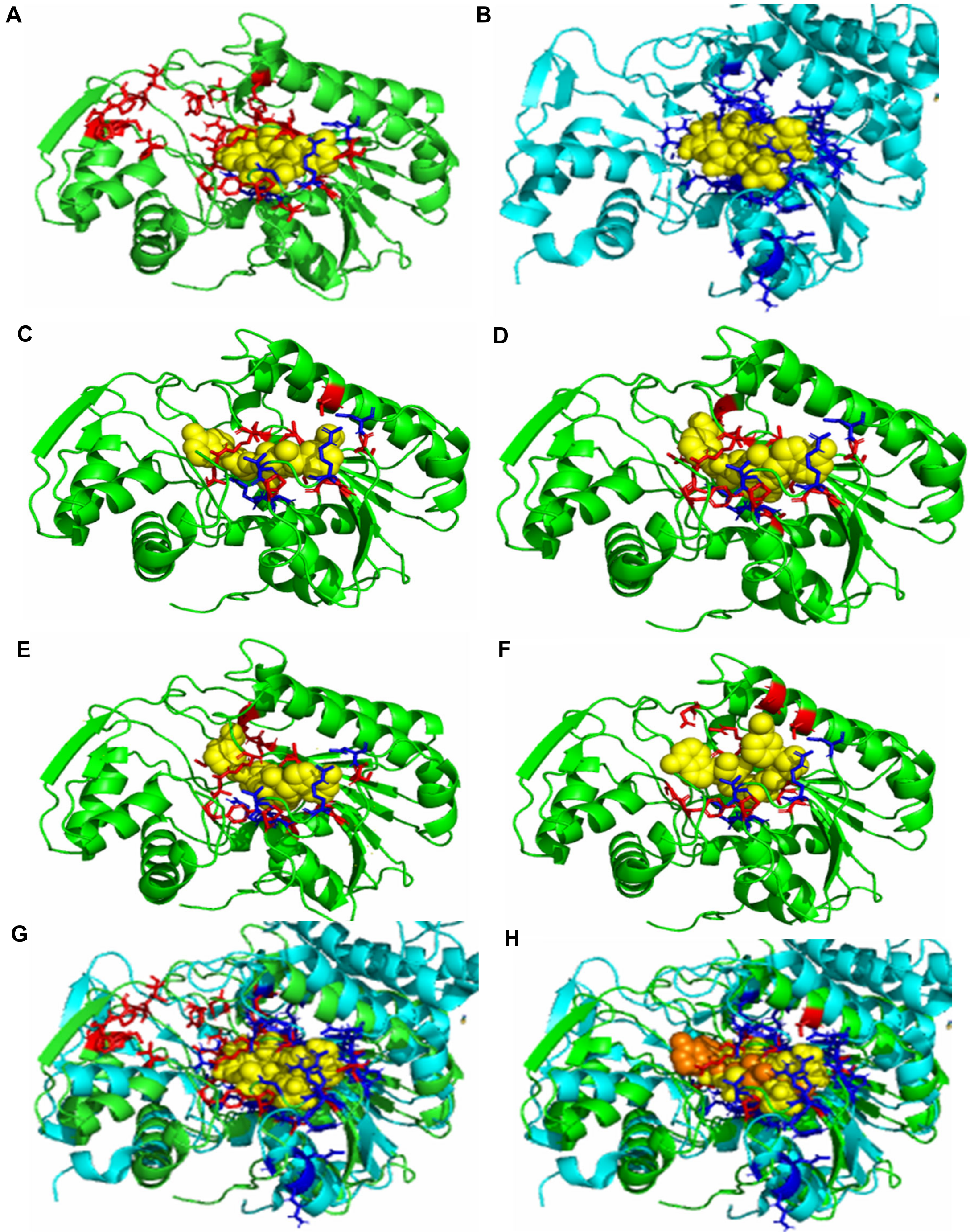


Figure 6.6. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0965 (PDB ID 6d2v)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAD ligand is shown as sphere and coloured yellow. BDT score of 0.35 and MCC score of 0.12 **(B)** The observed ligand binding site residues for T0965 (PDB ID 6d2v) shown as sticks and coloured blue, the NDP ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct prediction is shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D (22.5Å box calculation) with the predicted structure coloured green and the observed structure coloured cyan. NAD coloured orange and NDP coloured yellow. BDT and MCC score of 0.33 and 0.22, respectively

The predicted ligand-binding residues for T0965 by FunFOLD3 were 30,32,**33,34,35**,53,**54,** **56**,58,75,76,**77**,97,98,99,101,103,134,135,136,165,192,193,194,195,201,204,219,220,221, 226,264,286,291 (correct residues are in red and bold). The observed ligand-binding site residues were 10,12,13,14,15,33,34,35,54,55,56,57,77,78,79,81,114,115,116,145,149,172, 173,174,175 for the observed NDP ligand. The MCC and BDT score was 0.12 and 0.35, respectively. The results from the different grid box calculations are shown below in Tables 6.12-6.15.

Table 6.12. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0965 (PDB ID 6d2v)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	30,33,34,35,36,57,58,97,99,100,101,103,104, 195,201,202,203,204,205	0.12	0	0.22	-0.13
2	**30,33,34,35,53,54,57,58,76,77,97,98,99,101,113,136,192,194,195**	**0.22**	**0.1**	**0.33**	**-0.02**
3	33,34,35,36,57,58,97,99,100,101,102,103,104,201,203,204	0.15	0.03	0.22	-0.13
4	30,33,34,57,58,97,100,101,201,202,203,204,207,219,284	0.099	-0.021	0.16	-0.19
5	33,34,35,57,58,97,100,101,103,104,201,202,203,285,286,333	0.16	0.04	0.20	-0.15
6	101,103,104,136,137,138,165,193,194,202,203,204,207,288,333,334	-0.064	-0.184	0.029	-0.321
7	101,103,104,136,137,138,193,194,195,203,204,285,286,288,291	-0.064	-0.184	0.028	-0.322
8	101,103,194,195,202,203,204,206,207,210,327,328,333	-0.054	-0.174	0.024	-0.326
9	101,102,103,104,136,193,194,195,203,204,284,286	-0.057	-0.177	0.017	-0.333

Table 6.13. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0965 (PDB ID 6d2v)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding space

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**30,33,34,35,36,53,54,57,58,76,77,97,99,101,136, 137,138,192,193,194,195,201**	**0.20**	**0.08**	**0.33**	**-0.02**
2	30,33,34,35,36,57,58,97,98,99,100,101,136,137, 138,192,193,194,195,201	0.12	0	0.23	0.12
3	34,35,54,57,97,99,100,101,113,134,136,137,138, 192,193,194,195,201	0.13	0.01	0.21	-0.14
4	30,33,34,53,54,57,58,76,77,98,99,100,101,113,201, 202,333	0.20	0.08	0.28	-0.07
5	30,33,34,35,53,54,57,58,76,98,99,100,101,113,134, 136,165,192,194,195,201	0.16	0.04	0.29	-0.06
6	30,33,34,35,57,58,97,98,99,100,101,134,136,165,	0.15	0.03	0.21	-0.14

	195,201				
7	30,33,34,35,53,54,57,58,76,98,99,100,101,113,134,136,192,193,195,201	0.16	0.04	0.28	-0.07
8	30,34,35,53,54,57,58,76,97,98,99,100,101,136,137,138,165,192,193,194,195,201	0.10	-0.02	0.25	-0.1
9	30,33,34,35,53,54,57,58,76,97,98,99,100,101,113,134,136,165,193,195,201	0.16	0.04	0.29	-0.06

Table 6.14. 1Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0965 (PDB ID 6d2v)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding space

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	30,33,34,35,36,57,58,97,98,99,100,101,136,137,138,192,193,194,195,201	0.12	0.000	0.23	-0.12
2	**30,33,34,35,36,53,54,57,58,76,77,97,99,101,136,137,138,192,193,194,195,201**	**0.20**	**0.080**	**0.33**	**-0.02**
3	30,33,34,35,53,54,57,58,97,99,100,101,136,137,138,192,193,194,195,201	0.16	0.040	0.27	-0.08
4	30,33,34,35,36,54,58,97,98,99,100,101,113,136,137,138,192,193,194,195	0.12	0.0	0.25	-0.1
5	33,34,35,57,97,99,100,101,113,136,137,138,192,193,194,195,201	0.14	0.020	0.21	-0.14
6	30,32,33,34,53,54,57,58,75,76,77,97,98,99,100,101,113,201,202,333	0.17	0.050	0.31	-0.04
7	30,33,34,53,54,57,58,76,77,98,99,100,101,113,201,202,333	0.20	0.080	0.28	-0.07
8	33,34,35,54,57,58,97,98,99,100,101,103,136,192,193,194,195,201	0.18	0.060	0.25	-0.1
9	30,33,34,35,53,54,57,58,76,97,98,99,100,101,113,134,136,165,193,195,201	0.16	0.040	0.29	-0.06

Table 6.15. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0965 (PDB ID 6d2v)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding space

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	30,33,34,35,36,53,54,57,58,76,77,97,99,101,136,137,138,192,193,194,195,201	0.20	0.08	0.33	-0.02
2	30,33,34,35,36,57,58,97,98,99,100,101,136,137,138,192,193,194,195,201	0.12	0	0.23	-0.12
3	30,32,34,35,53,54,57,58,97,99,101,103,136,165,192,194,195,201,203,205	0.20	0.08	0.34	-0.01
4	30,32,34,35,53,54,57,58,97,99,101,103,136,165,192,194,195,201,203,205	0.12	0	0.24	-0.11
5	**30,33,34,35,54,57,58,77,97,98,99,100,101,104,109,110,113,201,202**	**0.22**	**0.1**	**0.30**	**-0.05**
6	30,33,34,35,57,58,97,98,99,100,101,136,192,193,194,195,201,203	0.13	0.01	0.21	-0.14
7	33,34,35,36,57,97,99,100,101,136,137,138,192,193,194,195,201	0.14	0.02	0.21	-0.14
8	30,33,34,35,54,57,58,97,98,99,100,101,113,136,137,138,165,192,193,194,195	0.16	0.04	0.27	-0.08
9	30,33,34,35,53,54,57,58,97,98,99,100,101,136,192,193,194,195,201	0.17	0.05	0.26	-0.09

CASP13 target T0965 predicted NAD as the biologically relevant ligand, however the observed ligands were NDP and Cl. Due to the closeness between NAD and NDP, and the cross over in correct ligand predictions, this target was selected for docking.

The results for this target were somewhat interesting, as across all four grid box calculations the BDT score did not increase but decreased, with some grid box calculations having modest decreases of 6%. However, the MDT scores increased in each one of the grid box calculations. Complex 2 using 22.5Å was provided the best MDT and BDT scores, although the BDT score decrease, the decrease was smaller compared to the decrease with other box calculations.

The increase in MDT score was likely driven by a decrease in the number of incorrect ligand-binding residue predictions. FunFOLD3 had predicted a total of 28 incorrect predictions, whereas the top-scoring model from AutoDock Vina had 13 incorrect predictions. As can be seen in Figure 6.6.

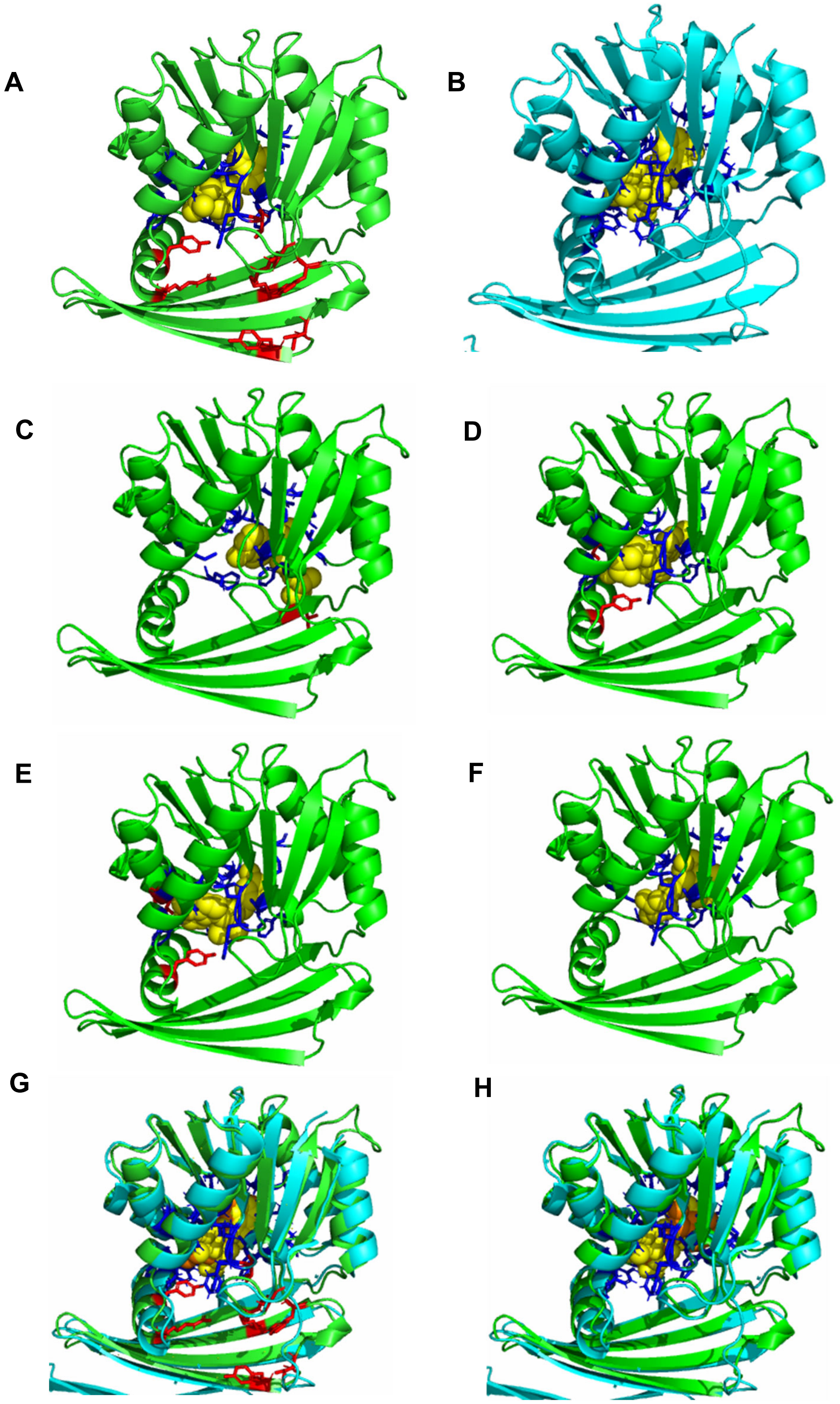


Figure 6.7. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0983 (PDB ID 6uk5)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the SAH ligand is shown as sphere and coloured yellow. BDT and MCC score of 0.715 **(B)** The observed ligand binding site residues for T0983 (PDB ID 6uk5) shown as sticks and coloured blue, the SAM ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct prediction is shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the predicted structure coloured green and the observed structure coloured cyan. Predicted ligand SAH is coloured orange for ease of identification **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D (50% box calculation) with the protein coloured green and the observed structure coloured cyan. SAH coloured orange and SAM coloured yellow. BDT and MCC score of 0.71 and 0.83,respectively

The predicted ligand-binding residues for T0983 by FunFOLD3 were **2**,14,17,**21,46,47,48, 52,67,68,69,72,88,89,90,105,106**,107,**108,111**,141,147,150,178,228 (correct residues are in red and bold). The observed ligand-binding site residues were 2,10,21,46, 47,48,52,66,67, 68,69,72,88,89,90,91,105,106,108,111,112 for the observed SAM ligand. The MCC and BDT score was 0.715 and 0.715, respectively.

The incorrect ligand-binding site residues are surrounding biologically irrelevant ligands, as determined by FunFOLD3, phenyl-uridine-5'-diphosphate, thymidine diphosphate phenol and thymidine. Whilst FunFOLD3 correctly identified the ligands and not being biologically relevant, the residues surrounding these ligands were incorrectly included in the prediction as shown in Figure 6.7A. Docking with AutoDock Vina correctly identified these residues not being part of the ligand-binding residue cluster (Figure 6.7C-F). The results from the grid box calculations are shown below in Tables 6.16-6.19.

Table 6.16. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0983 (PDB ID 6uk5)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**2,46,67,68,72,89,90,91,108,111,112,198**	**0.67**	**-0.045**	**0.53**	**-0.185**
2	1,2,14,46,67,68,69,72,90,105,106,108,111	0.64	-0.075	0.55	-0.165
3	1,2,46,67,68,88,90,111,197,198	0.45	-0.265	0.37	-0.345
4	17,106,107,110,111,140,150,152,165,167,169,178, 199,228	0.049	-0.666	0.17	-0.545
5	17,19,106,150,153,165,167,178,228	0.016	-0.699	0.077	-0.638
6	17,141,150,165,169,178,228	-0.054	-0.769	0.021	-0.694
7	2,67,68,90,111,197,198	0.38	-0.335	0.25	-0.465
8	17,19,24,140,150,178,226,228	-0.058	-0.773	0.044	-0.671
9	1,67,68,89,90,111,196,197,198	0.33	-0.385	0.27	-0.445

Table 6.17. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0983 (PDB ID 6uk5)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	2,14,21,46,47,48,49,50,51,52,67,68,72,88,90,108,111	0.66	-0.055	0.68	-0.035
2	2,14,21,46,47,48,49,50,51,52,67,68,72,106,108,111	0.63	-0.085	0.63	-0.085
3	2,14,21,46,47,48,49,51,52,67,68,105,106,108,111	0.65	-0.065	0.62	-0.095
4	1,2,14,21,46,47,48,49,50,51,52,67,68,72,106,108,111	0.60	-0.115	0.65	-0.065
5	2,45,46,67,68,72,88,89,90,105,106,108,111	0.71	-0.005	0.59	-0.125
6	1,2,14,19,21,67,68,69,72,90,106,108,111	0.58	-0.135	0.51	-0.205
7	**2,14,21,46,47,48,49,52,67,72,90,105,106,108,111**	**0.71**	**-0.005**	**0.64**	**-0.075**
8	2,14,21,46,52,67,68,72,90,106,108,111	0.67	-0.045	0.53	-0.185

Table 6.18. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0983 (PDB ID 6uk5)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	2,14,21,46,47,48,49,50,51,52,67,68,72,88,90,108,111	0.66	-0.055	0.68	-0.035
2	2,14,21,46,47,48,49,50,51,52,67,72,108,111	0.55	-0.165	0.53	-0.185
3	2,14,21,46,48,49,50,51,52,67,68,72,89,90,108,111,112	0.66	-0.055	0.68	-0.035
4	1,2,14,21,46,48,49,50,51,52,67,68,69,72,108,111	0.57	-0.145	0.60	-0.115
5	2,14,21,46,47,49,50,51,52,67,72,90,108,111	0.55	-0.165	0.53	-0.185
6	**2,14,21,46,47,48,49,51,52,67,68,72,90,105,106,108,111**	**0.72**	**0.005**	**0.71**	**-0.005**
7	1,2,14,46,47,48,49,67,68,69,72,90,108,111	0.62	-0.095	0.57	-0.145
8	2,46,67,68,90,106,108,111,112	0.64	-0.075	0.43	-0.285

Table 6.19. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0983 (PDB ID 6uk5)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	2,14,21,46,47,48,49,50,51,52,67,68,72,88,90,108,111	0.66	-0.055	0.68	-0.035
2	2,14,21,46,47,48,49,51,52,67,68,72,88,90,105,106,108,111	0.75	0.035	0.78	0.065
3	2,14,21,46,48,49,50,51,52,67,68,72,89,90,108,111	0.63	-0.085	0.63	-0.085
4	2,14,21,46,47,48,49,50,51,52,67,72,106,108,111	0.59	-0.125	0.58	-0.135
5	2,14,21,47,48,49,50,51,52,67,68,72,108,111	0.55	-0.165	0.53	-0.185
6	2,14,21,46,47,49,50,51,52,67,68,69,72,108,111	0.59	-0.125	0.58	-0.135
7	**2,46,48,67,68,72,88,89,90,91,105,106,108,111,112**	**0.83**	**0.115**	**0.71**	**-0.005**
8	1,2,14,48,49,67,68,69,72,90,106,108,111	0.58	-0.135	0.52	-0.195
9	2,14,46,47,48,49,52,67,72,90,105,106,108,111	0.68	-0.035	0.60	-0.115

T0983 was similar to T0965 in respect a different but related ligand being predicted SAH by FunFOLD3 and the observed ligand was in fact SAM. Additionally, a similar trend was seen whereby the MCC score increased however the BDT stayed the same/decreased slightly. FunFOLD3 achieved a MCC and BDT score of 0.715 for both scores. However, the top scoring complex; complex 7 from 50% box calculation achieved an MCC and BDT of 0.83 and 0.71, respectively. This was an increase in score of 0.115 for MCC and a decrease of -0.005 for BDT.

The increase in MCC score would be due to no incorrect ligand-binding residues being predicted in the FunFOLD3-D method compared to eight incorrect ligand-binding residues which was predicted by FunFOLD. Furthermore, two correct ligand-binding residues (PHE91 and GLU112) were predicted by FunFOLD3-D which were missed by FunFOLD3. Despite the addition of two correct ligand-binding residues the overall number of correct ligand-binding residues decreased from 17 with FunFOLD3 to 15 with FunFOLD3-D.

The next CASP13 target to be docked was T1003 and the predicted ligand-binding residues for T1003 by FunFOLD3 were 113,144,145,146,149,172,244,246,247,275,278,284,306, 308,474. The observed ligand-binding site residues were 257,258,259,262,285,287,328, 332,357,359,360,388,391,419,420,421. The MCC and BDT score was -0.04 and 0.06, respectively.

The FunFOLD3 prediction for T1003 and the observed ligand matched however, a poor MCC and BDT score was achieved. The MCC and BDT score was -0.04 and 0.06, respectively. Despite a correctly predicted ligand, docking failed to improve the MCC scores towards a more positive score. The best MCC and BDT score was obtained for complex 8 with grid box calculation 22.5Å, the MCC score improved to 0.06 and the BDT score to 0.15, thus giving a score increase of 0.1 and 0.09, respectively. Despite the impressive percentage increases the scores are still weak and therefore, no images have been produced. The FunFOLD3 prediction failed to predict any correct ligand-binding site residues. However, following docking with FunFOLD3-D this was increased to one correct ligand-binding residue (ILE391), this correct residue most likely drove the increase in MCC score. The BDT score increase was most likely increased due to two ligand-binding residues (GLN389 and ALA390) which were close to the observed ligand residues (VAL388 and ILE391).

The fifth CASP13 target to be docked was T1009 and the predicted ligand-binding residues by FunFOLD3 were **257,286,325,393,395**,396,**470,484,487,520,557** (correct ligand-binding residues in red and bold). The observed ligand-binding site residues were 257,285,286, 325, 393,395,470,484,487,520,557 for the XYS ligand. The MCC and BDT score was 0.91 and 0.94, respectively.

FunFOLD3 predicted the GLC ligand in two different locations with the one of the GLC ligands (identified as GLC ligand(2) - see Figure S.58 in Appendix 4. In comparison, the observed ligand had five observed ligands, with some predicted at different locations. The GLC ligand(2) shared 10 correct ligand-binding residues with XYS ligand and missed off one correct residue (285) and included one incorrect prediction (396), this this contributed to an MCC and BDT score of 0.91 and 0.94, respectively. As seen previously, in CASP11 with other top scoring MCC and BDT proteins, further improvement is unlikely. All four grid box calculations decreased the MCC and BDT scores, with 10% grid box, unsurprisingly, decreasing the MCC and BDT scores, the least with a 33% and 48% reduction, respectively. This target provides further support that there is a maximum threshold when it comes to improvement of FunFOLD3 predictions.

The sixth CASP13 target to be docked was T1014 and the predicted ligand-binding residues by FunFOLD3 were 57,58,60,61,85,88,90,98,103,104,105,114,115,116,117,118,119,120, 121,144,156. The observed ligand-binding site residues were 139,142,143,167,172,180, 185,186,187,199,200,201,202,228 for the observed ANP ligand. The MCC and BDT score was -0.05 and 0.05, respectively.

For T1014 FunFOLD3 predicted ADP and the observed ligand was ANP and Mg. The MCC and BDT score were on the poor side of the prediction with -0.05 and 0.05, respectively. Following docking, the MCC and BDT scores, did remain fairly poor. However, grid box calculation of 22.5Å improved the MCC and BDT score by 40% or 60%, depending on the complex but the BDT score decreased by 2%, 20% or 80%, once again depending on the complex. In comparison, complex 1 from 10% box grid calculation increased the BDT score by 160% to 0.13 with no effect on the MCC score. The clearest explanation for the increase in BDT is most likely due to the removal of ASN57, ALA58, LYS60 as ligand-binding residues and retainment of the ALA144 and PHE146 which are close to the GLN142 and GLY143 observed ligand-binding residues. This seemed to be common with other higher scoring complexes at other grid box calculations (e.g. complex 2 10% grid box calculation and complex 2 20% grid box calculation).

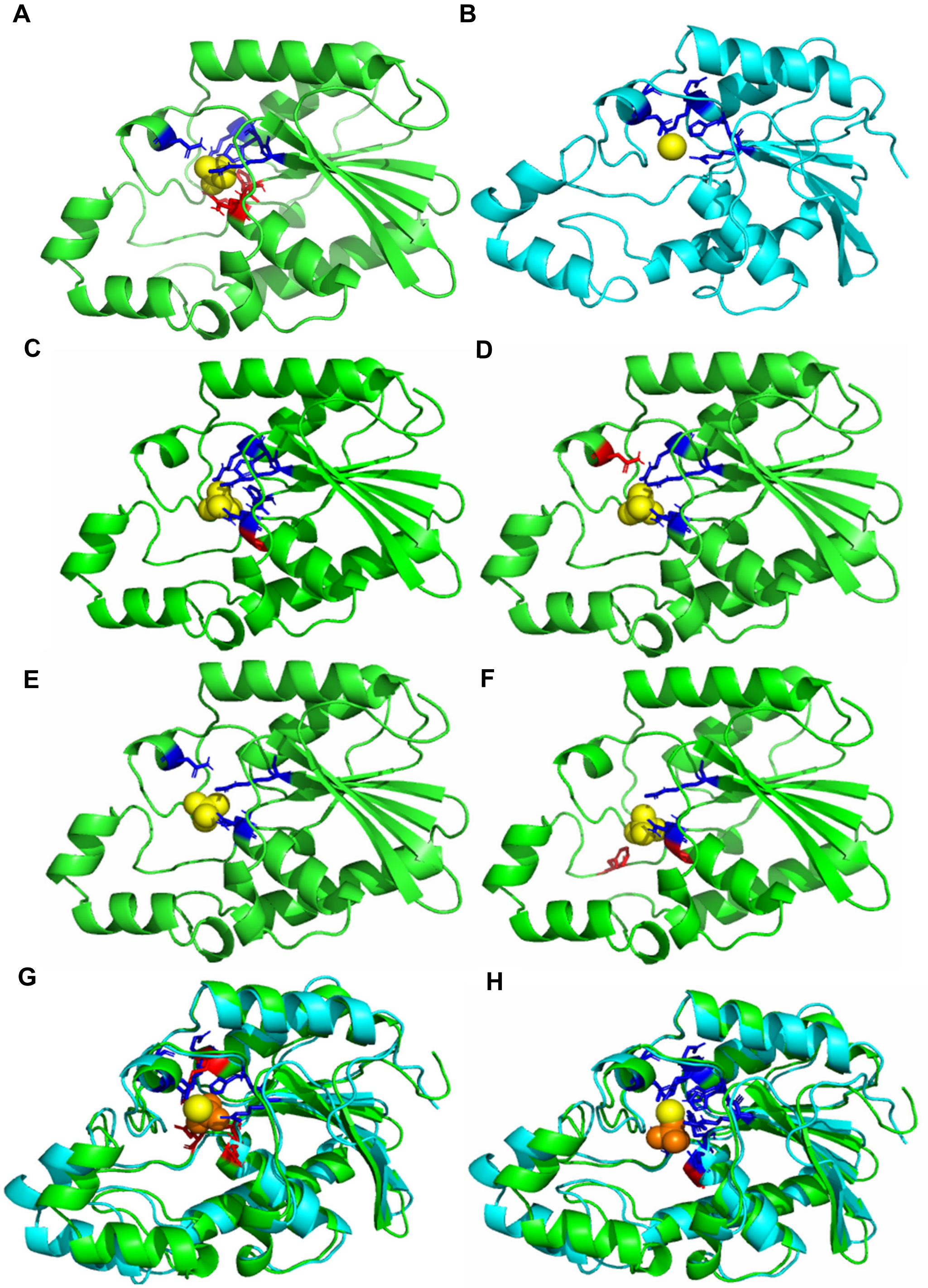


Figure 6.8. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T1016 (PDB ID 6e4b)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the PO4 ligand is shown as sphere and coloured yellow. BDT score of 0.646 and MCC score of 0.556 **(B)** The observed ligand binding site residues for T1016 (PDB ID 6e4b) shown as sticks and coloured blue, the CL ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct prediction is shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D (22.5Å box calculation) with the predicted structure coloured green and the observed structure coloured cyan. PO4 coloured orange and CL coloured yellow. BDT and MCC score of 0.91 and 0.85, respectively for 22.5Å

The predicted ligand-binding residues for T1016 by FunFOLD3 were **7,8,14,**19,20,21,**57,81**, 84,**149,150**. The observed ligand-binding site residues were 7,8,14,57,81,149,150 for the Cl ligand. The MCC and BDT score was 0.556 and 0.646, respectively.

Table 6.20. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T1016 (PDB ID 6e4b)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	7,57,81,149,150,151	0.76	0.204	0.77	0.214
2	**7,8,57,81,149,150,151**	**0.85**	**0.294**	**0.91**	**0.354**
3	Same as complex 2	-	-	-	-
4	Same as complex 2	-	-	-	-
5	Same as complex 2	-	-	-	-
6	Same as complex 2	-	-	-	-
7	Same as complex 2	-	-	-	-

Table 6.21. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T1016 (PDB ID 6e4b)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	7,81,150,151	0.56	0.004	0.48	-0.166
2	7,57,81,150,151	0.67	0.114	0.63	-0.016
3	Same as complex 2	-	-	-	-
4	Same as complex 2	-	-	-	-
5	Same as complex 2	-	-	-	-
6	Same as complex 2	-	-	-	-
7	Same as complex 2	-	-	-	-
8	**7,14,57,81,150**	**0.84**	**0.284**	**0.71**	**0.064**
9	Same as complex 8	-	-	-	-

Table 6.22. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T1016 (PDB ID 6e4b)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	7,81,150	0.64	0.084	0.43	-0.216
2	Same as complex 1	-	-	-	-
3	Same as complex 1	-	-	-	-
4	7,14,81,150,151	0.67	0.114	0.63	-0.016
5	7,81,150,151	0.56	0.004	0.48	-0.166
6	Same as complex 5	-	-	-	-
7	Same as complex 1	-	-	-	-
8	**7,14,81,150**	**0.75**	**0.194**	**0.57**	**0.076**
9	14,19,20,57,81	0.49	-0.066	0.46	-0.186

Table 6.23. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T1016 (PDB ID 6e4b)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**7,81,84,150,151**	**0.49**	**-0.066**	**0.50**	**-0.146**
2	Same as complex 1	-	-	-	-
3	Same as complex 1	-	-	-	-
4	Same as complex 1	-	-	-	-

5	Same as complex 1	-	-	-	-
6	Same as complex 1	-	-	-	-
7	Same as complex 1	-	-	-	-
8	Same as complex 1	-	-	-	-
9	Same as complex 1	-	-	-	-

The best CASP13 target to be docked was T1016 (PDB ID 6e4b), despite the predicted PO4 ligand not matching the observed Cl ligand as shown in Figure 6.8 above. As can be seen from Tables 6.20-6.23 the number of predictions per grid box calculation was decreased. This could be due to the size of the ligand and the ligand-binding space and therefore the number of rotations that is possible. Additionally, PO4 was the smallest ligand to be successfully docked so could highlight there is a size threshold for ligands when undergoing docking. The PO4 ligand has a molecular weight of 94.97, as mentioned previously in the chapter targets with small ligands such as Mg were unable to be docked. PO4 is almost four times the size of Mg and almost three times the size of the observed ligand Cl. The 22.5Å gave the best results and improvement in MCC and BDT score with increases from 0.556 and 0.646, respectively to 0.85 and 0.91, respectively. This gave a score increase of 0.294 and 0.354, respectively (and a percentage increase of 53% and 41%, respectively). Whilst this isn't the greatest percentage increase across the docking targets it is the best scores to be achieved following docking. With respect to the FunFOLD3 prediction there were seven correct ligand-binding residue predictions and four incorrect ligand binding residue predictions (see Table 3.9 Chapter 3). In comparison, 22.5Å from FunFOLD3-D correctly predicted six ligand-binding residues and had one incorrect ligand binding residue prediction which was not included in the FunFOLD3 prediction (GLY151). Only one of the correct ligand-binding residues was missed off (HIS8). All of these factors could have contributed to the improvement in MCC and BDT scores.

6.4 Summary and conclusion

As mentioned previously in the introduction, AutoDock Vina has been used previously in order to enhance the ligand-binding sites for COACH and was subsequently called COACH-D (Wu *et al.*, 2018). As a result, AutoDock Vina was selected as a method to improve the ligand-binding residues predicted by FunFOLD3. FunFOLD3-D is the proposed new method, which will utilise docking to improve the predicted ligand-binding sites by rotating the predicted ligand in the binding space, as identified using FunFOLD3. Therefore, FunFOLD3-D will not be a stand-alone method at this stage but an additional step following prediction with FunFOLD3. FunFOLD3-D outputs, up to a total of nine models for generation, as recommended by the developers of AutoDock Vina. However, it is worth noting, that the results in this chapter have shown in certain circumstances nine different ligand-binding sites might be unable to obtain due to the type of ligand involved for the quality of the original prediction. FunFOLD3-D has a clear difference between COACH-D with the utilisation of a box calculation method, a grid size of 22.5Å was chosen as the grid space for docking due to published information. However, the results in this chapter suggest that having various different grid box calculations is worthwhile and the grid box depends on a) the size of the ligand and b) how close the original prediction by FunFOLD3 was to the observed ligand-binding space. i.e. the more distal the predicted and observed ligand, the greater the grid box calculation needs to be. The converse would be true for more proximal ligands. A further consideration for docking is the similarity between the predicted and observed protein structure. For example, CASP13 target T1014 (PDB ID 6qrj) had a MCC and BDT score of -0.05 and 0.05, respectively. The TM score was 0.31010, corresponding towards a poor alignment and therefore with a poor alignment between the predicted and observed structure, this could have impacted on the ability of AutoDock Vina to improve the ligand-binding residues with docking.

FunFOLD3-D is the developed method, following on from the problems identified with FunFOLD3 with protein targets across the various CASP11, CASP12 and CASP13 experiments. Despite 12 of the 17 protein targets having an improvement in MCC and BDT scores. FunFOLD3-D is not without its limitations which is explored below:

1. **Semi-automated and manual curation:** The docking element of FunFOLD3-D is currently reliant on manually checking each of the nine 3D protein-ligand models produced by AutoDock Vina and the user selects the best model. Therefore, this stage of the process is time-consuming. Currently there is no scoring method to pre-select the best model. FunFOLDQA could be potentially used as a method to score these models. This will be an area which can be further developed. Further work is needed to overcome this limitation and this will make FunFOLD3-D a fully incorporated tool in the IntFOLD server pipeline.

2. **Variability in successfully docking ligands:** Not all ligands can be docked. If the ligand is a small metal ion, for example, which is bound in a tight space, there is therefore not enough space for the ligand to rotate and therefore will not benefit from docking, this does not present an immediate problem, as the prediction could stop at the FunFOLD3 stage with larger ligands and/or bigger ligand-binding pockets progressing onto FunFOLD3-D. Thus, FunFOLD3-D becomes an “add-on” method in certain circumstances i.e. a case-by-case basis.

3. **MCC and BDT score threshold:** Difficulty in improving an already good MCC and BDT score following prediction with FunFOLD3. As can be seen from the examples in this chapter if FunFOLD3 has predicted an MCC and BDT score of >0.70 then it is unlikely to be improved further by docking and there are no examples with the protein targets where docking with AutoDock Vina has improved these scores further and in certain circumstances, docking has been more detrimental to the MCC and BDT scores. Therefore,

this raises a potential consideration, of whether docking should be considered on an MCC and BDT score basis, which relates back to the limitations raised in point 2, albeit a different focus

4. **Time-consuming:** At this stage, FunFOLD3-D is currently not a stand-alone method but an add-on to the FunFOLD3 method, as a result this has made the refinement a relatively labour intensive task, although this could become more automated in future versions and as discussed in point 1 the plan is to male FunFOLD3-D fully automated in a similar way to FunFOLD3-D with only the input (refer to Figure 6.2B) requiring manual user input.

Despite the above-mentioned limitations, FunFOLD3-D has demonstrated that ligand-binding residue predictions can be improved even if the predicted and observed ligand do not match or if the predicted ligand is distal to the observed ligand-binding space.

Aside from the limitations above, in order to get to the end goal, this was not without problems. Overall, the success rate of docking software, depends on protein-ligand preparation and sampling of the search space. Protein-ligand preparation was not a problem faced by FunFOLD3-D, the receptor (protein) and ligand files can be easily separated either manually or with code. The output file from FunFOLD3, the lig2.pdb file is in a format which is user friendly and compatible with AutoDock Vina. The biggest problem faced by FunFOLD3-D was point number 2; sampling of the search space. As mentioned previously, the 22.5Å grid box calculation was published in literature. Therefore, it was an obvious decision to explore this, however in order to test the method and determine if there was a better option, other grid box calculations needed to be explored. The final decision was not an easy one and was based on trial and error of grid box calculations on a few selected targets. Initial grid box calculations ranged from 5% to as high as 100% of the protein model. The extremes were soon disregarded as 5% did not allow for much improvement if the MCC and BDT scores were at the lower end and 100% was incorporate too much of the protein model. In general, FunFOLD3 can be seen as identify the rough location of a ligand, with docking being utilised to make the ligand location more precise. Thus, having an 100% grid box calculation defeats this purpose and 5% was having a tendency to favour the FunFOLD3 methodology results, which would mean the method hadn't been developed sufficiently. Therefore, 50% was decided as this would take into account the original ligand-binding site and also part of the receptor (protein) which may be more suitable as the ligand-binding site. The 10% and 20% grid box calculations were selected, as there are two options when sampling a search space for docking; full protein or a region focused on the known binding site. As mentioned earlier, sampling on a full protein has limitations, so a region focused on the known binding site was preferred with 5% being ruled out, as mentioned previously and thus 10% and 20% seemed a reasonable search space.

The plan moving forward with FunFOLD3-D is to make it a downloadable software, like the predecessor FunFOLD3 and to add a scoring metric to select the best ligand-binding poses, rather than having to manually verify, with the idea, as mentioned in point 1 utilising FunFOLDQA. In the short-term, FunFOLD3-D be will assessed on further protein targets obtained from the CASP experiment, namely the CASP14 experiment.

The longer term vision is to obtain consistently good MCC and BDT scores (>0.70) across a variety of ligands (e.g. small metal ions and co-factors) and protein targets. Once this consistency has been reached it can assist with confidence of going into the "unknown" with protein -ligand and ligand-binding site predictions. Finally, the initial positive results from this chapter and the novelty of the FunFOLD3-D methodology, with the four different grid box calculations has been accepted as a book chapter. To further expand of the novelty of the FunFOLD3-D method, the code created for to explore the 10%, 20% and 50% grid box calculations, can be adapted to any grid box calculation and therefore highlights the versatility of the methodology for users. At the time of writing, this is in draft form but has been accepted by the publishers (refer to Appendix 1, Arvinas Book Chapter).

In conclusion, preliminary results from FunFOLD3-D shows that across 17 protein targets from CASP11, CASP12 and CASP13 with 20 ligands, FunFOLD3-D was able to improve the MCC and BDT scores for 12 of the ligand-binding sites with varying results. As a result, FunFOLD3-D is a viable option for refining the prediction of ligand-binding residues from FunFOLD3. Lastly, consideration needs to be given that the MCC and BDT score can only be determined, when there is an answer, e.g. an observed structure. Hence, why it is important to make improvements to FunFOLD3 in way that can be benchmarked so that FunFOLD3 can be applied to other aspects of ligand-binding site prediction and ligand-prediction.

Chapter six of this thesis will apply what has been learned about the strengths and limitations of the methods that have been developed and benchmarked (in Chapters 3-6) to discover more about the SARS-CoV-2 proteins with unknown structures and functions.

Chapter 7: Application of FunFOLD3: SARS-CoV-2 Case Report

7.1 Introduction

7.1.1 The SARS-CoV-2 Infection

In December 2019, a novel coronavirus infection emerged in Wuhan, China. At the time, it was given a provisional name of 2019 novel coronavirus (2019-nCoV) and rapidly spread across China and many other countries across the world (Feinstein & Brylinski, 2015). Other names for the virus were, novel coronavirus pneumonia or Wuhan pneumonia and fever was the most common symptom, followed by cough. On 11th February 2020, the World Health Organisation (WHO) announced a new name for the disease caused by the novel coronavirus – COVID-19 and the virus itself was renamed to severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses (Lai *et al.*, 2020).

Early reports suggested a link to a single local fish and wild animal or “wet” market as a possible source of emergence and thereby suggesting animal-to-human transmission, the rapid spreading of the virus globally demonstrated human-to-human transmission through droplets via the nose and/or mouth and direct contact (Lai *et al.*, 2020). On 30th January 2020, the WHO declared the COVID-19 outbreak as the sixth public health emergency of international concern. Following H1N1 (2009), polio (2014), Ebola in West Africa (2014), Zika (2016) and Ebola in the Democratic Republic of Congo (2019) and thereby making COVID-19 a global pandemic (Lai *et al.*, 2020). At the time of writing (February 2021), SARS-CoV-2 has affected more than 109,217,366 patients worldwide with 2,413,912 reported deaths in 223 countries (*World Health Organisation*, 2021).

Using molecular methods, SARS-CoV-2 was found to be a positive-sense, single-stranded RNA virus belonging to the genus *Betacoronavirus* (Lai *et al.*, 2020). Phylogenetic analysis revealed that SARS-CoV-2 is closely related (88–89% similarity) to two bat-derived SARS-like coronaviruses, namely bat-SL-CoVZC45 and bat-SL-CoVZXC21. However, it is more distant from SARS-CoV (~79% similarity) and Middle East respiratory syndrome coronavirus (MERS-CoV) (~50% similarity) (Chan, et al., 2020).

COVID-19 is characterised by three clinical patterns; no symptoms, mild to moderate disease, severe pneumonia requiring admission to intensive care unit and supplementary oxygen of which occurs in up to 31% of infected patients (Chan, et al., 2020). As SARS-CoV-2 is a novel virus with high virulence and mortality, further study to shed light on the immunopathogenesis of COVID-19 was required for the development or repurposing of antiviral medications to make the disease treatable and/or curable or vaccine development to make the disease preventable (Huang et al., 2020). Figure 7.1 illustrates the classification of disease states for COVID-19, highlighting the importance of the viral load early on in the disease.

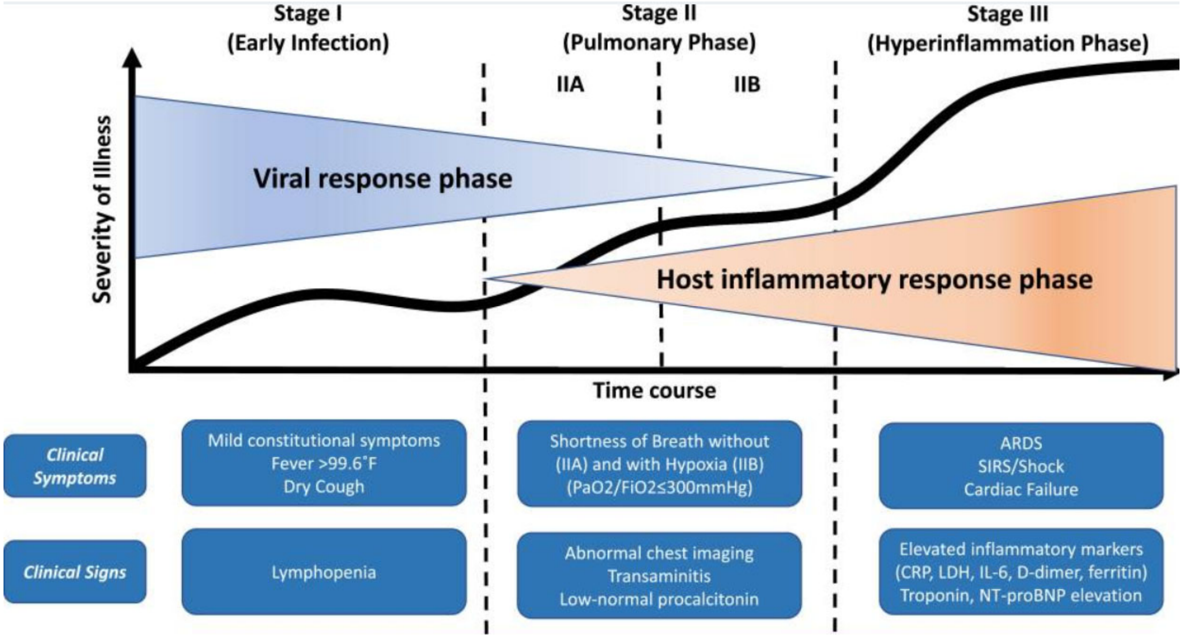


Figure 7.1. Classification of COVID-19 disease states
Figure 7.1 illustrates the three escalating phases of COVID-19 disease progression, with associated signs, symptoms, and potential phase-specific therapies. ARDS, acute respiratory distress syndrome; CRP, C-reactive protein; LDH, lactate dehydrogenase; NT-proBNP, N-terminal pro B-type natriuretic peptide; SIRS, systemic inflammatory response syndrome. Figure adapted from Seif et al., 2020

7.1.2 The structure of the virus

Coronaviruses belong to *Nidovirales* order, which are viruses that depend on a nested group of mRNAs for their replication (nido = nest). They contain the biggest known viral RNA genomes (27–32 kb in length) and are enveloped positive-sense single-stranded RNA viruses (Huang et al., 2020). The virus has a characteristic crown-like shape due to the presence of spike proteins on the surface and hence the name corona, which is served from the Latin *corona* meaning crown.

The spike or S protein is extensively glycosylated, forming a homotrimer and mediates virus entry through binding to specific receptors (e.g. angiotensin-converting enzyme 2 (ACE2)) and fusion with the cell membrane of the host (Figure 7.1). The S protein harbours the major antigen that stimulates the formation of neutralising antibodies, in addition to targets of cytotoxic lymphocytes. The membrane (M) protein (a glycoprotein, that spans the membrane bilayer, thus a transmembrane protein) (Thomas, 2020) has a major role in viral assembly (Saber-Ayad, Saleh and Abu-Gharbieh, 2020), the nucleocapsid (N) protein is responsible for the regulation of viral RNA synthesis and interacts with the M protein during the budding of the virus and forms part of the nucleocapsid (in association with the RNA).

SARS-CoV-2 has sequence homology with the influenza virus through hemagglutinin-esterase glycoprotein (HE). The HE binds to neuraminic acid on the host cell surface, leading to the adsorption of the virus on the cell surface. Such homology may denote an early recombination between the two viruses. Proteins M, N, and E are essential for the virus’ assembly and release (Tseng *et al.*, 2010) however, the exact function of the envelope (E) protein is not fully elucidated.

Additionally, Figure 7.2 demonstrates several pathways of the viral entry into the host cell which can be targeted by drugs and ultimately hinder viral replication and development of symptoms associated with COVID-19.

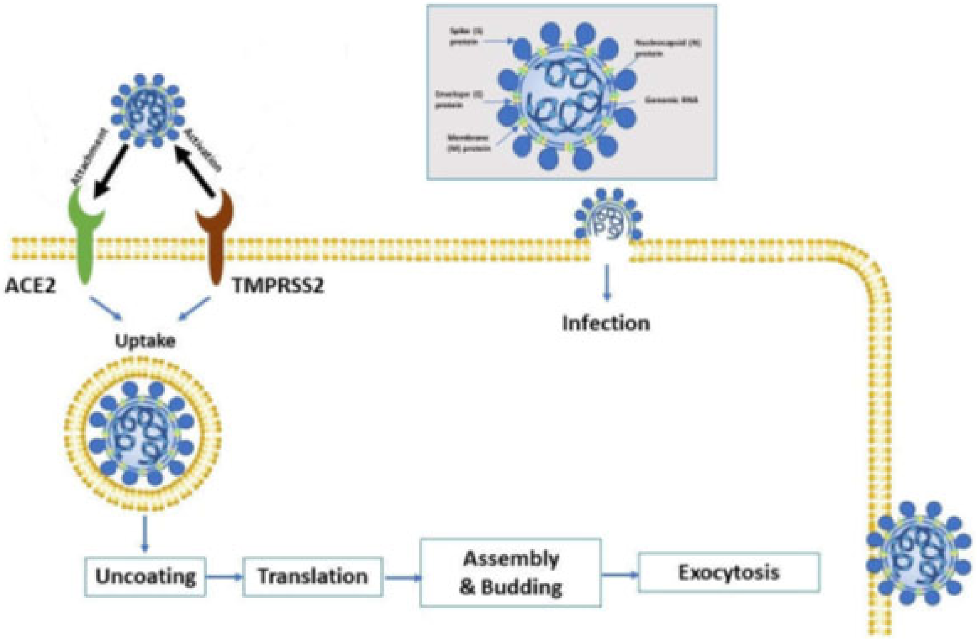


Figure 7.2. SARS-CoV-2 entry into the host cell
The attachment protein "-spike glycoprotein" of the severe acute respiratory syndrome-2 (SARS-CoV-2) uses a cellular attachment factor (angiotensin-converting enzyme 2 (ACE2)) and uses the cellular protease TMPRSS2 (transmembrane protease serine 2) for its activation. ACE2 can be activated via either losartan or recombinant human ACE 2 (rhACE2). Potential pharmacotherapeutic approaches include the use of camostat mesylate (which is a TMPRSS2 inhibitor) to block the priming of the spike protein, increasing the number of ACE2 receptors via losartan, and the use of soluble recombinant human ACE2 (which should slow viral entry into cells Is a competitive binding with SARS-CoV-2). The structure of SARS-CoV-2 is shown in the upper right. Figure taken from Siu et al., 2008

Table 7.1 below lists the function, or at least what is known, of the different proteins that form the SARS-CoV-2 genome. Where applicable solved experimental structures are also provided.

Table 7.1. Function annotation of SARS-CoV-2 genome
Table adapted from Zhang Lab (https://zhanggroup.org//COVID-19/index.html#download) based on UniProt curation of SARS-CoV-2 (Magrane & UniProt Consortium, 2011). Some of the functions have been identified specifically to the proteins which are part of the SARS-CoV-2 genome and some functions have been inferred from previous literature information

Protein name	Protein function	Solved experimental structure (where applicable)
nsp 1 (host translation inhibitor)	Inhibits host translation by interaction with the 40S ribosomal subunit, the C-terminal domain of nsp1 binds to the ribosomal mRNA channel to prevent host mRNA binding. By suppressing host gene expression, nsp1 facilitates efficient viral gene expression in infected cells and evasion from host immune response (Schubert *et al.*, 2020).	7k3n
nsp2 (non-structural protein 2)	Function not entirely known. May play a role in the modulation of host cell survival signalling pathway by interaction with host prohibitin 1 and 2 (PHB and PHB2).Protein is conserved in SARS-CoV. (Cornillez-Ty *et al.*, 2009). Prohibitin plays a role in maintaining the functional integrity of the mitochondria and protecting cells from various stresses	N/A
Papain-like proteinase	Together with nsp4 in the assembly of virally-induced cytoplasmic double-membrane vesicles required for viral replication (Lei *et al.*, 2020). Antagonises innate immune induction of type I interferon by blocking phosphorylation, dimerisation and subsequent nuclear translocation of host IRF3 (Lei *et al.*, 2020).	7kag (amino acid range 1-111) 6w6y (amino acid range 207-379) 6w9c (amino acid range 748-1060)

	Prevents host NF-kappa-B signalling (Lei *et al.*, 2020).	
nsp4 (non-structural protein 4)	Along with papain-like proteinase. Participates in the assembly of virally induced cytoplasmic double-membrane vesicles necessary for viral replication (Lei *et al.*, 2020).	N/A
Proteinase 3CL-PRO	Required for the processing viral polyproteins, that are translated from the viral RNA. This generates a functional replicase complex and enables viral spread (Zhang et al., 2020).	6lu7
nsp6 (non-structural protein 6)	Antagonises interferon production to evade host anti-viral defence (Cao et al., 2020).	N/A
nsp7 (non-structural protein 7)	Forms a hexadecamer with nsp8 that may participate in viral replication by acting as a primase (Peng et al., 2020; Konkolova et al., 2020).	6m71
nsp8 (non-structural protein 8)	Forms a hexadecamer with nsp7 (refer to nsp7) (Peng et al., 2020; Konkolova et al., 2020).	7cyq
nsp9 (non-structural protein 9)	May participate in viral replication, overall virulence and viral genomic RNA reproduction by binding to RNA, although exact mechanism is unknown (Littler *et al.*, 2020).	6w4b
nsp10 (non-structural protein 10)	Plays a role in viral mRNAs methylation potentially with nsp16 (Viswanathan *et al.*, 2021)	6w75
RdRp (RNA-directed RNA polymerase)	Responsible for replication and transcription of the viral RNA genome. Comprised of a catalytic subunit as nsp12 and two accessory subunits, nsp8 and nsp7 (Hillen *et al.*, 2020)	6m71
Helicase	RNA and DNA duplex-unwinding activities and activity of helicase is dependent on magnesium (Chen *et al.*, 2020)	5rl9
Proofreading exoribonuclease/Guanine-N7	Involved in viral mRNA cap synthesis. The RNA cap has several important biological roles in viruses as it is crucial for the stability of mRNAs, for translation and to evade the host immune response (Romano *et al.*, 2020)	N/A

NendoU (Uridylate-specific endoribonuclease)	Potentially has two roles: 1. Responsible for protein interface within the innate immune response (Deng *et al.*, 2017) 2. Degrades viral RNA to hide it from host defenses (Kim et al., 2020)	6vww
2'-O-methyltransferease	Plays an essential role in viral mRNAs methylation to improve viral protein translation which is essential to avoid host immune detection (Rosas-Lemus *et al.*, 2020)	6w75
Spike glycoprotein (S)	Attaches the virion to the cell membrane by interacting with the host receptor (ACE2). Binding to receptor causes internalisation of the virus (Hoffmann *et al.*, 2020)	6vyb (open state) 6vxx (closed state) 6lxt (amino acid range 912-988, 1164-1202)
ORF3a (open reading frame 3a)	Forms homotetrameric potassium sensitive ion channels (viroporin) and involved in virus release via lysosomal trafficking (Miao *et al.*, 2021)	6xdc
E (envelope)	Acts as a viroporin and self assembles in host membranes forming pentameric protein-lipid pores that allow ion transport. Additional function, as per membrane (M) protein with regulation of the localisation of the spike (S) protein at cis-Golgi, the place of virus budding (Boson et al., 2020)	7k3g
M (membrane)	A transmembrane Component of the viral envelope that plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins. Regulates the localisation of spike (S) protein at cis-Golgi, the place of virus budding (Boson et al., 2020)	N/A
ORF6 (open reading frame 6a)	Blocks expression of interferon stimulated genes that display antiviral activities (Xia et al., 2020)	N/A
ORF7a (open reading frame 7a)	Alters host immune response by binding to specific immune cells and is involved in the cytokine storm (Zhou *et al.*, 2021)	6w37 (amino acid range 16-82)

ORF8 (open reading frame 8)	Plays a role in modulating host immune response. Causes an increase in proinflammatory cytokines and contributes to cytokine storm in COVID-19 (Chan et al., 2020)	7jtl
N (nucleocapsid)	Packages the positive strand viral genome RNA into a helical RNP particle and plays a vital role during virion assembly. Plays an important role in enhancing the efficiency of viral RNA transcription and replication (Bai *et al.*, 2021)	6m3m (amino acid range 50-174) 6yun (amin acid range 249-364)
ORF10 (open reading frame 10)	Function is unclear (Bateman *et al.*, 2021)	N/A

7.1.3 CASP Commons

At the end of January 2020, the genome of the virus had been decoded and in order to help understand the virus spread and function, the CASP organisers launched the SARS-CoV-2 structure modelling initiative (Saber-Ayad, Saleh and Abu-Gharbieh, 2020). This initiative was part of CASP Commons and made the biggest contribution towards generating and evaluating models for the virus and in particular the proteins and domains for where there was no experimental structure available (The Protein Prediction Center, 2020). The goal was to obtain the best possible consensus 3D models and then use these models to assist in gaining further insight into the virus' structure and function. Additionally, a subsequent goal would be to identify possible epitopes that can be used for vaccine development and finally, evaluate drug targeting strategies by the identification of ligand-binding sites (The Protein Prediction Center, 2020). The identification of ligands, ligand-binding sites and GO terms will be done by FunFOLD3. If ligands, ligand-binding sites and GO terms can be identified then this can provide valuable insights into the relatively limited knowledge that is known about SARS-CoV-2 and could help in explaining the pathophysiology of COVID-19.

The objectives of the CASP Commons program was firstly, to provide the best quality predicted structures, which could be used to aid investigators working on these proteins.

Secondly, they would provide a basis for other CASP competition targets (e.g. CASP14), and thirdly for those proteins where experimental structure is not undertaken or unsuccessful, members of the CASP community will be invited to contribute structure models.

Aim: The aim of this chapter was to assist in the elucidation of function and potential ligand-binding sites for the ten CASP Commons SARS-CoV-2 protein targets using FunFOLD3. The purpose of this chapter is to apply FunFOLD3 into the unknown for the first occurrence in this thesis and to demonstrate how FunFOLDD3 can be applied outside of the CASP experiments.The identification of templates with possible ligands could help in identifying the function of the virus' proteins/domains in the absence of identification of ligands. Where ligands could be identified, then this could provide insight into possible targets for drug repurposing.

7.2 Materials and Methods

As with previous CASP competitions, the CASP Commons organisers provided the protein sequences for the ten protein targets of the SARS-CoV-2 targets that required 3D modelling and subsequent predictions of ligand-binding sites. The CASP Commons target ID, amino acid sequence and name of the protein targets are given in Table 7.2. CASP Commons occurred in two rounds in order to iteratively enhance the prediction of 3D protein/domain structures and ligand-binding sites. The methodology has been described previously in Chapter 2. Briefly, the top-ranked 3D model, a list of templates some of which may contain biologically relevant ligands and the fasta sequence were inputed to determine ligands and ligand-binding residues. The output was a list of templates containing biologically relevant ligands and if a ligand and ultimately a ligand-binding residues was identified this would be the output, along with GO terms. In instances where FunFOLD3 does not predict a biologically relevant ligand the functional predictions from IntFOLD6, where applicable, will be presented.

Starting models were from our groups updated manual tertiary structure prediction pipeline which was developed for CASP14 (see the McGuffin group CASP14 abstract for further details, https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf)

Ultimately, the targets were analysed in the same way CASP12 and 13 targets without PDB IDs have been reported (refer to Chapter 3). Additionally, given the evolution of knowledge with COVID-19, CASP Commons occurred in stages. Stage one focused on predictions for all the residues of the target, stage two focused on specific residues for some of the targets or some targets were modelled again in hopes of improving the modelling from the first round. A description of the stages is provided below:

Stage 1: Full chain sequences were modelled and scored. Then, where applicable, domains were released for modelling. Domain models were scored using QA methods, such as ModFOLD8 (McGuffin, Aldowsari and Alharbi, 2021).

Stage 2: A further stage of full chain sequence modelling, however not all methods or groups that participated in stage 1, participated in stage 2 hence the drop in scores between stage 1 and stage 2

Due to domains being scored on the full chain sequence to ensure consistency for residue numbering, the global scores for domains were lower than expected as normalised based on a full length sequence rather than shorter domains

Following modelling and scoring with ModFOLD8, structures were refined with ReFOLD (McGuffin & Adiyaman, 2021) and then FunFOLD3 was utilised to determine ligand and ligand-binding residues.

Across the two stages this yielded a total of 32 targets for ten protein targets. At the time of writing, two of the targets had PDB IDs associated; C1905 (PDB ID 6xdc) and C1908 (PDB ID 7tjl). Out of the 10 proteins targets, FunFOLD3 was able to predict ligand and ligand-binding site residues for two targets in round one and ten targets in round two. Even for targets without ligands and ligand-binding site residues, results will be presented around the templates in order to contribute towards this novel situation. Templates which were identified as containing a biologically relevant ligand are highlighted in the results tables.

Table 7.2. Amino acid sequence of CASP Commons targets

CASP Commons target ID	Amino acid sequence	Protein
C1901	AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKEIIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG	nsp2
C1902	KIVNNWLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFANKHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTILRTTNGDFLHFLPRVFSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVAYESLRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTSGRWVLNNDYYRSLPGVFCGVDAVNLLTNMFTPLIQPIGALDISASIVAGGIVAIVVTCLAYYFMRFRRAFGEYSHVVAFNTLLFLMSFTVLCLTPVYSFLPGVYSVIYLYLTFYLTNDVSFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFEEAALCTFLLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHLAKALNDFSNSGSDVLYQPPQTSITSAVLQ	nsp4
C1903	SAVKRTIKGTHHWLLLTILTSLLVLVQSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFVKHKHAFLCLFLLPSLATVAYFNMVYMPASWVMRIMTWLDMVDTSLSGFKLKDCVMYASAVVLLILMTARTVYDDGARRVWTLMNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQ	nsp6
C1904	KPANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSRVLGLKTLATHGLAAVNSVPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFFTLLLQLCTFTRSTNSRIKASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKLINIIIWFLLLSVCLGSLIYSTAALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPCSVCLSGLDSLDTYPSLETIQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQLFFSYFAVHFISNSWLMWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSSTCMMCYKRNRATRVECTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFISDEVARDLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPINVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAEVAVKMFDAYVNTFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVDSDVETKDVVECLKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIALIWNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGG	PL-PRO
C1905	MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFLYLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDYCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYYQLYSTQLSTDTGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL	ORF3a
C1906	MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPVTLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILLNVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYKLGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ	Transmembrane protein
C1907	MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEID	ORF6
C1908	MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIELCVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDFI	ORF8
C1909	MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT	ORF10
C1910	MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCHA	ORF7b

7.3 Results

Analysis of CASP Commons COVID-19 predictions

A total of ten protein targets were available for analysis. Given the uniqueness of COVID-19, suggested proteins targets where the experimental structure would not be immediately available, it is possible that additional experimental structures may be released at a later stage.

An overview of the ligands predicted by FunFOLD3 or the FunFOLD3 component of IntFOLD6 are given below in Table 7.3

Table 7.3. Predicted ligands for CASP Commons SARS-Cov-2
Predictions were made by either standalone FunFOLD3 or the FunFOLD3 component of IntFOLD6. For the two targets which had an actual crystal structure (C1905 and C1908) no biologically relevant ligands were identified. The protein name is provided in brackets

CASP Commons Target ID	Predicted ligand	Further information
C1901d2 (nsp2)	Palmitic acid (PLM)	Predicted by FunFOLD3 component of IntFOLD6
C1902d3 (nsp4)	Chlorophyll A (CLA)	Predicted by FunFOLD3 component of IntFOLD6
C1903d1 (nsp6)	HEME (HEM)	Predicted by FunFOLD3 component of IntFOLD6
C1903x2 (nsp6)	(2~{S})-2,3-bis(oxidanyl)propyl] (~{E})-undec-2-enoate (MUN) [(Z)-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG) Retinal (RET)	Predicted by FunFOLD3 component of IntFOLD6
C1904d1 (PL-PRO)	Palmitic acid (PLM)	Predicted by FunFOLD3 component of IntFOLD6
C1904d2 (PL-PRO)	[(Z)-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG)	Predicted by FunFOLD3 component of IntFOLD6
C1904d3 (PL-PRO)	Guanosine 5' Diphosphate (GDP)	Predicted by FunFOLD3 component of IntFOLD6
C1904x2 (PL-PRO)	ALA, SER, ASN	Predicted by FunFOLD3 component of IntFOLD6
C1905d1 (ORF3a)	[(Z)-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG)	FunFOLD3

C1906d1 (Transmembrane protein)	Duocarmycin A (DUO) [(Z)-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG)	FunFOLD3
C1906x2 (Transmembrane protein)	Glycerol (GOL), ASN, LEU	FunFOLD3
C1907 (ORF6)	Calcium (CA) HEME (HEM), Methyl 2-(2,5-dihydroxyphenyl)acetate (DCO) methyl 2-(3,5-dihydroxyphenyl)ethanoate (XQI)	FunFOLD3
C1908 (ORF8)	Calcium (CA)	FunFOLD3
C1909 (ORF10)	24-methylenecholesterol (94R)	Predicted by FunFOLD3 component of IntFOLD6
C1910 (ORF7b)	Chlorophyll A (CLA)	FunFOLD3

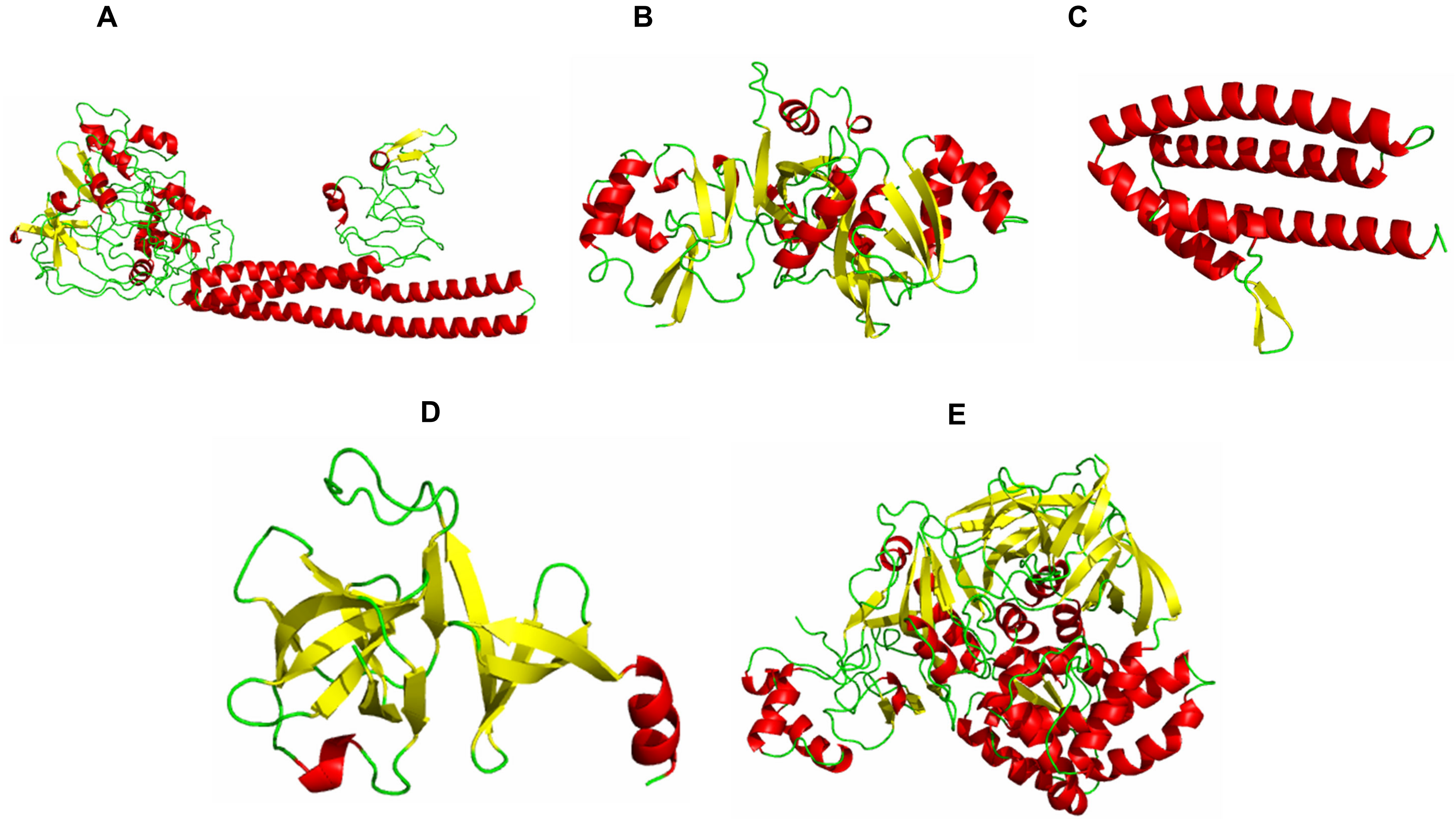


Figure 7.3. Top-scoring structure predictions from the McGuffin group for nsp2 (CASP Commons target C1901) full chain and individual domains
(A) Predicted structure for nsp2 (C1901) shown as cartoon and coloured by secondary structure with helices coloured red, sheets coloured yellow and loops coloured green. Residue length of 638 **(B)** Predicted structure for nsp2 (C1901d1) shown as cartoon and coloured by secondary structure with a residues 1-359 modelled **(C)** Predicted structure for nsp2 (C1901d2) shown as cartoon and coloured by secondary structure with residues 360-499 modelled **(D)** Predicted structure for nsp2 (C1901d3) shown as cartoon and coloured by secondary structure with residues 500-638 modelled **(E)** Predicted structure for nsp2 (C1901x2) shown as cartoon and coloured by secondary structure. This was the whole target of 638 residues following stage two modelling

The first predicted CASP Commons target (C1901) was nsp2 proteinand is described as a a non-structural protein 2. Non-structural protein 2 is one of four nsps and is an essential component of the replicative complexes. It contains helicase and RNA triphosphatase activities, required for RNA synthesis, and is also a protease that orchestrates sequential cleavages of non-structural polyprotein precursor P1234 (Frolova *et al.*, 2002). Additionally, nsp2 most likely acts by decreasing interferon (IFN) production and minimises virus visibility (Frolova *et al.*, 2002). Interferon gamma (IFNγ) is a cytokine that is critical for innate and adaptive immunity against viral, some bacterial and protozoal infections (Tau & Rothman, 1999).

There were five different variations of the structure and as can be seen from the results in Figure 7.3, no ligands were predicted across the five different variations by FunFOLD3. As per CASP Commons ID C1901 (nsp2) was the first stage modelling of the whole structure and C1901x2 was stage 2 modelling. The global model quality score had improved from 0.4146 for C1901 to 0.4701 for C1901x2 (nsp2) according to ModFOLD8 (McGuffin, Aldowsari and Alharbi, 2021) following the refinement with ReFOLD3 (McGuffin & Adiyaman, 2021). As there is no PDB ID associated with the target, no comparisons with the actual structure can be made at present.

As per the CASP Commons ID C1901d1, C1901d2, C1901d3 (all nsp2 protein) are sub-domains of the protein with global model quality scores of 0.2850, 0.1181 and 0.1146, respectively, although these scores are based on the full length structure. The global scores increased to 0.4887, 0.5270 and 0.5060, when scores are recalculated based on the shorter individual domains sequences. The global model quality scores range between 0 and 1. In general, scores <0.2 indicate there may be incorrectly modelled domains and scores >0.4

generally indicate more complete and confident models, which are highly similar to the native structure.

In comparison, IntFOLD6 obtained a prediction for C1901d2 (nsp2) and is depicted in Figure 7.4. In comparison, the global model quality score was 0.5280. The predicted ligand was palmitic acid (PLM). Palmitic acid is the most common saturated fatty acid found in the human body and can be provided in the diet or synthesised endogenously from other fatty acids, carbohydrates and amino acids (Carta *et al.*, 2017). Palmitic acid undergoes tight homeostatic control and this is likely related to its fundamental physiological role to guarantee membrane physical properties and in the lung is has efficient surfactant activity (Carta *et al.*, 2017).

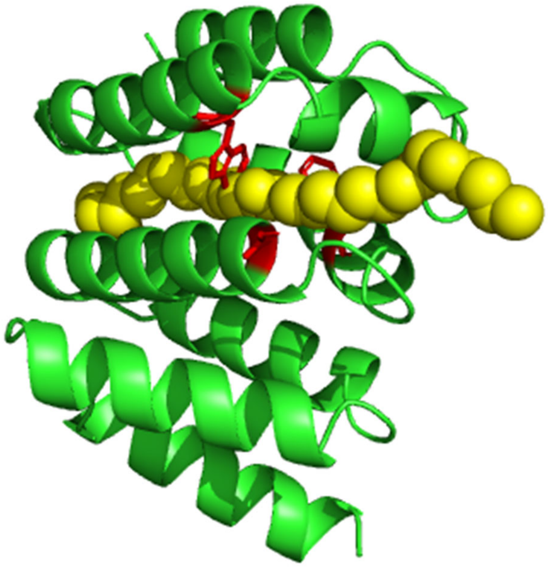


Figure 7.4. IntFOLD6 prediction for nsp2 (C1901d2)
Predicted structure for nsp2 (C1901d2) shown as cartoon and coloured green the PLM ligand is shown as sphere and coloured yellow. The ligand-binding site residues are 67,78,107 and are shown as sticks and coloured red

Table 7.4, below shows the templates associated with the target with the templates identified as having ligands by FunFOLD3 highlighted in yellow. The role/function of each protein and ligand as per the PDB entry is also depicted. How this relates to the description of the protein will be discussed later in the chapter.

Table 7.4. Template list for nsp2 (C1901)
Templates for nsp2 (C1901) with the role/function and ligand as per PBD ID. Templates identified by FunFOLD3 to contain biologically relevant ligands in the output are highlighted in yellow.

Template	Role/function	Ligand
1b3uA	scaffold protein	N/A
1dx5I	hydrolase/hydrolase inhibitor	CA
1is2A	oxidoreductase	FAD
1kloA	Glycoprotein	N/A
1yvlA	Signaling protein	TYR-ASP-LYS-PRO-HIS
2cvcA	electron transport	HEM
2d5bA	isomerase	ZN
2fd6U		
2fiyA	structural genomic/unknown function	FE
2jneA	Metal binding protein	ZN
2jrpA	structural genomic/unknown function	ZN
2k2dA	Metal binding protein	ZN
2lggA	Ligase/DNA binding protein	ZN
2wscF	photosynthesis	CLA
2yheA	Hydrolase	ZN, SO4
3h9bA	Ligase	ZN, NOT
3ld1A	Hydrolase	N/A
3lw5F	photosynthesis	CLA
3ouqA	electron transport	HEM
3txaA	cell adhesion	MG
4aybP	Transferase	ZN
4c8vA	Signalling protein	N/A
4cb8A	Apoptosis	SO4
4cb9A	Apoptosis	N/A
4fyeA	Hydrolase	PO4
4jc8A	Transport protein	N/A
4nurA	Hydrolase	ZN
4pdxA	Hydrolase	SO4, GOL
4tqlA	De novo protein	N/A
4ui9I	Cell cycle	ZN
4uosA	De novo protein	N/A
4uvkA	Cell cycle	N/A
4wrtC	Transferase/RNA	DNA/RNA
4ziqA	Membrane protein	GOL, CL
4zyaA	Ligase	ZN
5bptA	Cell cycle	N/A
5cwbA	De novo protein	N/A
5icuA	Chaperone	CU
5mspA	oxidoreductase	NAP
5owvC	Lipid binding protein	N/A
5urbA	Ligase	ZN
5vchA	Protein transport	N/A

5wgrA	Oxidoreductase/oxidoreductase inhibitor	FAD, PM7
5yjgA	cell adhesion	C, CA, MG
6ab7C	Viral protein	N/A
6ahcA	Hydrolase	N/A
6bzaA	Flavoprotein	13X, FAD
6dvwA	Membrane protein	N/A
6gapA	Viral protein	N/A
6jfkA	Membrane protein	GDP
6lthL	gene regulation	ZN
6nctA	Transferase	144, SO4
6nwfa	Membrane protein	RET, C14, BGC, D10, CL
6qp1a	Lyase	LCS
6rl5a	Transferase	PLP
6slfa	Hydrolase	LJ8, GOL, ACT, BTB, ZN
6snhx	Membrane protein	LMH
6u7kA	Viral protein	NAG-NAG-BMA-MAN-MAN-MAN, NAG-NAG, NAG
6ueha	Hydrolase	ACT, CA

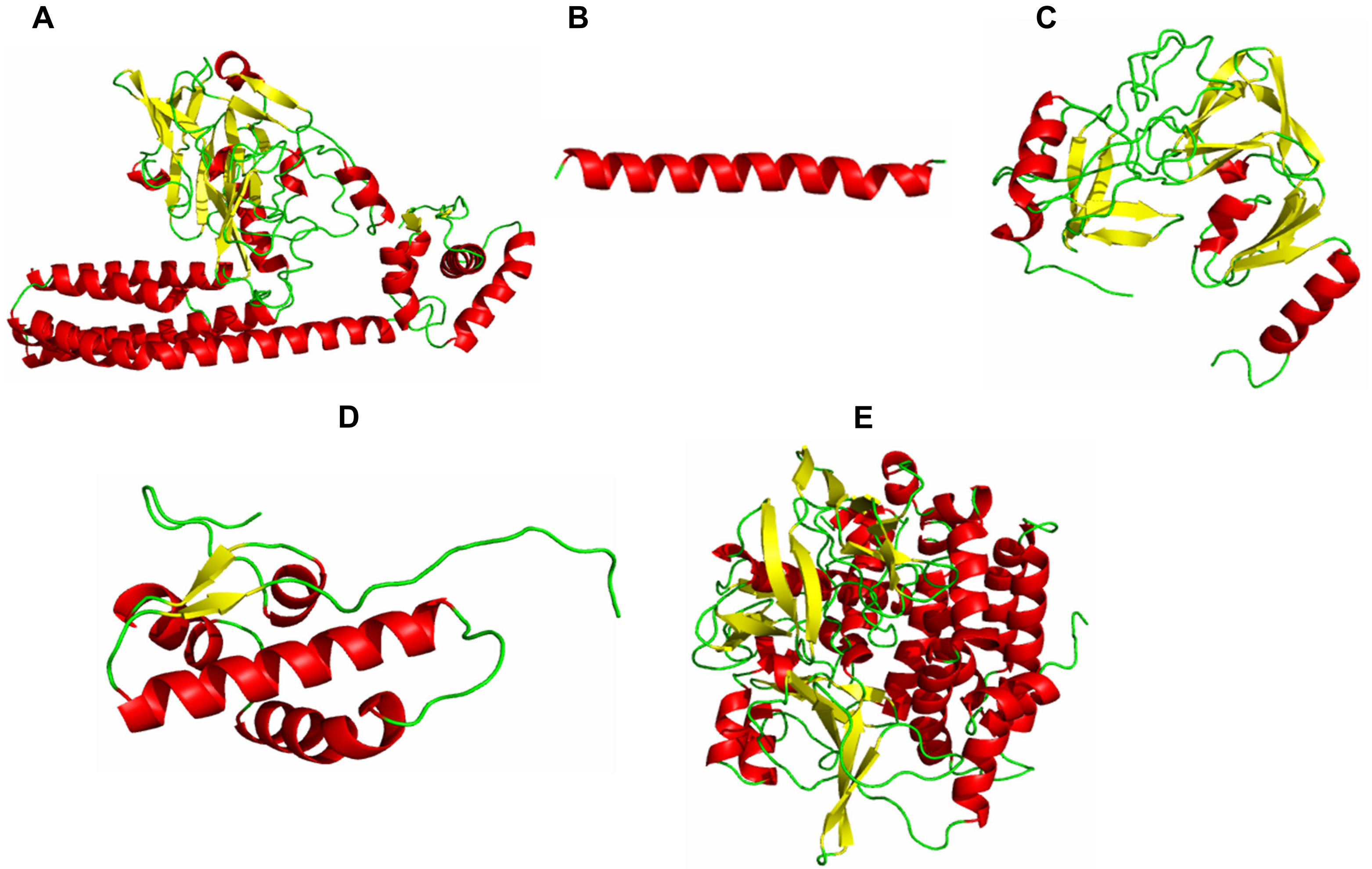


Figure 7.5. Top-scoring structure predictions from the McGuffin group for nsp4 (CASP Commons target C1902) full chain and individual domains
(A) Predicted structure for nsp4 (C1902) shown as cartoon and coloured by secondary structure with a residue length of 500 **(B)** Predicted structure for nsp4 (C1902d1) shown as cartoon and coloured by secondary structure with a residues 1-32 + 279-400 modelled **(C)** Predicted structure for nsp4 (C1902d2) shown as cartoon and coloured by secondary structure with residues 33-278 modelled **(D)** Predicted structure for nsp4 (C1902d3) shown as cartoon and coloured by secondary structure with residues 401-500 modelled **(E)** Predicted structure for nsp4 (C1902x2) shown as cartoon and coloured by secondary structure. This was the whole target of 500 residues following stage two modelling

The second predicted CASP Commons target is nsp4 or non-structural protein 4 (CASP Commons target C1902). Non-structural protein 4 can also be identified in *rotaviruses* as rotavirus nsp4 (Pham *et al.*, 2017). *Rotavirus* nonstructural protein 4 is an endoplasmic reticulum transmembrane glycoprotein that has a viroporin domain and nsp4 viroporin activity elevates cytosolic Ca^{2+} in mammalian cells (Pham *et al.*, 2017). Viroproteins have been predominantly identified in animal viruses and disruption of host cell Ca^{2+} homeostatis is critical for virus replication and pathogenesis (Pham *et al.*, 2017).

There were five different variations of the structure for prediction and as can be seen in Figure 7.5, there were no ligands or ligand-binding site residues predicted by FunFOLD3. As with C1901 (nsp2), C1902 (nsp4) was the first round of full structure modelling and achieved a global model quality score of 0.5057 and with C1902x2 (nsp4) the score had decreased to 0.4847. The global model quality scores were 0.1654, 0.236 and 0.1202 for C1902d1 (nsp4), C1902d2 (nsp4) and C1902d3 (nsp4), respectively, when based on the full chain reference sequence. As with C1901 (nsp2), no actual structure has been confirmed, so no comparisons with an actual structure can be made at present.

In comparison, IntFOLD6 obtained a prediction for the C1902d3 (nsp4) target sequence. In comparison, the global model quality score was 0.5942. The predicted ligand was chlorophyll A (CLA) and is depicted below in Figure 7.6.

It is worth mentioning that, chlorophyll and haem contain a common precursor, protoporphyrin IX. Both haem and chlorophyll contain metal containing components with iron being the metal complex in haems such as haemoglobin and myoglobin. In comparison, magnesium is characteristic of all chlorophylls and bacteriochlorophylls (Hendry & Jones, 1980). There are early literature reports attempting to unify the evolution of haemoglobin, cytochromes and chlorophylls with suggestions that porphyrin biosynthesis is essentially similar across the biological systems and the variation arose between haem and chlorophyll as a result of the adaptability of the porphyrin structure to differences in the biochemical requirements (Hendry & Jones, 1980). Therefore, it is likely that IntFOLD6 has identified a template in which chlorophyll has been identified as a biologically relevant ligand, however due to evolutionary changes it is likely that haem would be more suitable. Additionally, studies have shown that the non-structural proteins of SARS-CoV-2 can bind porphyrin, such as haem (Wenzhong & Hualan, 2020). Initially, viral proteins such as SARS-CoV-2 may attack haemoglobin, causing haem to dissociate into iron and porphyrins and then viral proteins capture the porphyrin (Wenzhong & Hualan, 2020). This attack causes respiratory distress and coagulation reaction (Wenzhong & Hualan, 2020). Furthermore, the virus and the porphyrin cause a complex which could enhance evasion of the virus (Wenzhong & Hualan, 2020).

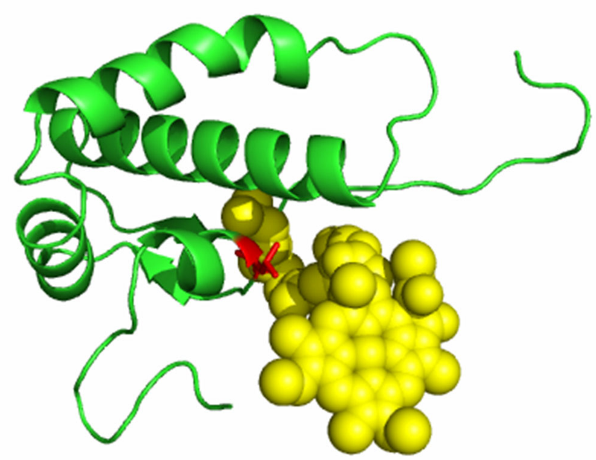


Figure 7.6. IntFOLD6 prediction for nsp4 (C1902d3)
Predicted structure for nsp4 (C1902d3) shown as cartoon and coloured green the CLA ligand is shown as sphere and coloured yellow. The ligand-binding site residues are shown as sticks and coloured red

Table 7.5, below shows the templates associated with the target with the templates identified as having biologically relevant ligands by FunFOLD3 highlighted in yellow. The role/function of each protein and ligand as per the PDB entry is also depicted.

Table 7.5. Template list for nsp4 (C1902)
Templates for nsp4 (C1902) with the role/function and ligand as per PBD ID. Templates identified by FunFOLD3 to be of biological relevance are highlighted in yellow

Template	Role/function	Ligand
1dgjA	Oxidoreductase	FES, 2MO-MCN
1f86A	Transport protein	T44
1jdhA	Transcription	N/A
1z7gA	Transferase	N/A
2a65A	Transport protein	BOG, LEU, NA, CL
2js3A	Structural genomics, unknown function	N/A
2k9jB	Membrane protein	N/A
2ketA	Antibiotic	N/A
2l1qA	Antimicrobial protein	N/A
2p22D	Transport protein	SO4
2rccA	Oxidoreducase	PG4, PEG, EDO, PGE, GOL and ZN
2zy9A	Membrane protein, metal transport	MG
3a6pA	Protein transport/nuclear protein/RNA	DNA/RNA (RA, RC, RU, RG)
3gzfA	Viral protein	T44
3pcvA	Lyase	GSH
3tshA	Allerfen, oxidoreductase	FDA
3vc8A	Viral protein	N/A
4al0A	Isomerase	GSH
4bpmA	Isomerase	GSH, LVJ
4bx8A	Protein transport	CL
4im7A	Oxidoreducase	NAI, CS2
4mlbA	Transport protein	CXE, CL
4or2A2	Signaling protein	CLR, FM9
4wrtC	Transferase	DNA/RNA (RA, RC, RU, RG)
5k47A and 5k47A1	Transport protein	NAG, NAG-NAG
5mkeA	Transport protein	CHS
5uz7R	Signaling protein	N/A
5w3sA	Transport protein	Y01
5z1wA and 5z1wA1	Membrane protein	NAG
6an7D	Transport protein	N/A
6d6tB	Transport protein	Y01
6e20A	Sugar binding protein	NDG
6ftg51	No entry on PDBsum	-
6gcs5	Oxidoreducase	SF4,FES, FMN, NDP, ZMP, CDL, 3PE and ZN
6jyjA	Gene regulation	FLC
6k7kA	Membrane protein	Y01, ADP, ALF
6lnwA and 6lnwB	Transport protein	N/A
6n29A	Blood clotting	CA
6qp1A	Lyase	LCS
6rl5A	Transferase	PLP
6slfA	Hydrolase	LJ8, GOL, ACT, BTB, ZN
6snhX	Membrane protein	LMH
6uehA	Hydrolase	ACT, CA

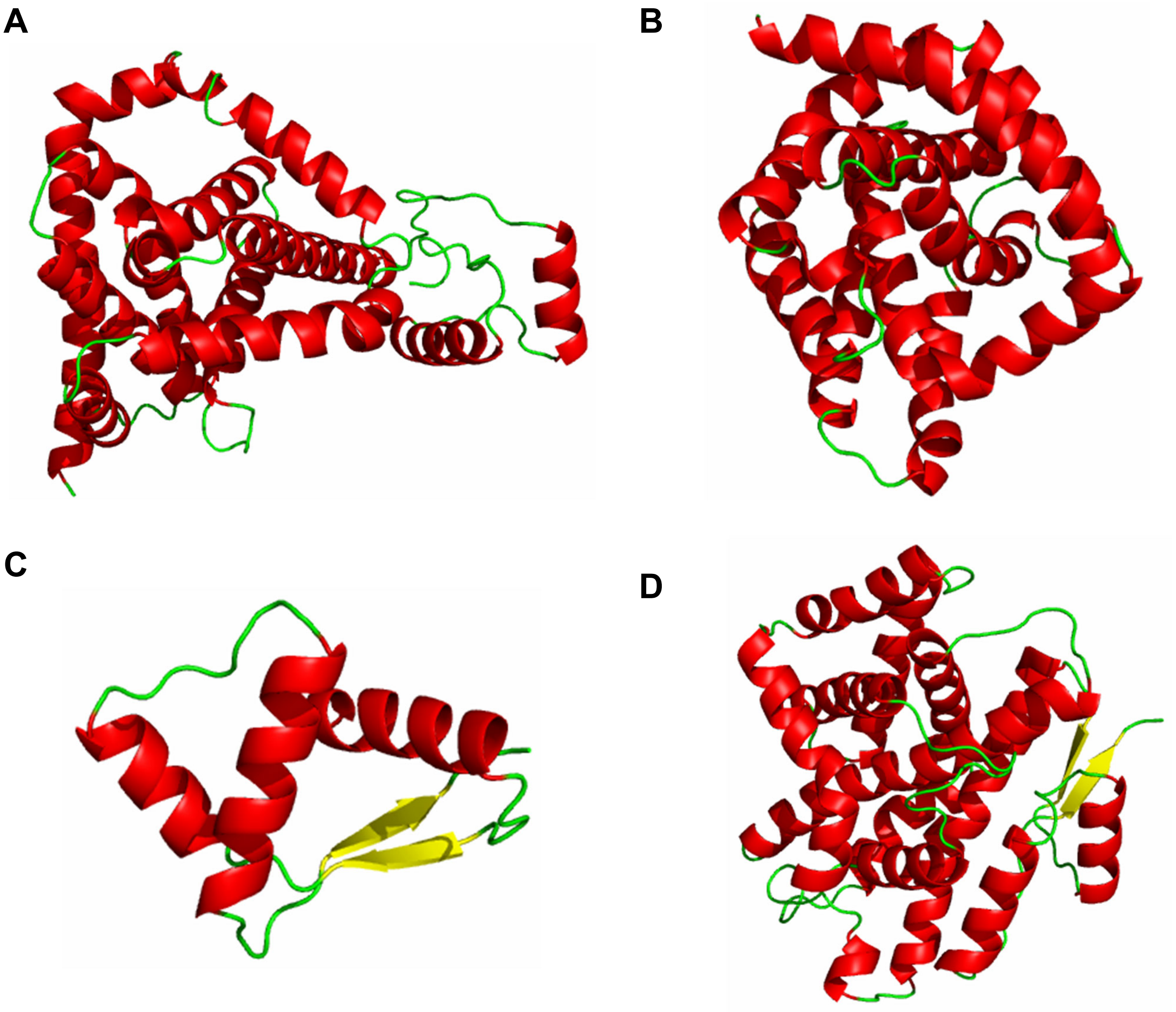


Figure 7.7. Top-scoring structure predictions from the McGuffin group for nsp6 (CASP Commons target C1903) full chain and individual domains
(A) Predicted structure for nsp6 (C1903) shown as cartoon and coloured by secondary structure with a residue length of 290 **(B)** Predicted structure for nsp6 (C1903d1) shown as cartoon and coloured by secondary structure with a residues 1-220 modelled **(C)** Predicted structure for nsp6 (C1903d2) shown as cartoon and coloured by secondary structure with residues 221-290 modelled **(D)** Predicted structure for nsp6 (C1903x2) shown as cartoon and coloured by secondary structure. This was the whole target of 290 residues following stage two modelling

The third predicted protein was nsp6 (CASP Commons targets C1903). Non-structural protein 6 can exist as a coronavirus non-structural protein 6. Autophagy, is activated by nsp6 and is a cellular response to starvation which generates autophagosomes to carry cellular organelles and long-lived proteins to lysosomes for degradation (Cottam *et al.*, 2011). Degradation through autophagy can provide an innate defence against virus infection, or autophagosomes can promote infection by facilitating assembly of replicase proteins (Cottam *et al.*, 2011). SARS-CoV can activate autophagy, an example is avian coronavirus infectious bronchitis virus (IBV) (Cottam *et al.*, 2011). Nsp6 may alter adaptive immune responses by directing immunomodulatory proteins synthesised by the ER into autophagosomes for degradation (Cottam *et al.*, 2011).

There were four different variations of the structure for prediction and as can be seen in Figure 7.8, there were no ligands and ligand-binding sites predicted by FunFOLD3. For the rounds of modelling of the full structure C1903 (nsp6) had a global quality score of 0.5232 and 0.5160 for C1903x2 (nsp6). For specific residues of the C1903 target (nsp6), C1903d1 (nsp6) and C1903d2 (nsp6) global quality score of 0.4139 and 0.1225, respectively, based on the full length sequences.

In comparison, IntFOLD6 obtained predictions for C1903d1, C1903d2 and C1903x2 (all nsp6 proteins) domain sequences, and is depicted below in Figure 7.9. The global model quality score was 0.4626, 0.4608 and 0.4618, respectively. The predicted ligands were haem (C1903d1), (3beta,14beta,17alpha)-ergosta-5,24(28)-dien-3-ol (94R) (C1903d2) and [(2~{S})-2,3-bis(oxidanyl)propyl] (~{E})-undec-2-enoate (MUN), [(2~{S})-2,3-bis(oxidanyl)propyl] (~{E})-undec-2-enoate (MPG) and retinal (RET) (1903x2).

Table 7.6, shows the templates associated with the target nsp6 (C1903) with the templates identified as having ligands by FunFOLD3 highlighted in yellow. The role/function of each template as per the PDB entry is also depicted.

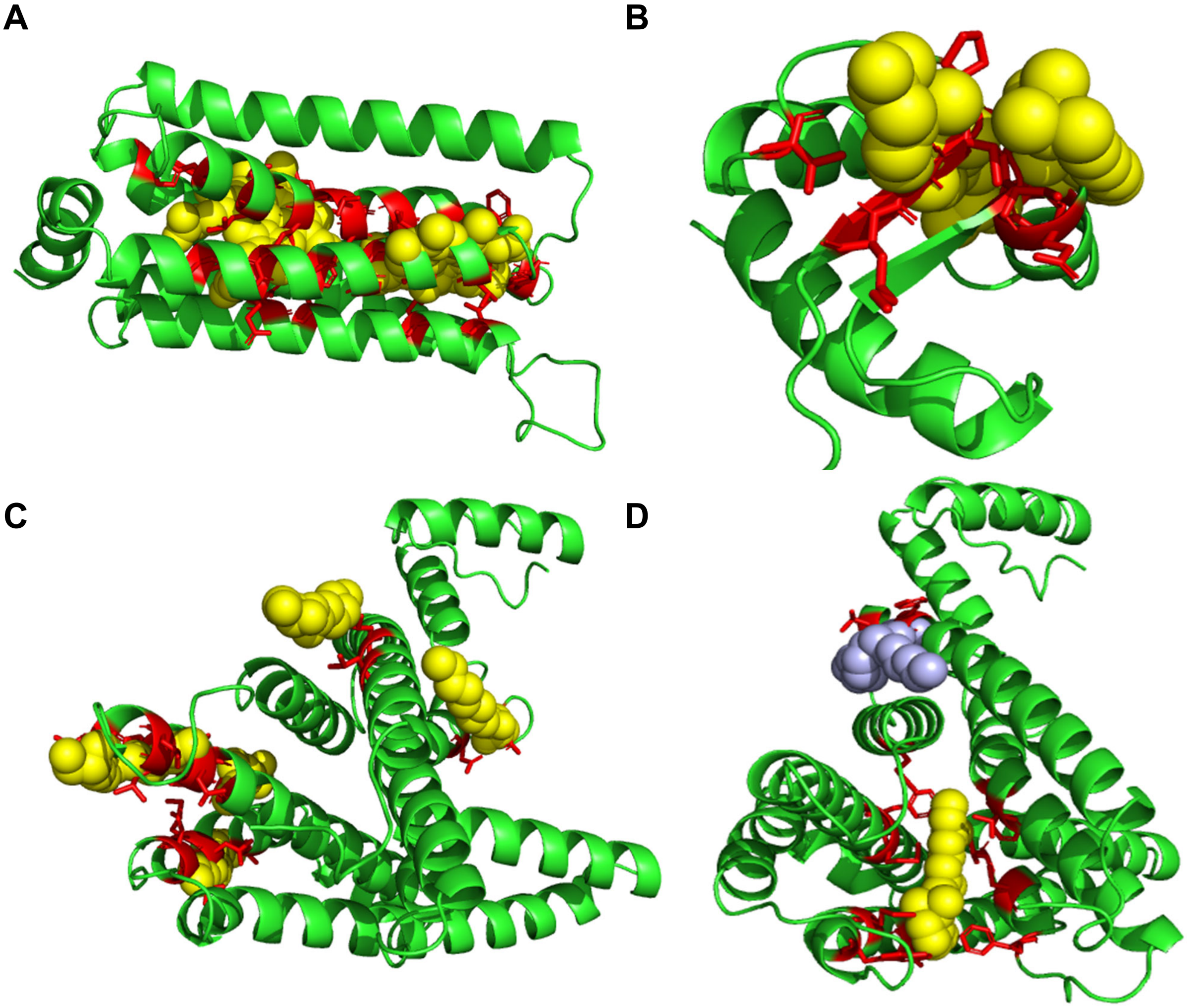


Figure 7.8. IntFOLD6 predictions for nsp6 (C1903) domains
(A) Predicted structure for nsp6 (C1903d1) shown as cartoon and coloured green with the predicted ligand HEM shown as sphere and coloured yellow. Ligand-binding sites are shown as sticks and coloured red **(B)** Predicted structure for nsp6 (C1903d2) shown as cartoon and coloured green with a residues with the predicted ligand 94R shown as sphere and coloured yellow. Ligand-binding sites are shown as sticks and coloured red **(C)** Predicted structure for nsp6 (C1903x2) shown as cartoon and coloured green with the predicted ligand MUN shown as sphere and coloured yellow, the MUN ligand was predicted in several locations in the protein. Ligand-binding sites are shown as sticks and coloured red **(D)** Predicted structure for nsp6 (C1903x2) shown as cartoon and coloured green with the predicted ligand MPG shown as sphere and coloured blue and the RET ligand shown as sphere and coloured yellow

Table 7.6. Template list for nsp6 (C1903)
Templates for nsp6 (C1903) with the role/function and ligand as per PBD ID. Templates identified by FunFOLD3 to be of biological relevance are highlighted in yellow

Template	Role/function	Ligand
1q06A	Transcription	AG
1r8dA	Transcription/DNA	DNA
1sp3A	Oxidoreductase	HEM
2ga1A	Unknown protein	GOL
2jihB1	Hydrolase	097
2l95A	Hydrolase	N/A
2ot3A	Protein transport	N/A
2r18A	Viral protein	N/A
3gp4A	Transcription regulator	MED
3rkoB, 3rkoC and 3rkoN	Oxidoreductase	CA7
3zuxa	Transport protein	NA, TCH
4c0oA	Transport protein/RNA binding protein	K
4jkvA	Membrane protein	1KS
4l6rA2	Membrane protein	
4mlqA	Transferase	DPM
4qi1A	Membrane protein	MPG, RET
4w6vA2	Transport protein	PEG
4yubA	Ligase	N/A
5ctgA	Transport protein	PE5, BNG, TRS
5dn6J	Hydrolase	ATP, ADP, MG
5ee7A1	Signalling protein	5MV
5hvdA	Transport protein	2CV
5i5fA	Membrane protein	N/A
5tqqA1	Transport protein	N/A
5xpdA1	Transport protein	DCM
5y6pgy	No entry	-
5y78A	Transport protein	P04
6bmlA	Transferase	ZN
6bmsA	Membrane protein	ZN
6csmA	Membrane protein	RET
6erdA	Transferase	GOL, CA
6eyuA	Membrane protein	RET, MUN
6gcs5	Oxidoreductase	SF4, FES, FMN, NDP, ZMP, CDL, 3PE, ZN
6gyhA	Proton transport	RET, CLR
6humD	Proton transport	BCR, LMG, SF4
6i1rA	Membrane protein	C5P
6k2ca	Oncoprotein	N/A
6n1zA	Transport protein	N/A
6nwfa	Membrane protein	RET, C14, BGC, D10, CL
6ob6A1	Transport protein	NBM
6pb1P	Signalling protein	CLR
6pw4A	Transport protein	PIO, CPL, PIK
6qv6B	Membrane protein	N/A
6rj8a	Hydrolase	PG4, PEG, EDO
6rl5A	Transferase	PLP
6tdxO	Membrane protein	N/A
6ukjA	Membrane protein	Y01
6vloa	Oxidoreductase	TYD, NAD, NI
6w08a	Toxin	ACY, EDO, FMT, K, CL

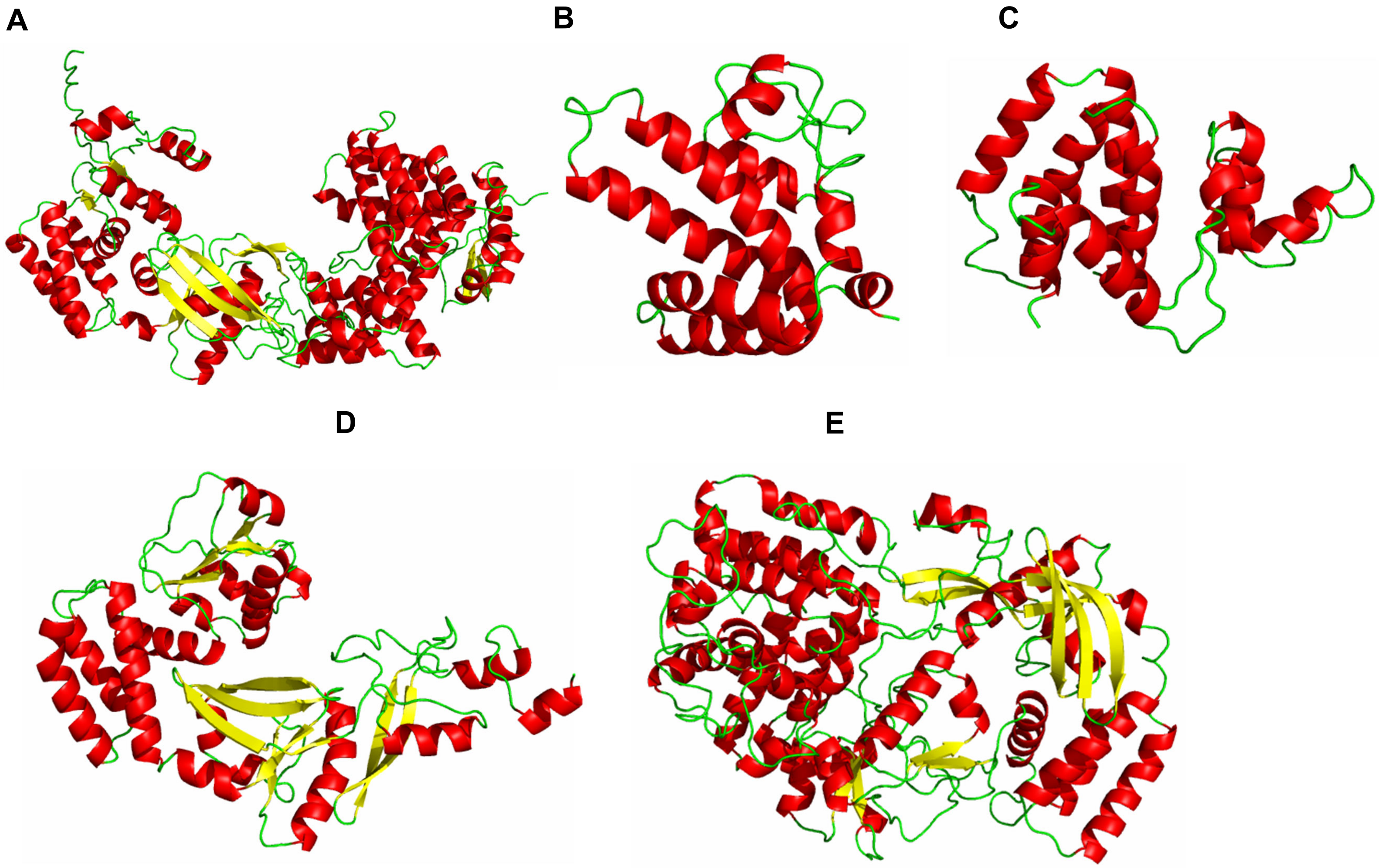


Figure 7.9. Top-scoring structure predictions from the McGuffin group for PL-PRO (CASP Commons target C1904) full chain and individual domains
(A) Predicted structure for PL-PRO (C1904) shown as cartoon and coloured by secondary structure with a residue length of 686 **(B)** Predicted structure for PL-PRO (C1904d1) shown as cartoon and coloured by secondary structure with a residues 1-151 modelled **(C)** Predicted structure for PL-PRO (C1904d2) shown as cartoon and coloured by secondary structure with residues 152-317 modelled **(D)** Predicted structure for PL-PRO (C1904d3) shown as cartoon and coloured by secondary structure with residues 318-686 modelled **(E)** Predicted structure for PL-PRO (C1904x2) shown as cartoon and coloured by secondary structure. This was the whole target of 686 residues following stage two modelling

The fourth predicted protein was PL-PRO (CASP Commons target was C1904), papain-like protease. The coronaviral proteases, papain-like protease and 3CL-like protease (3CLpro), are attractive antiviral drug targets because they are essential for coronaviral replication (Báez-Santos, St John and Mesecar, 2015). The primary function of PLpro and 3CLpro are to process the viral polyprotein in a coordinated manner (Báez-Santos, St John and Mesecar, 2015). PLpro has the additional function of stripping ubiquitin and interferon-induced gene 15 from host-cell proteins, to aid coronaviruses in their evasion of the host innate immune responses (Báez-Santos, St John and Mesecar, 2015). Thus, targeting PLpro with antiviral drugs may have an advantage in not only inhibiting viral replication, but also inhibiting the dysregulation of signalling cascades in infected cells that may lead to cell death in surrounding uninfected cells (Báez-Santos, St John and Mesecar, 2015). SARS-CoV PLpro antagonistic activities have been shown to block the production of important cytokines involved in the activation of the host's innate immune response such as CXCL10 and CCL5 (Báez-Santos, St John and Mesecar, 2015).

There were five variations of the structure available for prediction as shown in Figure 7.9, and there were no ligands and ligand-binding sites predicted by FunFOLD3. Based on the full length reference sequences, the global quality score was 0.4853 for C1904, 0.3278 for C1904d1, 0.1202 for C1904d2, 0.3038 for C1904d3 and 0.4842 for C1904x2 (all PL-PRO proteins).

IntFOLD6 predicted ligands and ligand-binding site residues for C1904d1, C1904d2, C1904d3 and C1904x2 (all PL-PRO proteins) and results are shown in Figure 7.10. The global quality score was 0.4332, 0.4507, 0.3969 and 0.3717, respectively.

Table 7.7 shows the templates associated with the protein PL-PRO (C1904) with the templates identified as having ligands by FunFOL3 highlighted in yellow. The role/function of each protein is also provided.

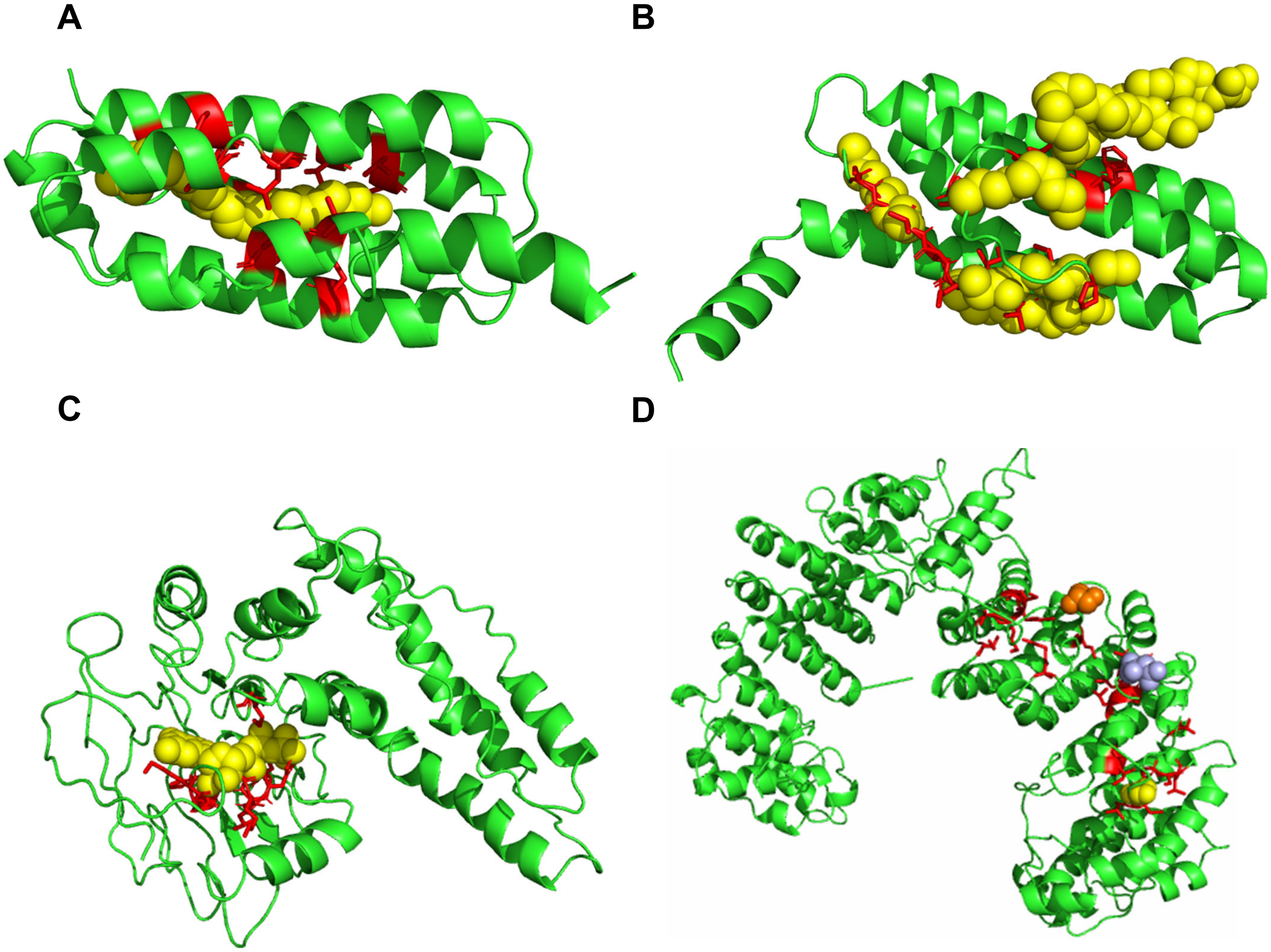


Figure 7.10. IntFOLD6 prediction for PL-PRO (C1904)
(A) Predicted structure for PL-PRO (C1904d1) shown as cartoon and coloured green with the predicted ligand PLM shown as sphere and coloured yellow. Ligand-binding sites are shown as sticks and coloured red **(B)** Predicted structure for PL-PRO (C1904d2) shown as cartoon and coloured green with a residues with the predicted ligand MPG shown as sphere and coloured yellow, the ligand has been predicted in three different locations. Ligand-binding sites are shown as sticks and coloured red **(C)** Predicted structure for PL-PRO (C1904d3) shown as cartoon and coloured green with the predicted ligand GDP shown as sphere and coloured yellow. Ligand-binding sites are shown as sticks and coloured red **(D)** Predicted structure for PL-PRO (C1904x2) shown as cartoon and coloured green with the predicted ligands ALA shown as sphere and coloured blue, SER ligand shown as sphere and coloured orange and the ASN ligand shown as sphere and coloured blue

Table 7.7. Template list for PL-PRO (C1904)
Templates for PL-PRO (C1904) with the role/function and ligand as per PBD ID. Templates identified by FunFOLD3 to be of biological relevance are highlighted in yellow

Template	Role/function	Ligand
1b3ua/1b3uA	Scaffold protein	N/A
1bm8A	Cell cycle	N/A
1d0qA	Transferase	ZN
1ek0A	Endocytosis/exocytosis	GNP, GDP, MG
1m7bA	Signalling protein	GTP, MG
1z1wA	Hydrolase	ZN
1z2aA	Protein transport	GDP
2atxA	Hydrolase	GNP,MG
2bptA	Nuclear transport	Random…
2fg5A	Signalling protein	GNP, ZN
2fgeA	Hydrolase/plant protein	MG, GNP
3ga8A	DNA binding protein	ZN
3jd8A	Membrane protein	CLR
3kkqA	Signalling protein	GDP
3opeA	Transferase	ZN, SAM
3qf4A	Transport protein	ANP, MG
3t5gA	Signalling protein, lipid binding protein	GDP, FAR
3uonA	Signalling protein/antagonist	QNB, BGC
3wajA	Transferase	ZN, SO4
4iuwA	Hydrolase	ZN, CO3
4kxfB	Immune system	ADP
4r5xA	Hydrolase/hydrolase inhibitor	ZN, R5X
4rnbA	Signalling protein	SUV
4ui9N/4ui90	Cell cycle	Random…
5ch1B	Transferase	ZN, SAH
5ee7A	Signalling protein	5MV
5fd3A	Transcription/DNA	ZN and DNA
5l22B	Protein transport	ADP, MG
5mscA	Oxidoreductase	AMP
5msoA	Oxidoreductase	NAP
5mspA	Oxidoreductase	NAP
6co7A3	Membrane protein	CLR, CA
6du7A	Oxidoreductase	FAD
6e3yR	Signalling protein	Random…
6itcA	Protein transport	BEF, ADP
6me2A	Membrane protein	JEV
6n29A	Blood clotting	CA
6nq0A	Transport protein	EUJ
6oh2A	Transport protein	C5P
6qp6A2	Membrane protein	CA
6rflK	Viral protein	RC, RA, RG, RU
6rl5a	Transferase	PLP
6snhx	Membrane protein	LMH
6ueha	Hydrolase	ACT, CA
6v6bC	Structural protein	N/A
6vbu0	Protein transport	CA

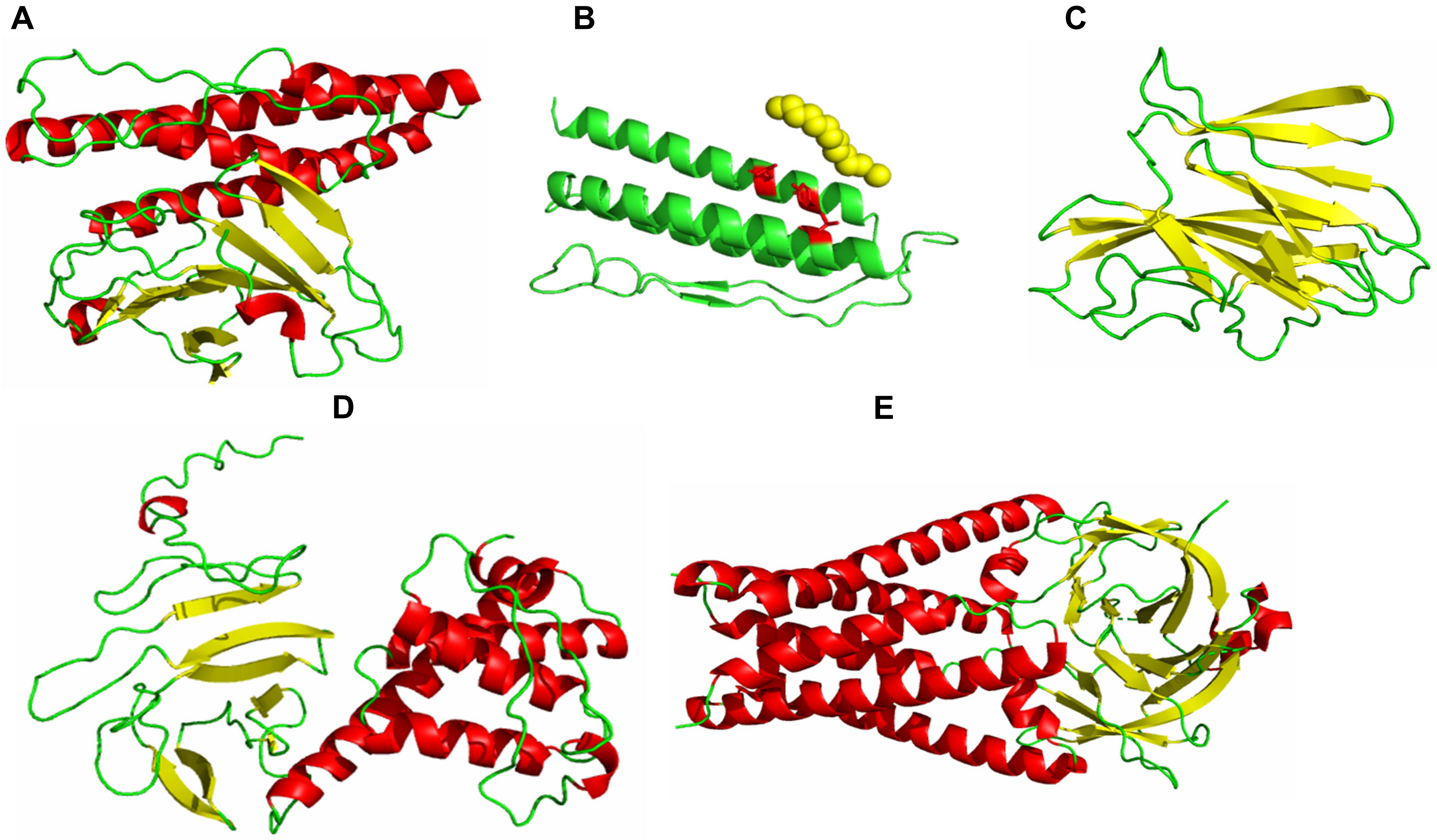


Figure 7.11. Top-scoring structure predictions from the McGuffin group for protein ORF3a (CASP Commons target C1905) full chain and individual domains
(A) Predicted structure for ORF3a (C1905) shown as cartoon and coloured by secondary structure with a residue length of 275 **(B)** Predicted structure for ORF3a (C1905d1) shown as cartoon and coloured green with a residues 1-130 modelled the predicted ligand MPG shown as sphere and coloured yellow with ligand-binding site residues shown as sticks and coloured red **(C)** Predicted structure for ORF3a (C1905d2) shown as cartoon and coloured by secondary structure with residues 131-275 modelled **(D)** Predicted structure for ORF3a (C1905x2) shown as cartoon and coloured by secondary structure. This was the whole target of 686 residues following stage two modelling **(E)** PDB structure (PDB ID 6xdc) for ORF3a (C1905) which is a dimer and the chains are coloured by secondary structure

The fifth predicted protein was ORF3a, open reading frame 3a (CASP Commons target C1905) . ORF3a is found in other coronaviruses along with ORF7b (C1910). Severe acute respiratory syndrome coronavirus (SARS-CoV) open reading frame 3a (ORF3a) accessory protein activates the NLRP3 inflammasome by promoting TNF receptor-associated factor 3 (TRAF3)–mediated ubiquitination of apoptosis-associated speck-like protein containing a caspase recruitment domain (ASC). ORF3a and three ion channel proteins; E are collectively required for viral replication and virulence (Siu *et al.*, 2019). SARS-CoV is capable of inducing a storm of proinflammatory cytokines (Siu *et al.*, 2019). A robust elevation of IL-1ß was seen during early infection of SARS-CoV and IL-1ß is a key proinflammatory cytokine (Siu *et al.*, 2019). Proinflammatory cytokines are double-edged swords that mobilise host defence and also drive pathologic inflammation (Siu *et al.*, 2019). Inflammation has both antiviral and proviral roles, on one hand it is part of the innate antiviral response that restricts viral replication and infection and conversely it facilitates viral dissemination by releasing a large number of virions (Siu *et al.*, 2019).

The UniProtKB entry for ORF3a states the function as, forms homotertrameric potassium sensitive ion channels (viroporin) and may modulate virus release (UniProt Consortium, 2019). Up-regulates expression of fibrinogen subunits FGA, FGB and FGG in host lung epithelial cells. Induces apoptosis in cell culture. Downregulates the type I interferon receptor by inducing serine phosphorylation within the IFN alpha-receptor subunit 1(IFNAR1) degradation motif and increasing IFNAR1 ubiquitination (UniProt Consortium, 2019). In terms of GO terms associated with ORF3a ion channel activity (GO:0005216) and pore formation by virus in membrane of host cell (GO:0039707)(UniProt Consortium, 2019). In comparison, FunFOLD3 predicted ion transport (GO:0006811) as the closest to ion channel activity. In terms of cellular component, integral component of membrane (GO:0016021) is associated with the UniProtKB entry and was also predicted by FunFOLD3.

There were four different variations of the structure available for prediction as shown in Figure 7.11 with one having a ligand and ligand-binding site residues predicted (Figure 7.11B). According to the full length reference sequence, the global quality model scores were 0.4392, 0.2252, 0.2311 and 0.4151 which relates to C1905, C1905d1, C1905d2 and C1905x2 (all ORF3a proteins), respectively.

This target was the first target of all the CASP Commons targets to not only have a predicted ligand by FunFOLD3 but also have an actual structure PDB associated. No ligand was observed in the actual structure and the predicted GO terms by FunFOLD3 have been outlined in Table 7.8 below. Additionally, TM scores and TM-align structures of the predicted structures by ModFOLD has been compared against the actual structure from PDB.

Table 7.8. Predicted GO terms for CASP Commons target ORF3a (C1905d1)
The GO terms for CASP Commons target ORF3a (C1905d1) and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component coloured purple

GO term	GO term domain	Function
GO: 0004129	Molecular function	cytochrome-c oxidase activity
GO:0016491	Molecular function	oxidoreductase activity
GO:0005506	Molecular function	iron ion binding
GO:0009055	Molecular function	electron transfer activity
GO:0020037	Molecular function	haeme binding
GO:0046872	Molecular function	metal ion binding
GO:0048039	Molecular function	ubiquinone binding
GO:0005507	Molecular function	copper ion binding
GO:0006811	Biological process	ion transport
GO:0022900	Biological process	electron transport chain
GO:0055114	Biological process	oxidation-reduction process
GO:0006119	Biological process	oxidative phosphorylation
GO:0009060	Biological process	aerobic respiration
GO:0006099	Biological process	tricarboxylic acid cycle
GO:0006121	Biological process	mitochondrial electron transport, succinate to ubiquinone
GO:0022904	Biological process	respiratory electron transport chain
GO:0005886	Cellular Component	plasma membrane
GO:0016020	Cellular Component	membrane
GO:0016021	Cellular Component	integral component of membrane
GO:0070469	Cellular Component	respirasome
GO:0005740	Cellular Component	mitochondrial envelope
GO:0005739	Cellular Component	mitochondrion
GO:0005743	Cellular Component	mitochondrial inner membrane
GO:0005749	Cellular Component	mitochondrial respiratory chain complex II, succinate dehydrogenase complex (ubiquinone)

Figure 7.12 below, shows the TMalign structural alignments for the predicted ORF3a protein targets; C1905 (A) and C1905x2 (B) and the observed structure from the PDB entry. A TM-score of 0.26105 and 0.30148 was achieved, respectively demonstrating poor structural alignment over the full chain structure. However, there was an improvement between stage one and stage two modelling. The domains achieved a TMscore of 0.26238 for domain 1 (C) and 0.22785 for domain 2 (D).

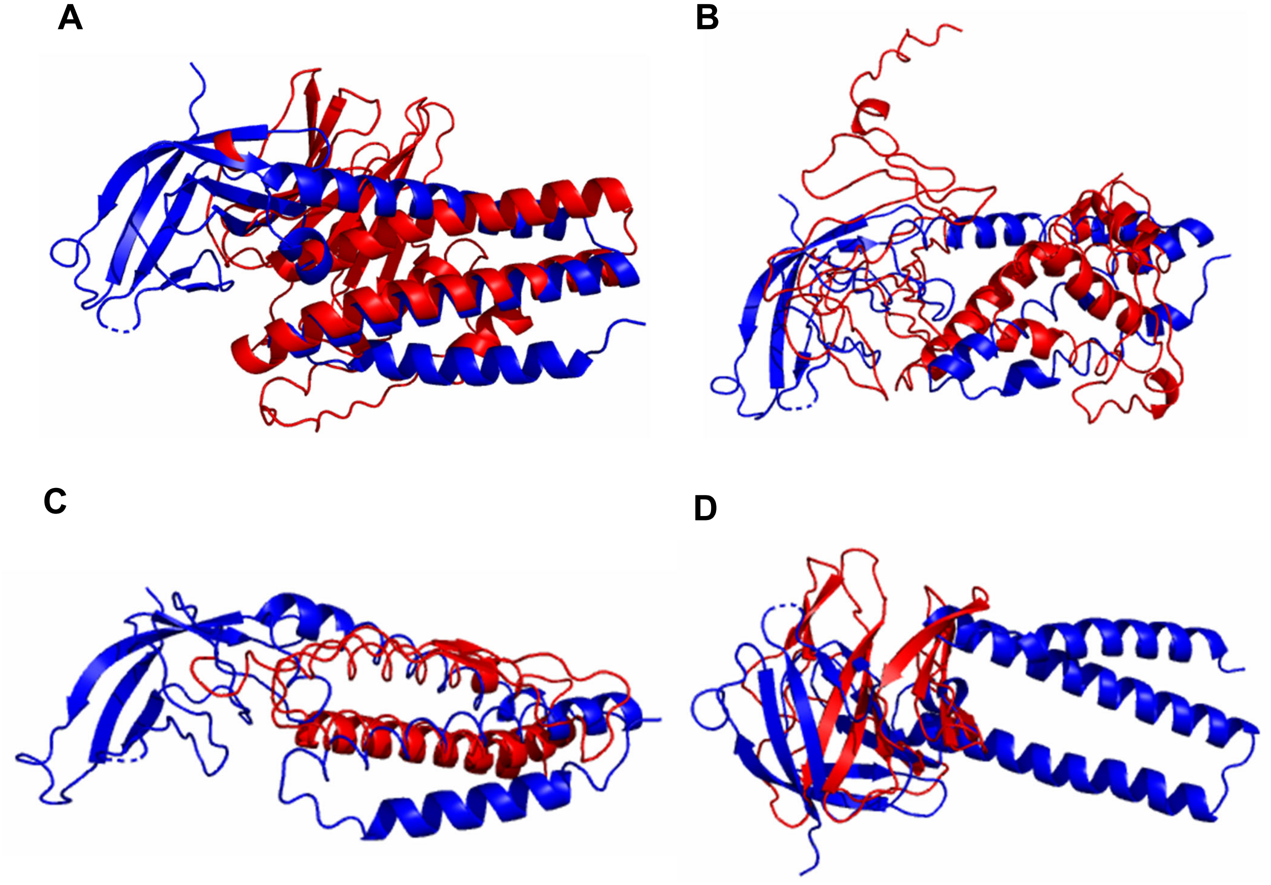


Figure 7.12. Comparison of TMalign superposition for CASP Commons target C1905 (ORF3a) and SARS-CoV-2 ORF3a (PDB ID 6xdc)
(A) The structure in blue is the structure of 6xdc and the predicted structure for ORF3a (CASP_Commons target C1905) is shown in red. A TM-align score of 0.26105 was achieved for the protein structures. This was normalised for SARS-CoV-2 ORF3a (PDB ID 6xdc) as it is the reference molecule **(B)** The structure in blue is the structure of 6xdc and the predicted structure for ORF3a (CASP_Commons target C1905x2) is shown in red. A TM-align score of 0.30148 was achieved for the protein structures **(C)** The structure in blue is the structure of 6xdc and the predicted structure for ORF3a (CASP_Commons target C1905d1) is shown in red. A TM-align score of 0.26238 was achieved for the protein structures. **(D)** The structure in blue is the structure of 6xdc and the predicted structure for ORF3a (CASP_Commons target C1905x2) is shown in red. A TM-align score of 0.22785 was achieved for the protein structures

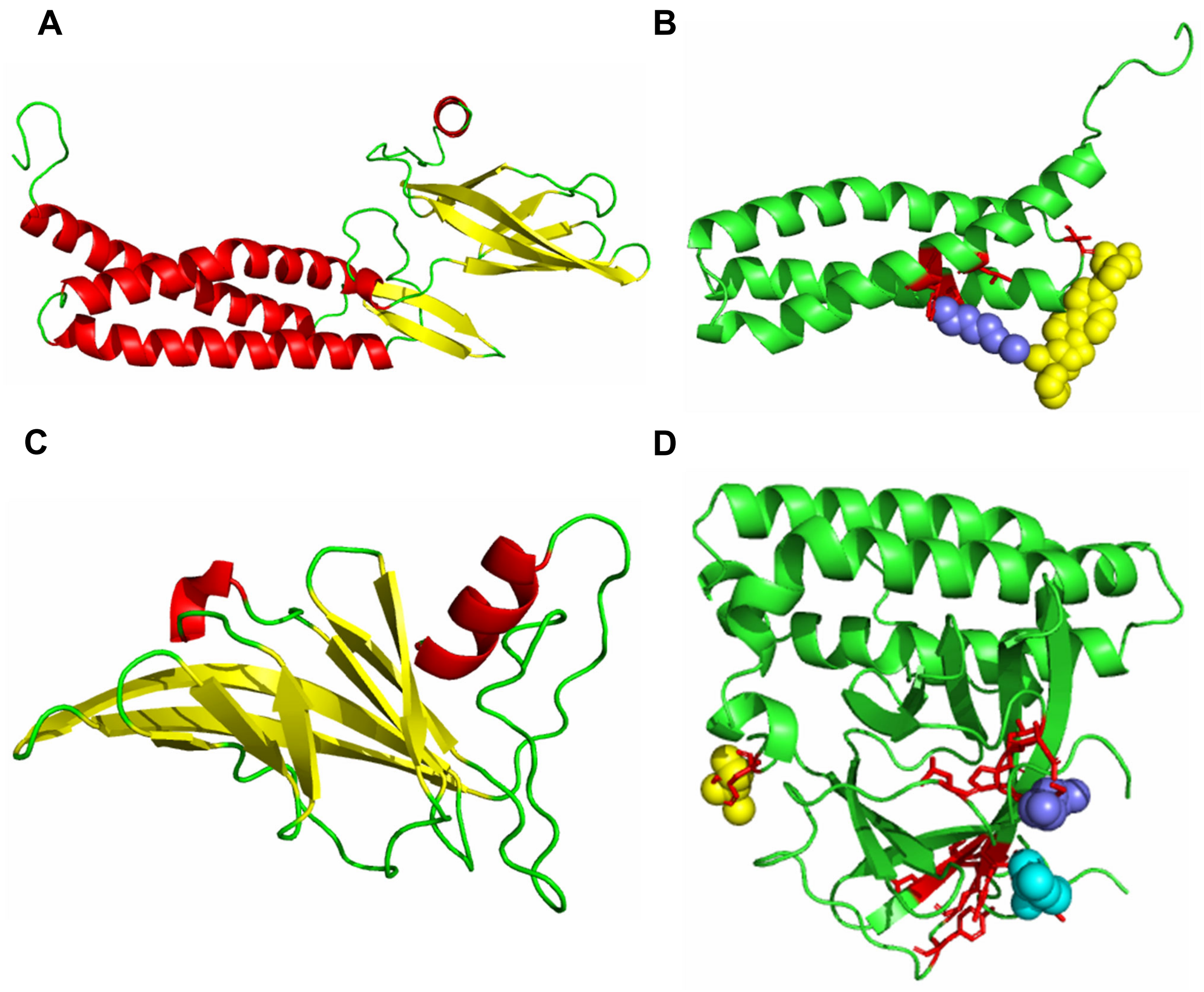


Figure 7.13. Top-scoring structure predictions from the McGuffin group for transmembrane protein (CASP Commons target C1906) full chain and individual domains
(A) Predicted structure for transmembrane protein (C1906) shown as cartoon and coloured by secondary structure with a residue length of 222 **(B)** Predicted structure for transmembrane protein (C1906d1) shown as cartoon and coloured green with a residues 1-103 modelled the predicted ligand DUO (duocarmycin) shown as sphere and coloured yellow and the ligand MPG shown as sphere an coloured blue. Ligand-binding site residues shown as sticks and coloured red **(C)** Predicted structure for transmembrane (C1906d2) shown as cartoon and coloured by secondary structure with residues 104-222 modelled **(D)** Predicted structure for transmembrane (C1906x2) shown as cartoon and coloured green. This was the whole target of 222 residues following stage two modelling the GOL ligand is shown as sphere and coloured yellow, the ASN ligand shown as sphere and coloured blue and the LEU ligand shown as sphere and coloured cyan. Ligand-binding site residues are shown as sticks and coloured red

The sixth predicted protein was transmembrane protein (CASP Commons target was C1906). On a whole, “membrane protein” is quite a broad description for a protein however, when specifically related to SARS-CoV-2, membrane protein is a major structural protein as mentioned in the introduction (Tseng *et al.*, 2010). The most abundant structural protein of SARS-CoV-2 is the M glycoprotein (Thomas, 2020) and spans the membrane bilayer, leaving a short NH2-terminal domain outside the virus and a long COOH terminus (cytoplasmic domain) inside the virion (Thomas, 2020). The M protein can bind to all other structural proteins and cooperates with the S protein (Thomas, 2020) despite this, the role of M protein is not fully understood (Thomas, 2020).

The UniProtKB entry on membrane protein states the function, as a component of the viral envelope that plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins (UniProt Consortium, 2019). Additionally, GO terms related to structural constituent of virion (GO:0039660) and mitigation of host immune response by virus (GO:0030683) were predicted by FunFOLD3. In terms of cellular components, the following GO terms were associated; host cell Golgi membrane (GO:0044178), integral component of membrane (GO:0016021), viral envelope (GO:0019031) and virion membrane (GO:0055036)(UniProt Consortium, 2019).

Membrane protein is defined as a homomultimer and (1) interacts with envelope E protein in the budding compartment of the host cell, which is located between endoplasmic reticulum and the Golgi complex. (2) Forms a complex with HE (hemagglutinin-esterase) and S proteins (UniProt Consortium, 2019), (3) interacts with nucleocapsid N protein. This interaction probably participates in RNA packaging into the virus and (4) interacts with the accessory proteins 3a and 7a (UniProt Consortium, 2019).

Two ligands were predicted for C1906d1; duocarmycin and MPG. Duocarmycins are a group of antineoplastic agents with low picomolar potency (Guerrero *et al.*, 2020). They are thought to act by binding and alkylating double-stranded DNA in AT-rich regions of the minor groove (Guerrero *et al.*, 2020). Octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate or MPG is classed as experimental and belongs to the class of organic compounds known as fatty alcohol esters (Wishart *et al.*, 2018). Three ligands were predicted for C1906x2, compared to no ligands predicted in the first round with C1906 (all transmembrane proteins). However, the ligands were ASN and LEU which are amino acids and glycerol (GOL) which despite being a crystallisation additive can also bind to proteins as a substrate (Yamanishi *et al.*, 2012).

According to the full length reference sequence, the global quality model scores were 0.4172, 0.2532, 0.2966 and 0.5149 for targets C1906, C1906d1, C1906d2 and C1906x2, respectively (all transmembrane proteins).

To provide further insight into the function of the protein the GO terms associated are listed below in Table 7.9, no GO terms were predicted for C1906d1. Due to C1906 (transmembrane protein) having predicted ligands, no template list has been provided.

Table 7.9. Predicted GO terms for transmembrane protein (CASP Commons target C1906x2)
The GO terms for CASP Commons target C1906x2 (transmembrane protein) and their associated term domains and function are shown below. Molecular function coloured green and biological process coloured red

GO term	GO term domain	Function
GO: 0003824	Molecular function	catalytic activity
GO:00088152	Biological process	metabolic process

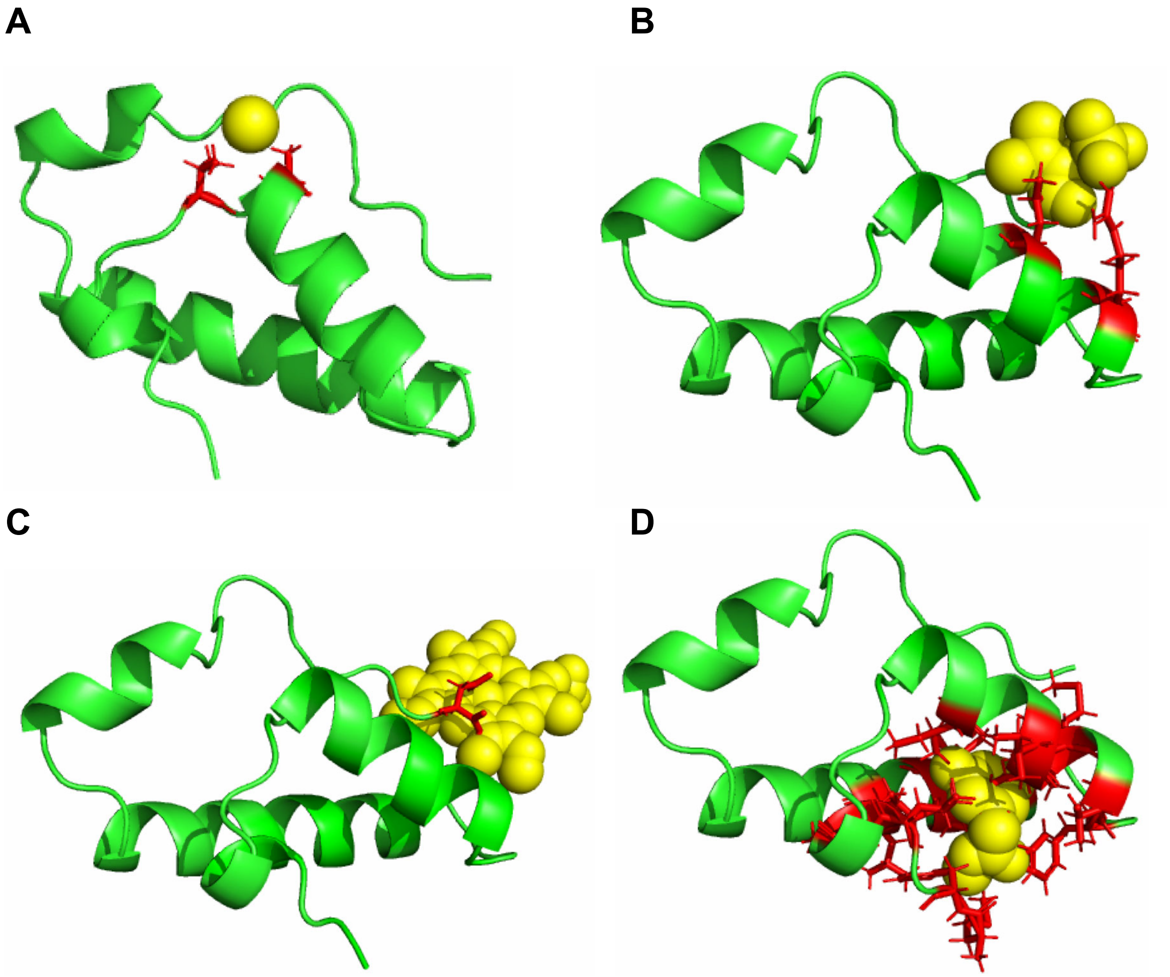


Figure 7.14. Top-scoring structure predictions from the McGuffin group for protein ORF6 (CASP Commons target C1907) **(A)** Predicted structure for ORF6 (C1907) shown as cartoon and coloured green with a residue length of 61 with the predicted ligand calcium shown as sphere and coloured yellow. Ligand-binding site residues shown as sticks and coloured red **(B)** Predicted structure for ORF6 (C1907) shown as cartoon and coloured green with a residue length of 61 with the predicted ligand 3,3-Dichloro-2-phosphonomethyl-acrylic acid (DCO) shown as sphere and coloured yellow. Ligand-binding site residues shown as sticks and coloured red **(C)** Predicted structure for ORF6 (C1907) shown as cartoon and coloured green with a residue length of 61 with the predicted ligand haemoglobin shown as sphere and coloured yellow. Ligand-binding site residues shown as sticks and coloured red **(D)** Predicted structure for ORF6 (C1907) shown as cartoon and coloured green with a residue length of 61 with the predicted ligand methyl 2-(2,5-dihydroxyphenyl)acetate (XQI) shown as sphere and coloured yellow. Ligand-binding site residues shown as sticks and coloured red

The seventh predicted protein was ORF6, open reading frame 6 (CASP Commons target was C1907),. ORF6 protein is one of the eight accessory proteins of SARS-CoV (Gunalan, Mirazimi and Tan, 2011). Furthermore, it has been suggested that accessory genes have subtle effects on SARS-CoV replication that may be more important for viral replication or pathogenesis *in vivo* (Gunalan, Mirazimi and Tan, 2011). ORF6, encodes for a ~7kDa protein with a hydrophobic N-terminal that has been suggested to have a N-endo-C-endo conformation (Gunalan, Mirazimi and Tan, 2011). Additionally, ORF6 has been shown to interact with nsp8 protein from the SARS replicase complex (Kumar *et al.*, 2007), able to increase infection titre during early infection at low multiplicity of infection (Zhao et al., 2009), increase the rate of cellular gene synthesis (Geng *et al.*, 2005), inhibit interferon production (Kopecky-Bromberg *et al.*, 2007) and inhibit the nuclear translocation of STAT1 by interacting with karyopherin alpha2 (Frieman *et al.*, 2007). Most recently, the ORF6 protein has been suggested to induce intracellular membrane rearrangements resulting in a vesicular population in the infected cell which should potentially serve some role in increasing replication (Zhou *et al.*, 2010).

The UniProtKB entry for ORF6, identifies the gene as a major DNA-binding protein (UniProt Consortium, 2019) and outlines the function as several crucial roles in viral infection (UniProt Consortium, 2019). Participating in the opening of the viral DNA origin to initiate replication by interacting with the origin-binding protein. May disrupt loops, hairpins and other secondary structures present on ssDNA to reduce and eliminate pausing of viral DNA polymerase at specific sites during elongation. Promotes viral DNA recombination by performing strand-transfer (UniProt Consortium, 2019).

In terms of ligand prediction, calcium is essential for virus entry, viral gene replication, virion maturation and release (Chen, Cao and Zhong, 2019). The alteration of host cells Ca^{2+} homeostasis is one of the strategies that viruses use to modulate host cell signal transduction mechanisms in their favour by to achieve successful replication via multiple routes; for instance, viral proteins directly bind to Ca^{2+} to disturb the membrane permeability for Ca^{2+} by manipulating Ca^{2+} apparatus (Chen, Cao and Zhong, 2019). With respect to haemoglobin; envelope protein, nucleocapsid phosphoprotein and ORF3a all have haem linked sites, with Arg134 of ORF3a, Cys44 of envelope protein and Ile304 of nucleocapsid phosphoprotein have been identified as the haem-iron sites, respectively (Liu, 2020a). Of which, none were predicted by FunFOLD3 as ligand-binding site residues. No information could be found on 3,3-Dichloro-2-phosphonomethyl-acrylic acid or methyl 3,5-dihydroxyphenylacetate with respect to SARS-CoV-2 or similar viruses.

The global quality model score for the structure was 0.5865.

To provide further insight into the function of the protein the GO terms associated are listed below in Table 7.10.

Table 7.10. Predicted GO terms for ORF6 (CASP_Commons target C1907)
The GO terms for CASP_Commons target C1907 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component coloured purple

GO term	GO term domain	Function
GO:0005524	Molecular function	ATP binding
GO:0051082	Molecular function	unfolded protein binding
GO:0003824	Molecular function	catalytic activity
GO:0008964	Molecular function	phosphoenolpyruvate carboxylase activity
GO:0016829	Molecular function	lyase activity
GO:0003677	Molecular function	DNA binding
GO:0043565	Molecular function	sequence-specific DNA binding
GO:0004497	Molecular function	monooxygenase activity
GO:0005496	Molecular function	steroid binding
GO:0005506	Molecular function	iron ion binding
GO:0008395	Molecular function	steroid hydroxylase activity
GO:0009055	Molecular function	electron transfer activity
GO:0016491	Molecular function	oxidoreductase activity
GO:0016705	Molecular function	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0016712	Molecular function	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor,

		and incorporation of one atom of oxygen
GO:0019825	Molecular function	oxygen binding
GO:0019899	Molecular function	enzyme binding
GO:0020037	Molecular function	haeme binding
GO:0030343	Molecular function	Vitamin D3 25-hydroxylase activity
GO:0033780	Molecular function	taurochenodeoxycholate 6alpha-hydroxylase activity
GO:0034875	Molecular function	caffeine oxidase activity
GO:0046872	Molecular function	metal ion binding
GO:0047638	Molecular function	albendazole monooxygenase activity
GO:0050591	Molecular function	quinine 3-monooxygenase activity
GO:0050649	Molecular function	testosterone 6-beta-hydroxylase activity
GO:0070576	Molecular function	vitamin D 24-hydroxylase activity
GO:0009822	Biological process	alkaloid catabolic process
GO:0006457	Biological process	protein folding
GO:0006950	Biological process	response to stress
GO:0006099	Biological process	tricarboxylic acid cycle
GO:0008152	Biological process	metabolic process
GO:0015977	Biological process	carbon fixation
GO:0006629	Biological process	lipid metabolic process
GO:0006706	Biological process	steroid catabolic process
GO:0008202	Biological process	steroid metabolic process
GO:0008209	Biological process	androgen metabolic process
GO:0006805	Biological process	xenobiotic metabolic process
GO:00016098	Biological process	monoterpenoid metabolic process
GO:0017144	Biological process	drug metabolic process
GO:0042737	Biological process	drug catabolic process
GO:0042738	Biological process	exogenous drug catabolic process
GO:0044281	Biological process	small molecule metabolic process
GO:0046483	Biological process	heterocycle metabolic process
GO:0055114	Biological process	oxidation-reduction process
GO:0070989	Biological process	oxidative demethylation
GO:0005737	Cellular Component	cytoplasm
GO:0005783	Cellular Component	endoplasmic reticulum
GO:0005789	Cellular Component	endoplasmic reticulum membrane
GO:0009986	Cellular Component	cell surface
GO:0016020	Cellular Component	membrane
GO:0016021	Cellular Component	integral component of membrane
GO:0031090	Cellular Component	organelle membrane
GO:0043231	Cellular Component	intracellular membrane-bounded organelle
GO:0005749	Cellular Component	mitochondrial respiratory chain complex II, succinate dehydrogenase complex (ubiquinone)

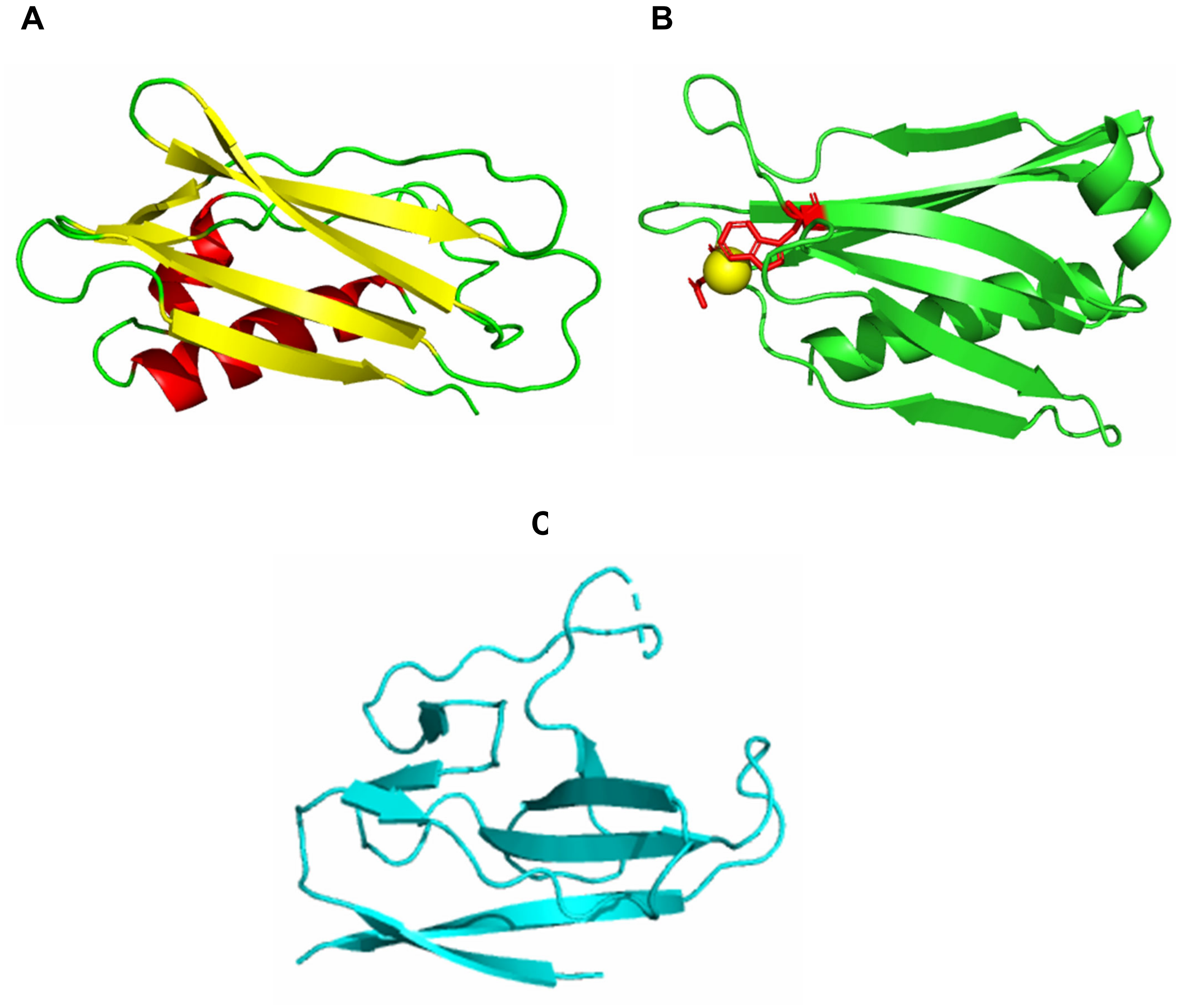


Figure 7.15. Top-scoring structure predictions from the McGuffin group for protein ORF8 (CASP Commons target C1908) **(A)** Predicted structure for ORF8 (C1908) shown as cartoon and coloured by secondary structure with a residue length of 121 **(B)** Predicted structure for ORF8 (C1908x2) shown as cartoon and coloured green with a residue length of 121 following stage 2 modelling and the CA ligand shown as sphere and coloured yellow with ligand-binding site residues shown as sticks and coloured red **(C)** PDB structure for ORF8 (PDB ID 7jtl) shown as cartoon and coloured cyan. Only domain A has been shown for comparison purposes

The eighth predicted protein was ORF8, opening reading frame 8 (CASP Commons target C1908),. As with ORF6, ORF8 is also one of the accessory proteins for SARs-CoV-2 and is one of the several transcribed non-structural proteins (Liu, 2020b). Spike proteins, ORF8a and ORF3a proteins are significantly different from other known SARS-like coronaviruses and may cause more serious pathogenicity and transmission differences than SARS-CoV (Liu, 2020b).

The function of ORF8 as per the UniProtKB entry is (1) may play a role in modulating host immune response and (2) may play a role in blocking host IL17 cytokine by its interaction with host IL17RA (Magrane & UniProt Consortium, 2011). Literature information available specifically for SARS-CoV-2 ORF8 shows it disrupts IFN-1 signalling when exogenously overexpressed in cells (Li *et al.*, 2020).

Predictions on this target were done on the same residue length after two rounds of modelling, with the second round predicting the calcium ligand. The global model score decreased from 0.4935 for C1908 to 0.4721 for C1908x2 (all ORF8 proteins), this could be explained by some groups who participated in round one not partaking in round two. Despite the decrease in the global model score the stage two model predicted a ligand, however no ligand was observed in the actual structure.

To provide further insight into the function of the protein, from a FunFOLD3, perspective the GO terms associated are listed below in Table 7.11.

Table 7.11. Predicted GO terms for ORF8 (CASP Commons target C1908)
The GO terms for ORF 8 (CASP_Commons target C1908) and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component coloured purple

GO term	GO term domain	Function
GO:0007155	Biological process	cell adhesion
GO:0006954	Biological process	inflammatory response
Go:0045087	Biological process	innate immune response
GO:0005604	Cellular Component	basement membrane
GO:0005615	Cellular Component	extracellular space
GO:0005576	Cellular Component	extracellular region

Figure 7.16 below, shows the TMalign structural alignments for the predicted C1908 (A) and C1908x2 (B) (all ORF8 proteins) and the observed structure from the PDB entry. A TM-score of 0.34318 and 0.38410 was achieved, respectively demonstrating poor structural homology..

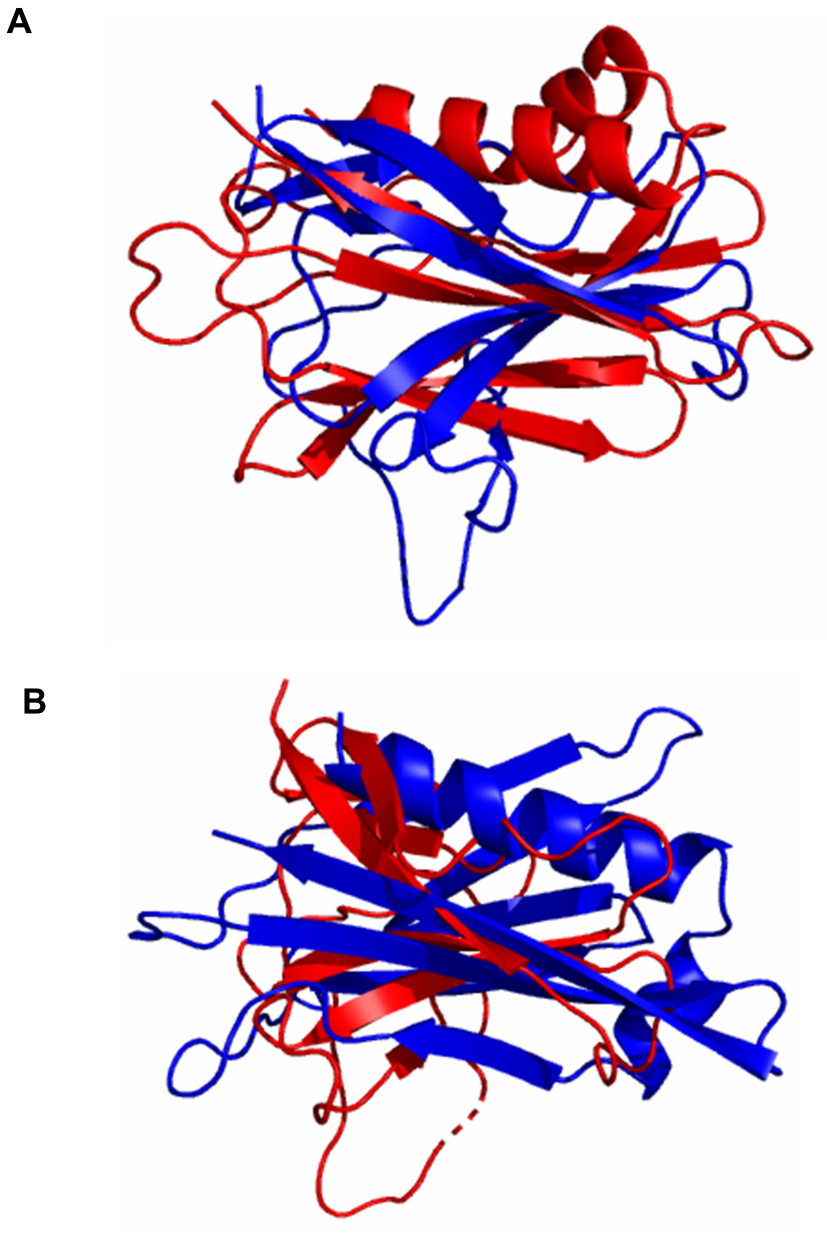


Figure 7.16. Comparison of TMalign superposition for predicted ORF8 (CASP Commons target C1908) and SARS-CoV-2 ORF8 (PDB ID 7jtl)
(A) The structure in blue is the structure of 7jtl and the predicted structure for the predicted ORF8 (CASP Commons C1908) is shown in red . A TM-align score of 0.34318 was achieved for the protein structures. This was normalised for SARS-CoV-2 ORF8 (PDB ID 7jtl) as it is the reference molecule **(B)** The structure in blue is the structure of 6xdc and the predicted ORF8 structure (CASP Commons C1908x2) is shown in red . A TM-align score of 0.38410 was achieved for the protein structure.

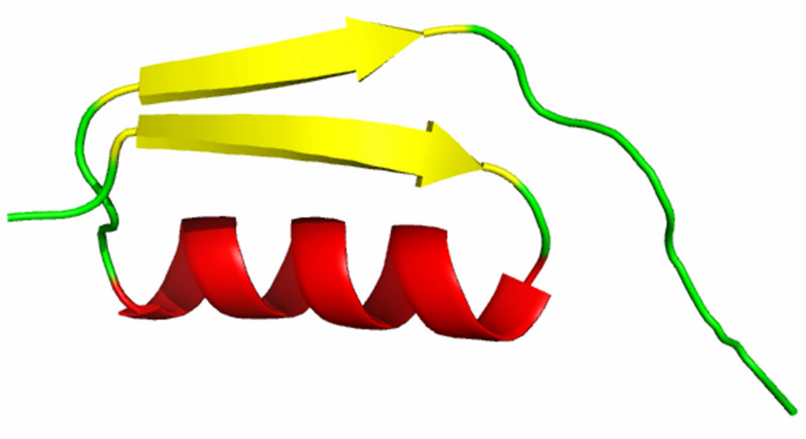


Figure 7.17. Top-scoring structure predictions from the McGuffin group for protein ORF10 (CASP Commons target C1909)
Predicted structure for ORF10 (C1909) shown as cartoon and coloured green with a residue length of 38 residues

The nineth predicted protein was ORF10, open reading frame 10 (CASP Commons target C1909). Recent experiments to characterise SARS-CoV-2 gene functions, predicted nine accessory proteins ORFs (3a, 3b, 7a, 7b, 8, 9b, 9c and 10)(Gordon *et al.*, 2020) and ORF10 is believed to be involved in ubiquitin ligases with SARS-CoV-2 (Gordon *et al.*, 2020). However, DNA nanoball sequencing concluded that SARS-CoV-2 expresses only five canonical accessory ORFs (3a, 6, 7a, 7b, 8)(D. Kim *et al.*, 2020).

The role of ORF10 has been studied in Kaposi's sarcoma-associated herpesvirus (Bisson, Page and Ganem, 2009). Viral protein replication timing regulatory Factor 1 (RIF) is a product of ORF10 and RIF is a potent and specific suppressor of interferon signalling (Bisson, Page and Ganem, 2009). Type I interferon are important mediators of innate antiviral defence and function by activating a signalling pathway through their cognate type I receptor type I receptor and this in turn triggers activation of a signalling pathway that generates a plethora of proteins with broad-spectrum antiviral activities (Bisson, Page and Ganem, 2009).

A FunFOLD3 prediction was made on one protein structure and the global model quality score was 0.4061.

IntFOLD6 predicted ligands and ligand-binding site residues for the target and results are shown in below in Figure 7.18.

Table 7.12, shows the templates associated with the protein ORF10 (target C1909) with the templates identified as having ligands by FunFOL3 highlighted in yellow. The role/function of each protein is also provided.

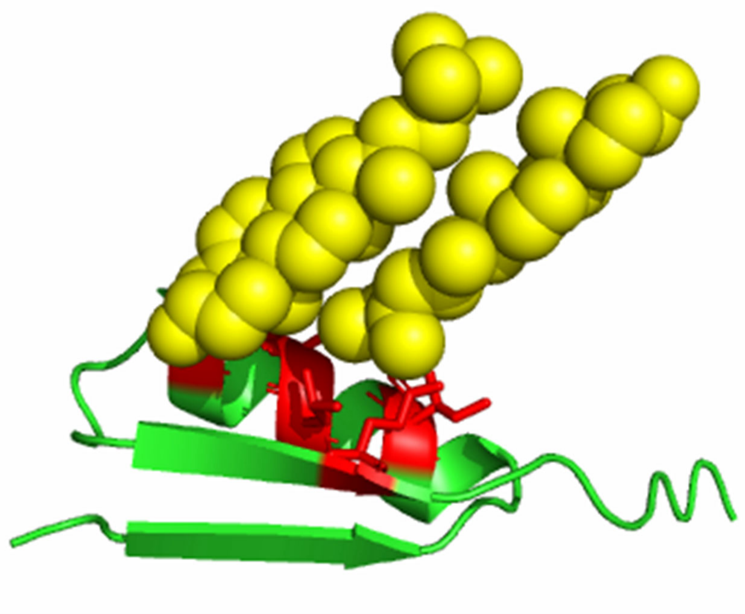


Figure 7.18. IntFOLD6 prediction for ORF10 (C1909)
Predicted structure for ORF10 (C1909) shown as cartoon and coloured green with the predicted ligand (3beta,14beta,17alpha)-ergosta-5,24(28)-dien-3-ol (94R) shown as sphere and coloured yellow. Ligand-binding sites are shown as sticks and coloured red

Table 7.12. Template list for ORF10 (C1909)
Templates for ORF10 (C1909) with the role/function and ligand as per PBD ID. Templates identified by FunFOLD3 to be of biological relevance are highlighted in yellow

Template	Role/function	Ligand
1f60B	Translation	N/A
1i8nA	Toxin	ROP
1m93A	Viral protein	PO4
1maeH6	Oxidoreductase	HDZ
1pl7A2	Oxidoreductase	ZN
1ytfC1	Transcription	DNA
2abyA	Unknown function	N/A
2lzlA	Membrane protein	N/A
2ot9A	Unknown function	SRT, NA
2wjvd	-	-
3fdfA1	Unknown function	N/A
3gdzA	Ligase	EDO
3ig9A	Viral protein	IMD
3iibA	Hydrolase	PGE, PEG, ZN
3jcuX	Membrane protein	CLA, LUT, NEX, LHG, LMG, BCR
3m6iA2	Oxidoreductase	ZN, NAD
3wcxA	Hydrolase	EPE, SIN
4cu5A	Hydrolase	N/A
4hizA	Hydrolase/viral protein	SLB, SIA, CA
4m4dA	Lipid binding protein	NAG, PC1
4o66A	DNA binding protein	SO4, NA
4q9tA	Protein transport	-
4u7nA	Transferase	-
5hy3B	Toxin/antitoxin	-
5js4A	Viral protein	MLA
5kc1c	Endocytosis	NO3, ED0, NH4, SO4, CL, IOD, NA
5o9eA	Ribosome	EDO
5yylC	Signalling protein	94R
6fbsA	RNA binding protein	DNA/RNA, ZN
6ovkb	Signalling protein	TLA
6pvra	Viral protein	-
6tdvL	Membrane protein	CDL, LMT, LPP, TRT
6vqva	RNA binding protein/RNA/inhibitor	DNA/RNA
6vqwa	RNA binding protein/RNA/inhibitor	DNA/RNA
6vqxa	RNA binding protein/RNA/inhibitor	DNA/RNA
6vz6a	Metal binding protein	HEM

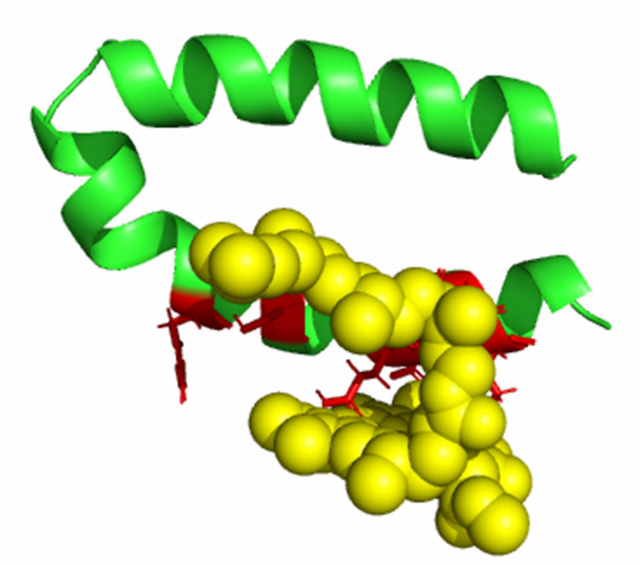


Figure 7.19. Top-scoring structure predictions from the McGuffin group for protein ORF7b (CASP Commons target C1910)
Predicted structure for ORF7b (C1910) shown as cartoon and coloured green with a residue length of 43 residues the CLA ligand is shown as sphere and coloured yellow. The ligand-binding site residues are shown as sticks and coloured red

The tenth predicted protein was ORF7b, or opening reading frame 7b (CASP Commons target C1910). As mentioned in the previous target, (C1909), ORF7b is one of the nine accessory proteins associated with SARS-CoV-2. The role of ORF7b, has been identified in SARS-CoV, where it is a putative viral accessory protein encoded on subgenomic (sg) RNA (Schaecher, Mackenzie and Pekosz, 2007). The SARS-CoV ORF7b protein is predicted to be a 44-amino-acid, highly hydrophobic protein and due to the hydrophobic nature of ORF7b, it has been hypothesised that ORF7b is a transmembrane protein and possibly a viral structural protein (Schaecher, Mackenzie and Pekosz, 2007). ORF7b localises to the Golgi complex, is an integral membrane protein, is translated for the gene 7 mRNA via ribosomal leaky scanning, is associated with intracellular virus particles and is present in purified virus particles (Schaecher, Mackenzie and Pekosz, 2007). Additionally, ORF7b has been shown to be deleted in SARS-CoV-2 (Su *et al.*, 2020).

A FunFOLD3 prediction was made on one protein structure and the global model quality score was 0.4854 and one ligand was predicted with chlorophyll A (CLA). The role of CLA.

was explored previously with nsp4 (C1902).

The UniProtKB entry identifies the subcellular location as host membrane and single-pass membrane protein. The GO terms predicted by FunFOLD3 are given below in Table 7.13. UniProtKB had GO:0016021 associated with the entry as well as GO:0033644 which was host cell membrane which was not predicted by FunFOLD3.

Table 7.13. Predicted GO terms for protein ORF7b (C1910)
The GO terms for CASP Commons target C1910 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component coloured purple

GO term	GO term domain	Function
GO:0005391	Molecular Function	sodium:potassium-exchanging ATPase activity
GO:0006754	Biological Process	ATP biosynthetic process
GO:0006813	Biological Process	potassium ion transport
GO:0006814	Biological Process	sodium ion transport
GO:0046034	Biological Process	ATP metabolic process
GO:0015979	Biological Process	photosynthesis
GO:0019684	Biological Process	photosynthesis, light reaction
GO:0016020	Cellular Component	membrane
GO:0016021	Cellular Component	Integral component of membrane
GO:0009523	Cellular Component	photosystem II
GO:0009579	Cellular Component	thylakoid
GO:0042651	Cellular Component	thylakoid membrane

7.4 Discussion

COVID-19 provided a unique opportunity for FunFOLD3 to be used on novel proteins and aid in the elucidation of function from structure for the lesser known proteins from the SARS-CoV-2 virus. Across the 32 targets (composed of the 10 full length stage 1 sequences, the 15 domain sequences and the 7 stage two full length sequences). FunFOLD3 made ligand-binding site predictions for six of the targets (C1905d1 (ORF3a), C1906x2 (transmembrane protein), C1906d1 (transmembrane protein) C1907 (ORF6), C1908x2 (ORF8), C1910 (ORF7b). For some targets, where FunFOLD3, using the top 3D model and templates, did not identify any ligands and ligand-binding site residues, using the IntFOLD6 models FunFOLD3 was able to make predictions (C1901d2 (nsp2), C1903d1 (nsp6), C1903d2 (nsp6), C1903x2 (nsp6), C1904d1 (PL-PRO), C1904d2 (PL-PRO), C1904d3 (PL-PRO), C1904x2 (PL-PRO) and C1909 (ORF10). For targets where no ligand predictions were made by FunFOLD3, the templates identified by FunFOLD3 as containing biologically relevant ligands could potentially be useful in the elucidation of function.

For target C1901 (nsp2), the largest of all the CASP Commons structures, no ligands were predicted by FunFOLD3. In terms of templates with biologically relevant ligands, there was a diverse spread ranging from membrane proteins (6jfkA), cell adhesion proteins (5yjgA, 3txaA), hydrolase (1dx5l, 2yheA, 4fyeA and 4nurA) to templates with activities related to DNA or RNA (2lggA and 4wrtC, respectively). In a publication about the role of nsp2 in the pathogenesis of COVID-19, it was found that the nsp2 differs from bat coronavirus for 11 residues (Angeletti *et al.*, 2020). Additionally, some regions of nsp2 has been shown to be structurally homologous to other known viral proteins, for example PDB ID 3ld1 (Angeletti *et al.*, 2020). This template was identified as a similar template but no ligands are associated with the protein. Furthermore, nsp2 is believed to have transmembrane helices and the predicted topology is shown in Figure 7.20 below (Angeletti *et al.*, 2020).

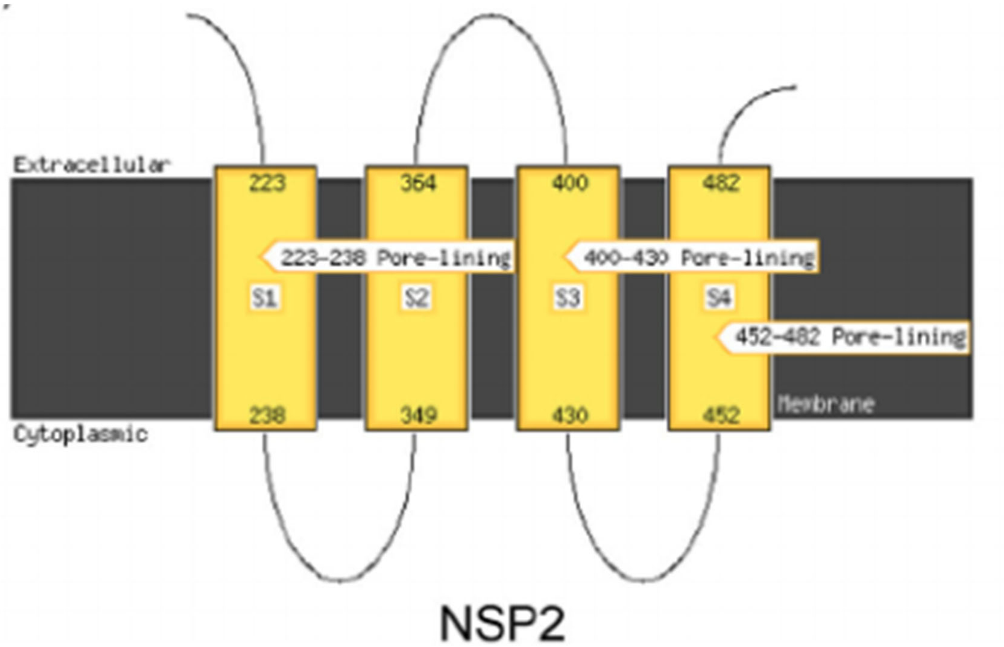


Figure 7.20. Diagram of the topology of predicted transmembrane helices
Figure taken from Angeletti et al., 2020

Position 321 of the nsp2 protein has a polar amino acid and can be speculated, that due to its side chain length, polarity and potential to form H-bonds the glutamine amino acid may confer higher stability to the protein (Angeletti *et al.*, 2020). Mutations fall within the protein nsp2 on the region homologous to the endosome-associated protein, similar to the avian infections bronchitis virus (PDB ID 3ld1) that plays a key role in the viral pathogenicity (Angeletti *et al.*, 2020). When relating this to the function identified in literature for nsp2 in decreasing interferon production to minimise virus visibility (Frolova *et al.*, 2002), it could be argued that this is aligned with the role of PDB ID 3ld1, in playing a key role in viral pathogenicity. Interferon (IFN)-γ is a critical antiviral mediator and central to the elimination of viruses. Secreted IFN-γ stimulates adaptive antigen-specific immunity and activates innate cell-mediated immunity, particularly through the activation of macrophages (Kang, Brown and Hwang, 2018). Overall, IFN-γ is a broad-spectrum anti-microbial agent and a crucial regulatory of overall inflammatory responses to pathogens (Kang, Brown and Hwang, 2018).

Hence, how inhibition of IFN-γ, minimises the virus's visibility and enables profound infection with the virus (Kang, Brown and Hwang, 2018).

Additionally, PDB ID 3ld1 is a hydrolase and the information in literature has shown that the nsp2 protein of SARS-CoV-2 contains polar amino acids (Angeletti *et al.*, 2020). Serine, is a polar amino acid and serine hydrolases are known to perform crucial functions in bacteria and viruses where they contribute to pathogen life cycle (Shahiduzzaman & Coombs, 2012). All serine hydrolases possess a common catalytic mechanism that involves activation of a conserved serine nucleophile for attack on a substrate ester/thioester/amide bond to form an acyl-enzyme intermediate, followed by water-catalysed hydrolysis of this intermediate to liberate the product (Shahiduzzaman & Coombs, 2012). Thus, the role of this target in the pathogenesis of SARS-CoV-2, could as a serine hydrolase which attaches to a host cell and by the catalytic mechanism prevents IFN-γ from exhibiting its effects as an anti-viral agent. Most likely, by preventing IFN-γ from binding to Janus kinase/signal transducer and activator of transcription (JAK/STAT) protein signal transduction pathways (Kang, Brown and Hwang, 2018). IFN-γ can target various stages within a viral life cycle which includes entry, replication, gene expression, stability, release and reactivation (Kang, Brown and Hwang, 2018). Hence, why it is an attractive target for viruses. The full antiviral mechanism of IFN-γ is depicted below in Figure 7.21 (Kang, Brown and Hwang, 2018).

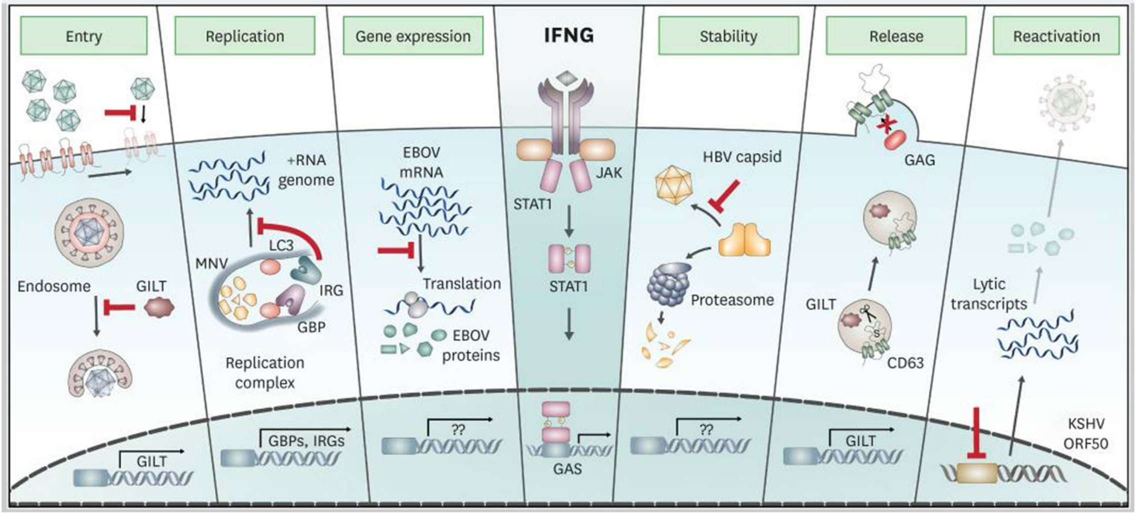


Figure 7.21. Antiviral mechanisms of IFN-γ
IFN-γ obstructs the various stages of viral life cycle in the cells. Representative examples are depicted here: IFN-γ inhibits viral entry at both extracellular and intracellular stages, replication by disrupting replication niche, gene expression by hindering translation, stability by impeding nucleocapsid assembly, release by breaking the disulfide bond of a necessary cellular interaction partner, and reactivation by suppressing the transcription of a viral master regulator. Red colour bars signify inhibiting function of IFN-γ. Figure taken from Kang, Brown and Hwang, 2018

Protein nsp4 (CASP Commons target C1902), as with nsp2 (CASP Commons C1901) no ligands were predicted by FunFOLD3 based on the top selected model. However, the FunFOLD3 component of IntFOLD6 predicted chlorophyll A. The potential role of this was mentioned with nsp4 (CASP Commons C1902). In addition, the application of porphyrins and their use in the inactivation of viruses has been published (Sh Lebedeva *et al.*, 2020) and the abnormal concentration of porphyrins in the serum of COVID-19 patents has also been reported (San Juan *et al.*, 2020).

In terms of templates, the term “transport protein” is by far the most common with frequency and specifically templates with ligands (1f86A, 2a65A, 3a6pA, 5mkeA, 5w3sA and 6d6tB).

As a whole coronaviruses, consisting of HCoV-OC43, which causes the milder common-cold like symptoms, SARS-CoV-1 (emerged in 2002), MERS-CoV (in 2012) and the recent SARS-CoV-2 possess the largest known RNA viral genomes and the 5’ 20 kb region of the genome encodes for two open reading frames (ORF1a/1ab) that produces 16 non-structural proteins (nsp1-16) needed to form the viral replication complex, while the 3’ proximal region encodes for the structural proteins and several accessory factors with varying roles, as shown below in Figure 7.22 (Davies, Jonathan; Almasy, McDonald and Plate, 2020). NSP4 is 80% identical between SARS strains but only 42% between SARS and OC43 strains (Davies, Jonathan; Almasy, McDonald and Plate, 2020).

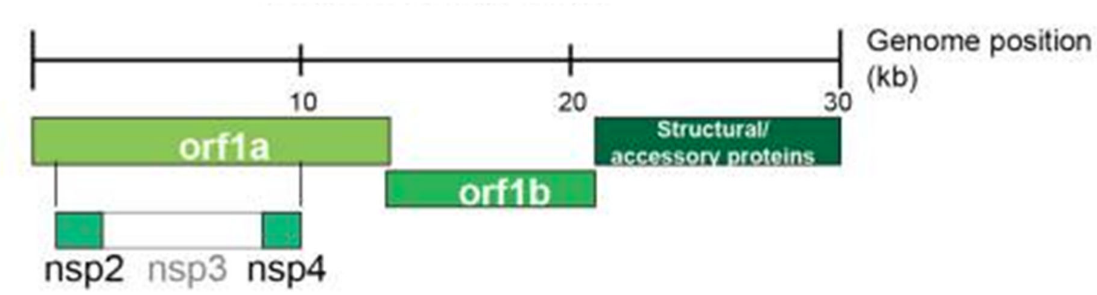


Figure 7.22. Schematic of SARS-CoV-2 genome
Figure taken from Davies, Jonathan; Almasy, McDonald and Plate, 2020

As can be seen in Figure 7.23, nsp4, is a transmembrane glycoprotein and is better defined most notably in formation of the double-membrane vesicles associated with replication complexes and unlike nsp2, has a high degree of sequence similarity across human coronavirus strains (Davies, Jonathan; Almasy, McDonald and Plate, 2020). Nsp4 contains hydrophobic stretches and is predicted to be integral membrane protein and is likely to function in anchoring the replication complexes to the lipid bilayer (Davies, Jonathan; Almasy, McDonald, 2020). This could explain why the term “transport protein” has been the common among predicted templates. Figure 7.23 below is a schematic representation of the intracellular localisation and cellular transport for nsp4 and the SARS-CoV M protein (Oostra *et al.*, 2007). Nsp4 is transported to the double membrane vesicles without passing through the Golgi compartment, once again supporting the common templates for transport proteins. Furthermore, the commonly associated ligand with transport protein templates, cholesterol hemisuccinate is an acidic cholesterol ester that self-assembles into bilayers in alkaline and neutral aqueous media, (Hafez & Cullis, 2000) this sounds similar to double-membrane vesicles (DMV) which is depicted in Figure 7.23.

Additionally, GO-terms associated with the nsp4 showed terms related to biological processes for cell organisation and biogenesis, transport and metabolic process (Davies, Jonathan; Almasy, McDonald and Plate, 2020). In evaluation of cellular compartment GO-terms, nsp4 interactors are enriched in membranes of the endoplasmic reticulum and the mitochondria (Davies, Jonathan; Almasy, McDonald and Plate, 2020). In particular ERLIN1/2 and RNF170. SARS-CoVs may use ERLIN1/2 to regulate ER Ca^{2+} signalling and the myriad of downstream host processes controlled by this signalling pathway (Davies, Jonathan; Almasy, McDonald and Plate, 2020).

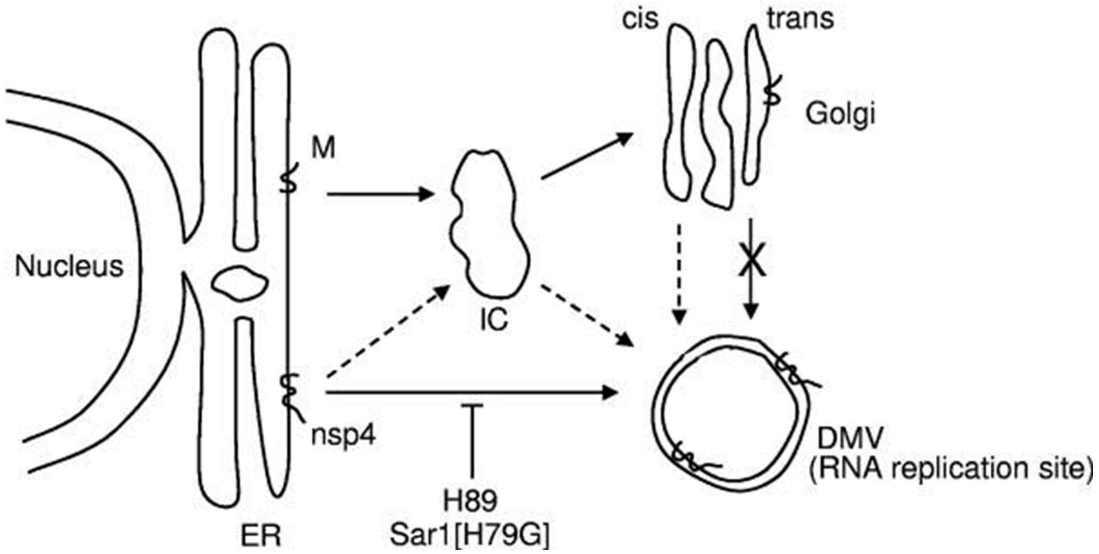


Figure 7.23. Intracellular localisation and celular transport for nsp4
Intracellular localisation and cellular transport for nsp4 and SARs-CoV M protein. The M protein is transported from the endoplasmic reticulum to the Golgi compartment via the intermediate compartment (IC), whereas nsp4 is transported to the double membrane vesicles (DMV) without passing through the IC. Figure taken from Oostra et al., 2007

Nsp6 is the last of the non-structural proteins available for prediction; (CASP Commons ID C1903) . As with nsp2 (C1901) and nsp4 (C1902), FunFOLD3 did not predict any ligands when using the top selected 3D model. However, the FunFOLD3 component of IntFOLD6 predicted a diverse range of ligands with haemoglobin, 24-methylenechloresterol (94R), monoundecenion (MUN), [Z-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG) and retinal (RET). Membrane protein was the commonly predicted term associated with templates (4jkvA, 4qi1A, 6bmsA, 6csmA, 6eyuA and 6i1rA) followed by transport protein (3zuxa, 5hvdA, 5xpdA1, 5y78A and 6pw4A). As can be imagined, literature is fast-paced in the SARS-CoV-2 space and a recent publication identified non-structural proteins as critical elements of the replication and transcription complex (RTC), as well as immune system evasion (Santerre *et al.*, 2020). Co-expression of all three SARS-CoV NSPs (nsp3, nsp4 and nsp6) is essential to induce DMVs and these proteins contain particular multiple transmembrane domains that help the virus replication complex via the recruitment of intracellular membranes (Santerre *et al.*, 2020). Figure 7.24 below shows the nsp6 domain similarity between SARS-CoV and SARS-CoV-2.

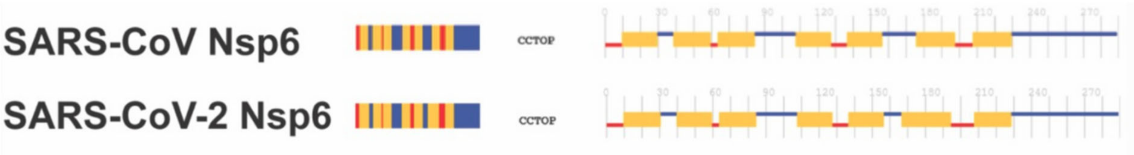


Figure 7.24. Nsp6 domain similarities
The 1D and 2D panels show the amino acid sequence coloured by the consensus topology. Colours are based on the localisation: grey, black, blue red and yellow and orange for transit sequence, signal peptide, extra-cytosolic, cytosolic, membrane, and re-entrant loop regions, respectively. Figure taken from Santerre et al., 2020

Nsp6 is a membrane protein (Santerre *et al.*, 2020), which fits in with the commonly occurring template and is approximately 34 kDa with six transmembrane helices and a highly conserved C-terminus. When inserted in the ER membrane, it associates with nsp3 and nsp4 multi-pass transmembrane proteins during the assembly of coronavirus replication complex to form DMVs (Santerre *et al.*, 2020). Nsp6 may modify adaptive immune responses by sending immunomodulatory proteins synthesised by the ER into autophagosomes for degradation and the SARS-CoV-2 nsp6 protein also interacts with the sigma receptor, which is known to participate in ER stress response(Santerre *et al.*, 2020). A further mechanism is that nsp6 binds TANK binding kinase 1 (TBK2) to supress interferon regulatory factor 3 (IRF3) phosphorylation (Xia, et al., 2020). An example of the inhibitory role of nsp6 is shown below in Figure 7.25. The role of interferon, specifically IFN-γ has been discussed previously. Furthermore, nsp6 along with other non-structural proteins and open reading frames may suppress STAT1 and STAT2 phosphorylation and thus suppress the nuclear translocation of STAT1 during IFN signalling (Xia, et al., 2020).

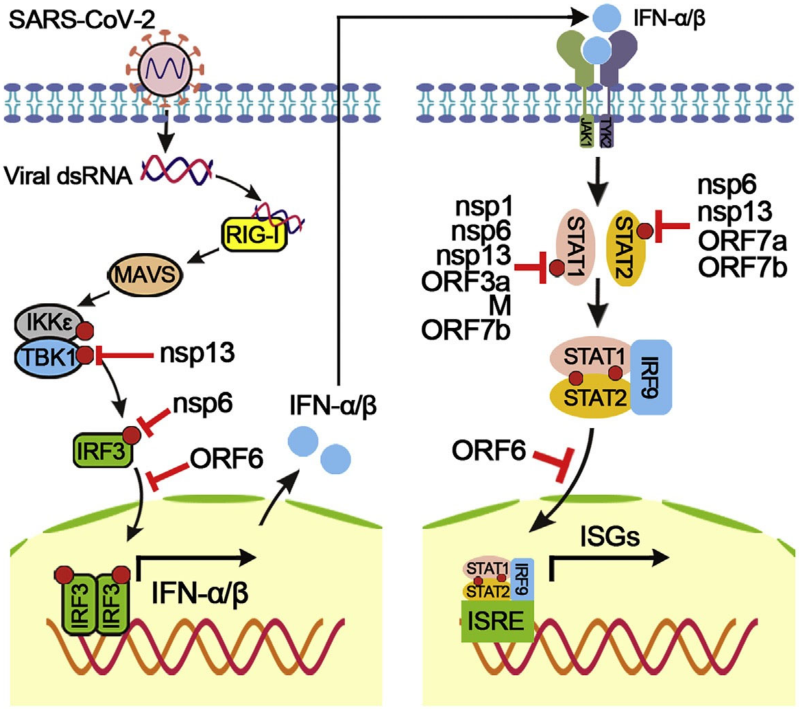


Figure 7.25. Summary of antagonism of IFN production
The inhibitory steps are indicated for individual viral proteins. Figure taken from Xia et al., 2020

PL-PRO (C1904) is the only target for papain-like protease, encoded by nsp3, PLpro is one of two known CoV proteases and is required for efficient cleavage of nsp2 and nsp3 from the viral polyprotein, a process essential for viral genome transcription and replication. Both SARS-CoV-2 and SARS-CoV critically relies on the activity of viral proteases (Shin *et al.*, 2020). Papain-like proteases are the protease domain from the membrane anchored multi-domain protein nsp3 and generate a functional replicase complex to enable viral spread (Shin *et al.*, 2020). The crystal structure of SARS-CoV-2 PLpro has been determined and is shown below in Figure 7.26. As can be seen in Figure 7.26, there is a clear finger and thumb region of the protein with the beta sheets and alpha helixes, respectively. There was an increase in the number of beta sheets from the first prediction in Figure 7.9A to Figure 7.9E, however neither structures look as ordered, as expected, compared to the crystal structure.

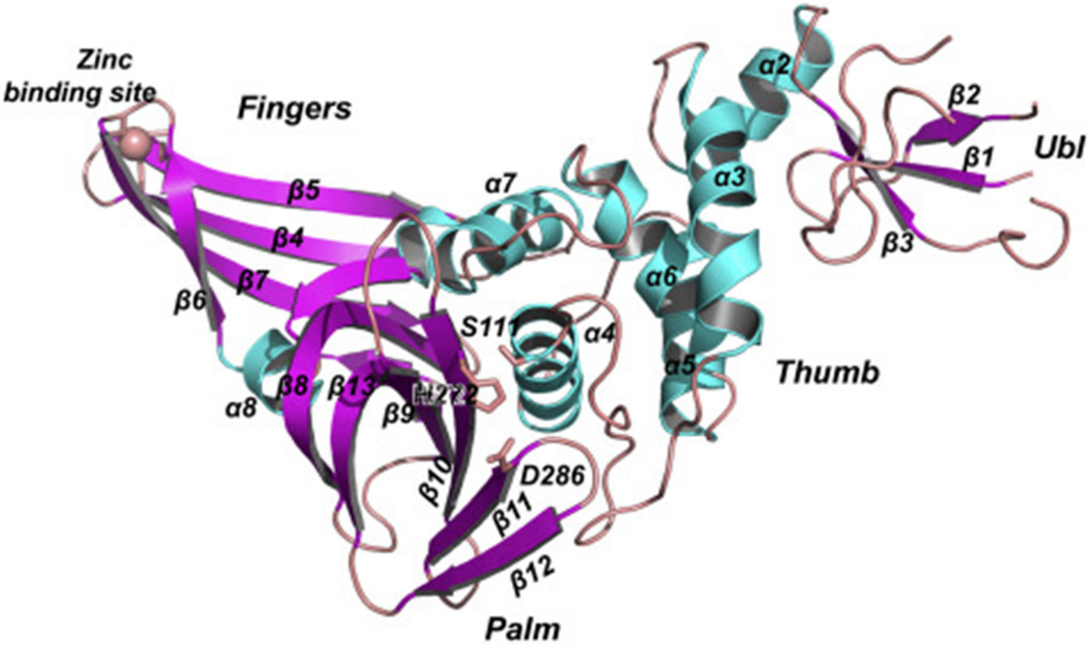


Figure 7.26. Crystal structure of the unliganded SARS-CoV-2 PLpro
Ribbon model of the unliganded SARS-CoV-2 PLpro. β-strands magenta, α-helices cyan. PLpro subdomains Ub1, Thumb, Finger and Palm are indicated. Four cysteine residues forming the zine finger on the Finger subdomain are shown with stick model. Figure taken from Gao et al., 2020

In terms of templates, the common terms were oxidoreductase (5mscA, 5msoA, 5mspA and 6du7A) and transport protein/protein transport (1z2aA, 3qf4A, 6ltcA, 6nq0A and 6oh2A). SARS-CoV-2 PLpro and the two ISG15 domains are compared to MERS PLpro (PDB ID 6bI8),(Shin *et al.*, 2020) the latter protein is classified as a hydrolase/substrate. It is worth noting, that PDB ID 6bI8 was not identified as a template with similar folds. The structural and accessory proteins are variable between SARS-CoV and SARS-CoV-2, however, the virally-encoded replicases are highly conserved (Gao *et al.*, 2020).

PLpro along with main protease are responsible for the processing of viral polyproteins (pp1a and pp1ab) yielding mature viral proteins and has been suggested as an attractive drug target for treating COVID-19. Additionally, PLpro suppresses innate immunity through reversing the ubiquitination and ISGylation events (Gao *et al.*, 2020) and SARS-CoV-2 preferentially reduced the appearance of ISG15-conjugated (ISGylated) protein substrates (Shin *et al.*, 2020). Ubiquitination and ISGylation play important roles in the regulation of innate immune responses to viral infection, ubiquitination is a post-translational modification characterised by the addition of ubiquitin chains to lysine residues of a protein, which regulates its activity, notability via its targeting to proteasomal degradation (McClain & Vabret, 2020). ISGylation involves interferon-stimulated gene 15 (ISG15), a small protein highly induced by IFN-α and γ, by viral infection and double-strand as well as ischaemia, DNA damage and aging, is conjugated to target proteins and modulates their functions (Villarroya-Beltri, Guerra and Sánchez-Madrid, 2017). ISGylation ultimately blocks the entry, replication or release of different intracellular pathogens (Villarroya-Beltri, Guerra and Sánchez-Madrid, 2017). The induction of ISG15 in viral infection is given below in Figure 7.27

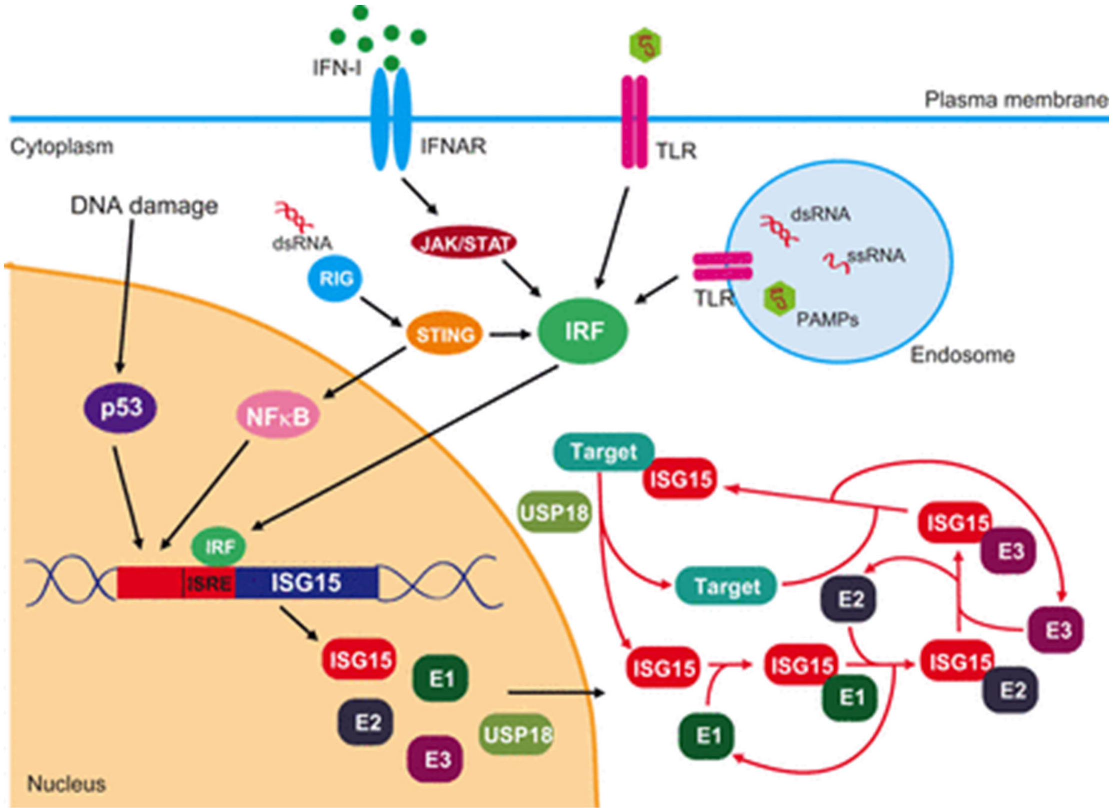


Figure 7.27. ISG15 induction and conjugation pathways
ISG15 expression is induced upon binding of interferon response factors (IRF) to the interferon-stimulated response element (ISRE) located in the ISG15 promoter. This binding is induced by type I interferon (IFN-I) through activation of IFN receptor (IFNAR) and JAK/STAT signalling, as well as by single strand (ss)RNA, double-strand (ds)RNA, or other viral compounds (pathogen-associated molecular patterns; PAMPs). Figure taken from Villarroya-Beltri, Guerra and Sánchez-Madrid, 2017

Information available in literature (Gao *et al.*, 2020) has found that GRL-0617, is an inhibitor of SARS-CoV-2 PLpro and ultimately inhibits the deubiquitination and deISGylation activities of SARS-CoV-2 (McClain & Vabret, 2020) and thus the proteolytic processing of the viral polypeptide. The structure is given below in Figure 7.28 and the effect in the viral replication pathway is shown in Figure 7.29 (McClain & Vabret, 2020). The closest example is GRL-0617 would be the prediction of GDP by the FunFOLD3 component of IntFOLD6 and the figures are shown below. Whilst not exact, the structures are similar in terms of containing benzenes rings and amine groups.

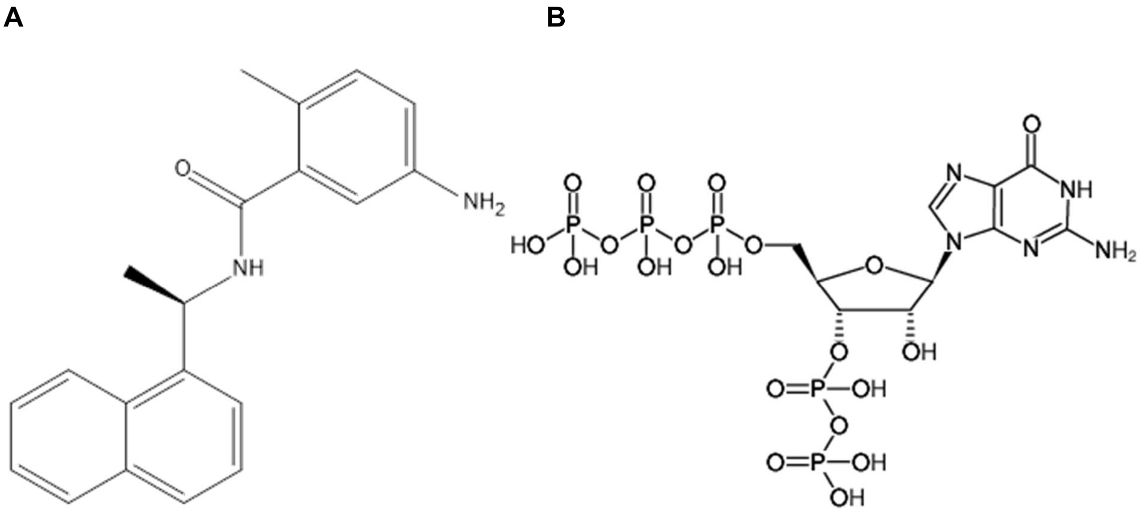


Figure 7.28. Comparison of GRL-0617 and GDP ligand
The difference between GRL-0617 (A) and GDP (B) ligand. The GRL-0617 ligand has been shown as a ligand in literature and the GDP ligand was a biologically relevant ligand identified in the predicted structure by the FunFOLD3 component of the IntFOLD6 server. Figure for GRL-0617 taken from Báez-Santos, St John and Mesecar, 2015

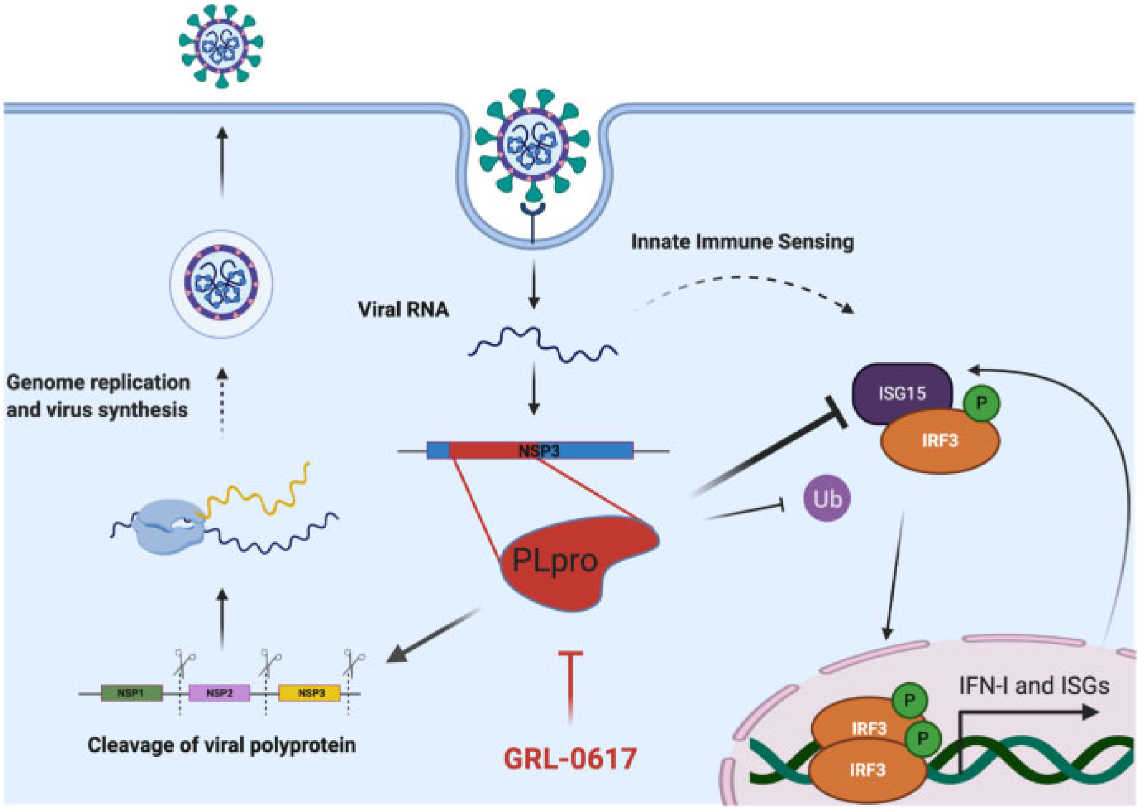


Figure 7.29. Dual role of SARS-CoV-2 protease PLpro in viral replication and inhibition of innate sensing
PLpro is required for the processing of SARS-CoV-2 polyprotein into mature sub-units to generate a functional replicase complex. Additionally, PLpro antagonizes the ISGylation of cellular proteins, including IRF3, leading the dysregulation of innate immune sensing. GRL-0617 targets PLpro and prevents genome replication and virus synthesis and the dysregulation of innate immune sensing. Figure taken from McClain & Vabret, 2020

ORF3a (C1905) was the first CASP Commons target with a ligand predicted by FunFOLD3. The ligand (Z)-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG) was predicted in the C1905d1 target spanning the 1-130 residues for the whole target. Additionally, this target also had a solved structure as per PDB. No ligands were identified in the solved structure and there was also a poor overall structural superposition between the predicted and observed structure.

In total there are six accessory proteins (ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10) in SARS-CoV-2 (Majumdar & Niyogi, 2020). SARS-CoV-2 ORF3a bears 72.4% sequence identity and 85.1% sequence similarity with that of SARS-CoV (Majumdar & Niyogi, 2020). The role of ORF3a in both SARS-CoV and SARS-CoV-2 are ion channels (viroporins) and involved in virion assembly and membrane budding (Tan, Schneider, Shukla, Chandrasekharan, Aravind and Zhang, 2020). Additionally, *in vitro* studies have shown that SARS-CoV-2 ORF3a can efficiently induce apoptosis in cells (Ren *et al.*, 2020). Furthermore, it was found that ORF3a induces apoptosis via the extrinsic pathway, in which activated caspase-8 cleaves Bid to tBid and in turn induces the release of mitochondrial cytochrome c, resulting in apoptosome formation and caspase-9 cleavage/activation and it’s pro-apoptotic activity is weaker than SARS-CoV ORF3a (Ren *et al.*, 2020) which could potentially add to the virus’s pathogenicity. On drug bank the predicted ligand MPG is predicted to be a weak inhibitor of hERG (Wishart *et al.*, 2018), which is a voltage sensitive K+ channel with a fundamental role in cardiac action potential repolarisation (Butler *et al.*, 2020) Although, this would not specifically be relevant to SARS-CoV-2 ORF3a, targeting ion channels to inhibit virion and membrane budding could provide some basis particularly because ORF3a forms homotetrameric potassium sensitive ion channels (UniProt Consortium, 2019).

C1906, or transmembrane protein was the next target with FunFOLD3 ligand predictions with a total of five predictions across two targets. Despite no PDB ID being associated with the target, there is published information of the structure and function of the membrane protein, which happens to the most abundant structural protein of SARS-CoV-2 (Thomas, 2020). The M protein of SARS-CoV-2 is 98.6% similar to the M protein of bat SARS-CoV (Thomas, 2020) and spans the membrane bilayer, leaving a short NH2-terminal domain outside the virus and a long COOH terminus (cytoplasmic domain) inside the virion (Mousavizadeh & Ghasemi, 2020). Binding with M protein helps to stabilise N proteins and promotes complete of viral assembly by stabilising the N protein-RNA complex, inside the internal virion (Astuti & Ysrafil, 2020).

The predicted structure from Thomas (Thomas, 2020) is shown below in Figure 7.30 as predicted using I-TASSER, as can be seen the structure does bare some similarities with the top ranked predicted structure selected from ModFOLD8 (C1906x1) with three distinct alpha-helices and beta-sheets are also present, however the I-TASSER model has more flexible loops (Thomas, 2020). Comparisons with 3D models and the models produced by the McGuffin group are shown in Appendix 5.

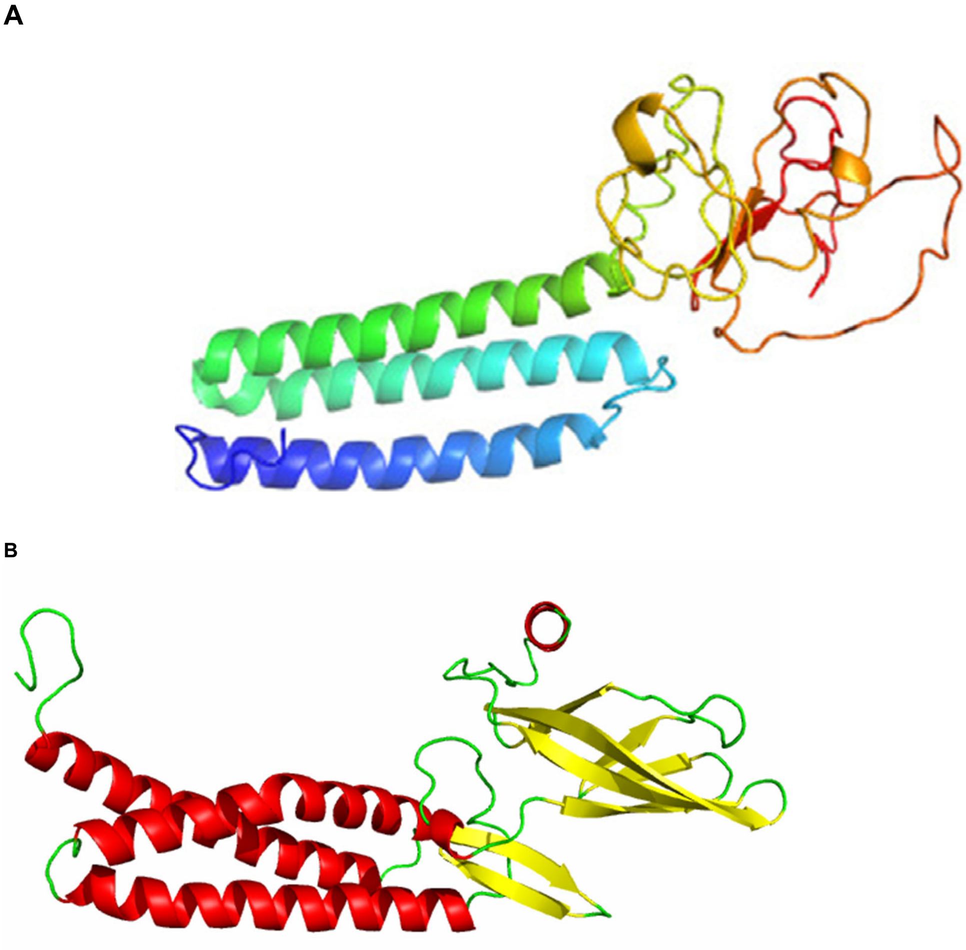


Figure 7.30. Predicted M protein structure of SARS-CoV-2 and transmembrane protein (C1906x1) from the McGuffin group
(A) Predicted structure of membrane protein using the software I-TASSER. Similarities between the top selected model for C1906x1 **(B)** can clearly be seen with the alpha helices. Figure taken from Thomas, 2020

In silico analyses of the M protein demonstrated that it has a triple-helix bundle and forms a continuous three-transmembrane domain. The M protein has a short amino terminal domain inside the viral envelopes as shown below in Figure 7.31.

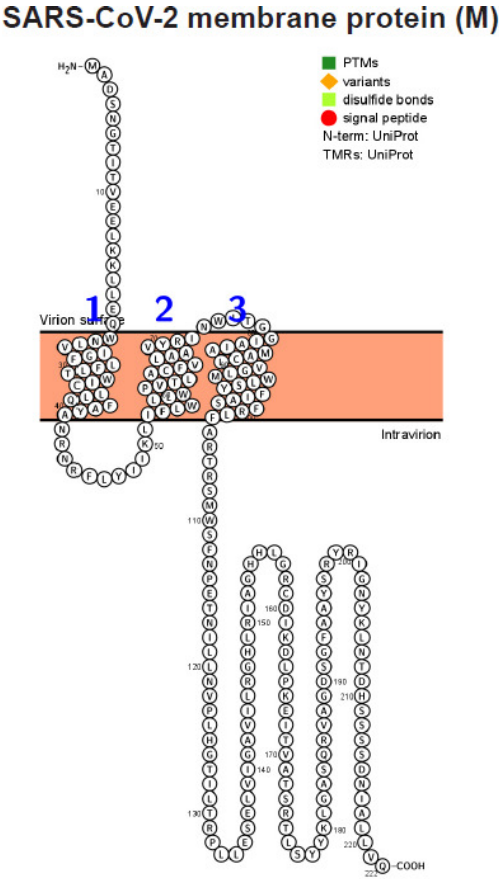


Figure 7.31. Membrane topology of Membrane protein (snake diagrams)
Membrane glycoprotein of SARS-CoV-2 has a triple helix bundle and formed a single three-transmembrane domain. Figure taken from Thomas, 2020

Additional, *in silico* data analysis has shown that the M protein resembles the SemiSWEET sugar transport of prokaryotes, which mediate the movement of amino acids across lysosomal membranes, which could explain why amino acids were predicted as binding ligands by FunFOLD3 for transmembrane protein (C1906x2). In the study by Thomas (Thomas, 2020), it was hypothesised that that the sugar transport-like structure of the M protein influences glycosylation of the S protein. Sucrose is involved in endosome and lysosome maturation and may also induce autophagy, pathways that help the entry of the virus (Thomas, 2020). However, further experiments would be needed to validate these findings (Thomas, 2020).

The first accessory proteins and thus, open reading frame targets is C1907 with ORF6. ORF6 of SARS-CoV-2 has 69% amino acid identity with SARS-CoV (Miorin *et al.*, 2020).

Based on studies with SARS-CoV, the role of ORF6 has been implicated in virus replication due to interaction with nsp8 (Kumar *et al.*, 2007). SARS-CoV ORF6 is localised in the ER and Golgi membranes in infected cells and binds to karyopherin alpha 2 and karyopherin beta 1 proteins and hinders STAT1 nuclear import and ultimately STAT1 function (Frieman *et al.*, 2007). Following interferon stimulation STAT1:STAT1 or STAT1:STAT2 dimers translocate into the nucleus by binding to the import receptor karyopherin alpha 1, KPNA1. The STAT:KPNA1 complex interacts with KPNB1, which mediates docking of the import complex to the nuclear pore complex (NPC).(Miorin *et al.*, 2020). Studies have shown that SARS-CoV ORF6 interferes with IFNAR signalling by tethering KPNA2 and KPNB1 to the endoplasmic reticulum/Golgi membrane to block STAT1 nuclear import (Kopecky-Bromberg et al., 2007;Frieman et al., 2007). Experimentally, SARS-CoV-2 ORF6 has been shown to interact with both KPNA1 and KPNA2 and shown a change in localisation of KPNA1 and KPNA2 form the nucleus to the cytoplasm in ORF6-expressing cells (Miorin *et al.*, 2020). However, overexpression of KPNA1 or KPNA2 could not rescue the ORF6-dependent block of STAT1-GFP nuclear translocation, thereby suggesting another factor which is involved in ORF6 blocking IFN signalling. The C-terminal region of ORF6 directly binds the nucleoporins98 (Nup98) Nup98-Rae1 complex (Miorin *et al.*, 2020). ORF6 is associated with Nup98 both at NPCs present at the nuclear envelope and at annulate lamellae in the cytoplasm. These data are consistent with ORF6 altering nuclear transport functions of Nup98 at NPCs and it has been shown that ORF6 specifically targets Nup98 to block STAT nuclear import (Miorin *et al.*, 2020). In a paper by Gordon et al, which looked at targets for drug repurposing, Selinexor was identified as a potential pharmacological target due to the interaction of ORF6 and the mRNA nuclear export complex Nup98-Rae1 (Gordon *et al.*, 2020).

C1908, or open reading frame 8, ORF8 was the second of the open reading frames to have a predicted ligand and also has an observed structure and a PDB ID associated. The ligand was predicted following the second round of modelling.

Unlike ORF6, ORF8 shares a low sequence homology with SARS-CoV of 26% (Zhang et al., 2020). ORF8 is an immunoglobin-like protein that modulates pathogenesis (Zinzula, 2020) and is part of a hypervariable genomic region of ~430 bp in length and is a recombination hotspot and is highly susceptible to deletions and nucleotide substitutions (Zinzula, 2020).

Computational analysis of ORF8 revealed that its structural organisation resembles the one observed among members of the immunoglobin-like domains containing protein superfamily (IgSF)(Zinzula, 2020). Encoded by a variety of viruses, IgSF proteins seem to evolved from host-acquired genes and so have evolved to mimic the original host function, which consist of cell-to-cell adhesion or ligand-receptor recesses, thereby interfering with and acting as molecular traps immunomodulatory properties (Zinzula, 2020). Indeed, this relates to the GO terms predicted by FunFOLD3, with cell adhesion being predicted as a biological process. The crystallised structure of ORF8 (PDB ID 7jtl) reveals a covalently-bound dimer, held by an intermolecular disulphide bridge between the two cysteine residues at position 30. In each monomer are a ß-sheet core of eight antiparallel ß-strands is held together by three intramolecular disulphide bridges, while two ß-strands from each core are involved in hydrophobic interactions with their counterparts in the other monomer to further stabilise the dimer interface (Zinzula, 2020). Figure 7.32 below shows the two hypothetical states of ORF8.

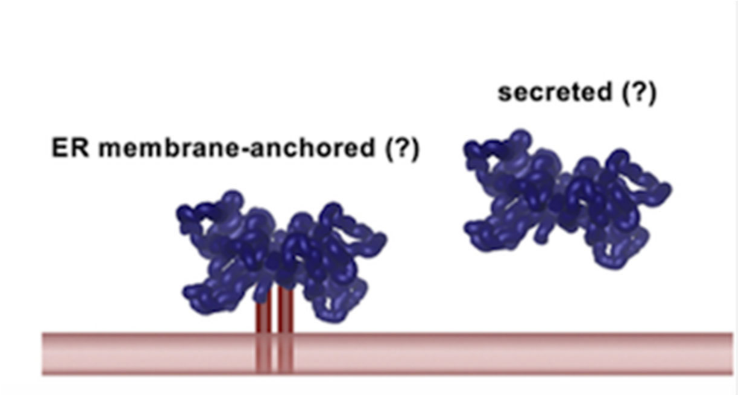


Figure 7.32. Schematic diagram showing the two hypohetical states of ORF8
Membrane-anchored and secretory states of ORF8. Figure taken from Zinzula, 2020

Cytotoxic T lymphocytes (CTLs) are important for the control of viral infections by directly eradicating the virus-infected cells. In a virus-infected cell, MHC-1 molecules present peptides derived from a variety of viral proteins (Zhang et al., 2020). Once the T cell receptor on CD8$^+$ T cells recognise the signal presented by MHC-1-peptide complex, the CTL releases varies toxic substances including perforins, granzyme and FasL which directly induce the death of viral-infected cells, as well as many other cytokines such as IFN-γ, TNF-α and IL-2 (Berke, 1995). Figure 7.33 shows the inhibition of the type 1 IFN antiviral response.

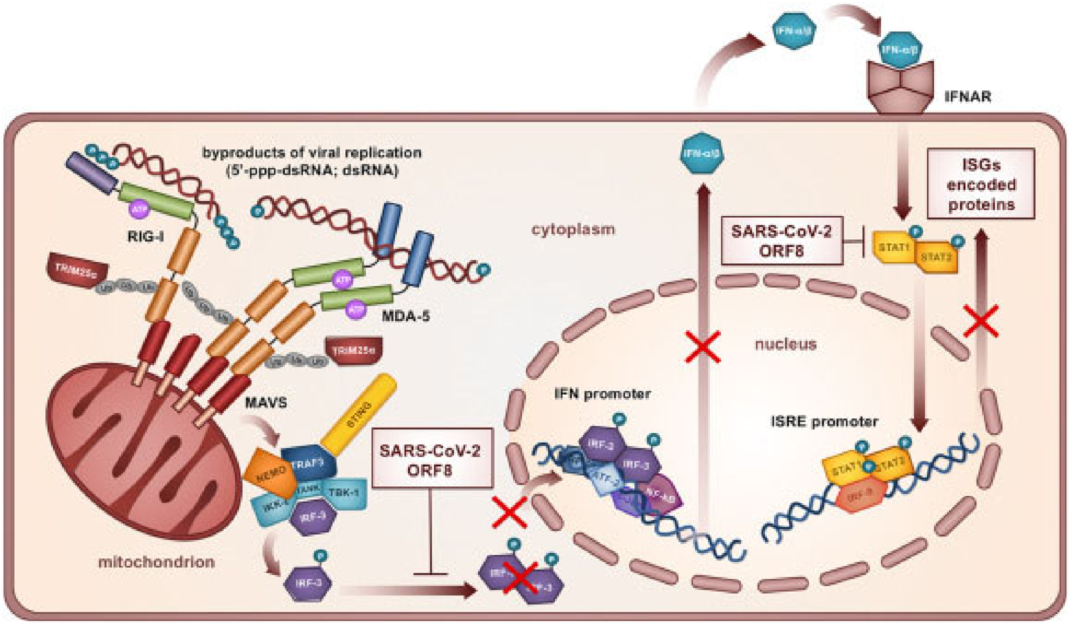


Figure 7.33. Inhibition of the type I IFN antiviral response by ORF8
Schematic diagram describing how ORF8 suppresses IFN stimulated expression of ISGs. Figure taken from(Zinzula, 2020)

Experimental studies have shown that the overexpression of ORF8 in 293T cells significantly down-regulates MHC-1 molecules.

The mechanism by which ORF8-mediated MHC-I reduction is possibly due to lysosomal degradation by an autophagy-dependent mechanism and in ORF8-expressing cells, the surface expression of MHC-1 was almost abrogated and redistribute into cytoplasm (Zhang et al., 2020). Thus, allowing the virus to evade immune surveillance and increase in viral load. Additionally, data has shown that ORF8 can bind directly to MHC-I molecules and targets for lysosomal degradation. The authors concluded that based upon the available data, instead of regular routing through Golgi to plasma membrane, MHC-I at ER is captured by ORF8 and is re-routed to autophagosome and subsequently to autolysosome for degradation (Zhang et al., 2020). SARS-CoV-2 utilises its ORF8 as a unique mechanism to alter the expression of, but not limited to, surface MHC-I expression to evade immune surveillance (Zhang et al., 2020). This mechanism supports the GO term prediction by FunFOLD3 of innate immune response. Figure 7.34 below shows the dysregulation of the MHC-I mediated antigen presentation by ORF8.

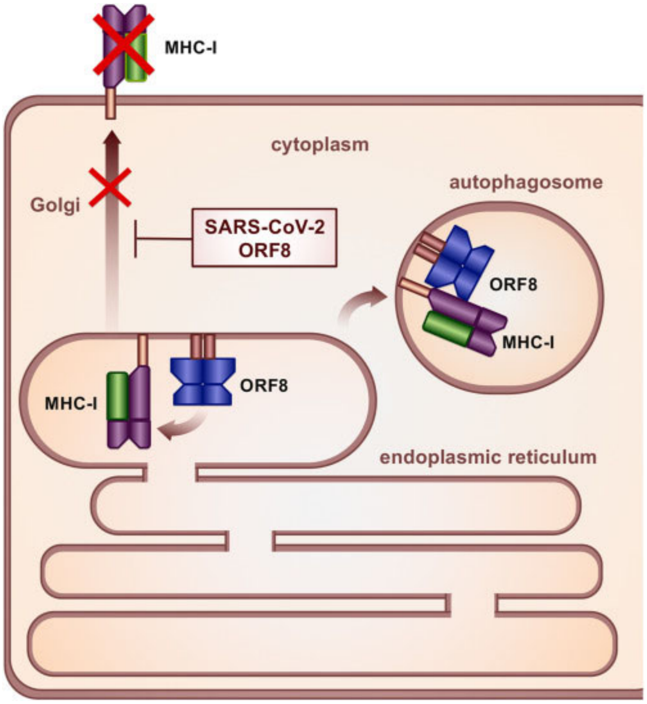


Figure 7.34. Downregulation of the MHC-I mediated antigen presentation by ORF8
Schematic diagram describing how ORF8 mediates the MHC-I degradation via an autophagy-dependent pathway. Figure taken from Zinzula, 2020 which is adapted from Zhang et al., 2020

In a paper by Gordan et al, (Gordon *et al.*, 2020) that focused on drug repurposing, rapamycin was identified as a potential target for ORF8 (Gordon *et al.*, 2020). Several proteins are found to interact with ORF8 and these include complex formed by transforming growth factor-β 1 (TGFβ1), latency associated peptide (LAP) and latent TGFβ binding protein 1 (LTBP1), and with the complex formed by integrin subunit alpha 3 (ITGA3) and serpin family E member 1 (SERPINE1). Potentially, these could be investigated as targets.

C1909, ORF10 is proposed to be unique to SARS-CoV-2 (Wu *et al.*, 2020) and there is no data to provide evidence that the protein is expressed during SARS-CoV-2 infection (Michel *et al.*, 2020) and at 38-residue peptides, is the smallest accessory protein (Hassan *et al.*, 2020). New viruses can originate from existing proteins acquired through horizontal gene transfer or through gene duplication, or can be generated de novo (Michel *et al.*, 2020). Michel et al., (Michel *et al.*, 2020) suggested that the ORF10 of SARS-CoV-2 evolved via the mutation of a stop codon (TAA) at nucleotide 76 and the addition of a new *X* motif of length 15 nucleotides in the 3’ region. Of all the putative ORF proteins, ORF10 has the highest number of immunogenic epitopes, therefore making it a potential target for vaccine development (Kiyotani *et al.*, 2020).

ORF10 consists of a molecular recognition feature region from amino acid residue 3-7, which is a molecular recognition site for interaction with other proteins (Giri *et al.*, 2020). Molecular recognition features, are intrinsic disorder-based protein-protein interaction sites that are commonly utilised by proteins for interaction with specific partners (Giri *et al.*, 2020). One of the critical properties of intrinsically disordered proteins that allow proteins to adapt an ensemble of conformations when bound to different proteins, and hereby permits interaction with multiple proteins (Uversky, 2011). Through high-throughput analysis it has been revealed that ORF10 can interact with a large number of hot proteins and most likely due to the MoRF region (Giri *et al.*, 2020). Through bioinformatic techniques it has been reported that ORF10 interacts with multiple members of the Cullin-ubiquitin-ligase complex and controls host-ubiquitin machinery for viral pathogenesis (Gordon *et al.*, 2020). Specifically the CUL2^{ZYG11B} complex, in particular the α-helical region. The ubiquitin transfer to a substrate requires neddylation of CUL2 by NEDD8-activating enzyme (NAE), which is a druggable target (Gordon *et al.*, 2020). One of the druggable targets which has been identified is pevonedistat, which inhibits NAE. Whilst FunFOLD3 did not predict any ligands, the FunFOLD3 component of IntFOLD6 predicted 94R, or 24-methylenecholesterol and the comparison between pevonedistat and 94R is given below:

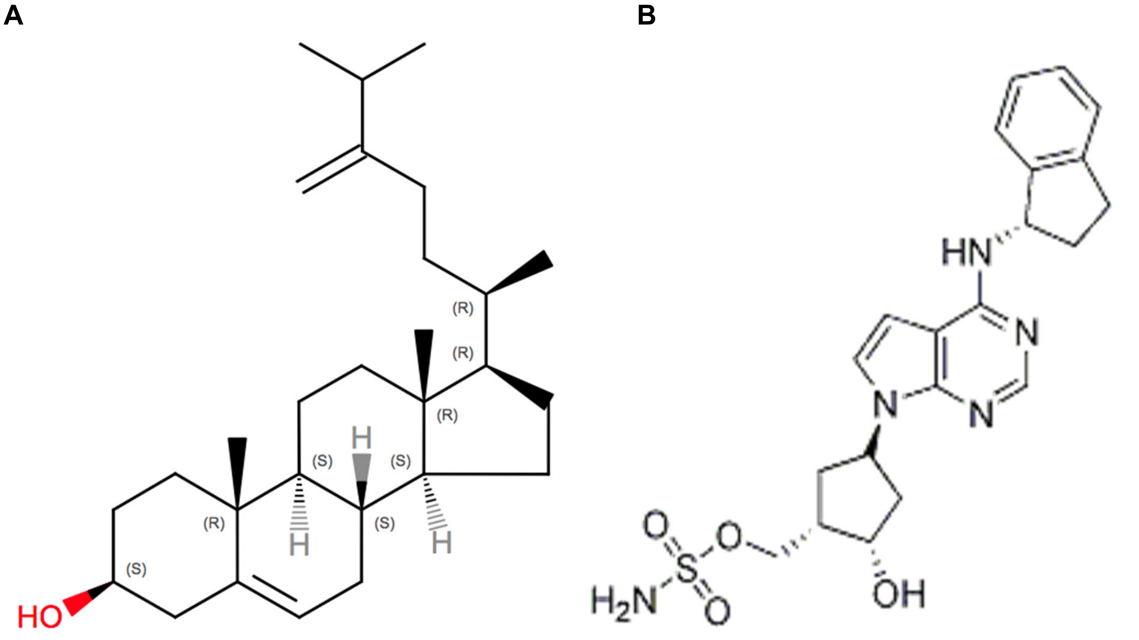


Figure 7.35. Comparison between 94R and peveonedistat
(A) Predicted 94R ligand from the FunFOLD3 component of the IntFOLD6 server and **(B)** pevonedistat which inhibits NAE and due to ORF8 potentially also interacting with this enzyme it has been propsed as a ligand

C1910, ORF7b is the final CASP Commons target and one of the ten proteins that significantly suppressed IFN-alpha signalling, along with nsp1, nsp6, nsp7, nsp13, nsp14, ORF3a, M, ORF6 and ORF7a (Xia, Cao, Xie, Zhang, Chen, Wang, Vineet D. Menachery, *et al.*, 2020) in terms of sequence identity it shares 85.4% with SARS-CoV and 97.2% similarity (Yoshimoto, 2020).

In vitro studies found that ORF7b, along with some of the aforementioned proteins inhibited STAT2 phosphorylation by 33-50% (Xia, Cao, Xie, Zhang, Chen, Wang, Vineet D. Menachery, *et al.*, 2020). Consistent with inhibition of STAT2 phosphorylation, ORF7b suppressed nuclear translocation of STAT1 during IFN-I signalling (Xia, Cao, Xie, Zhang, Chen, Wang, Vineet D. Menachery, *et al.*, 2020) and is localised in the Golgi compartment (Yoshimoto, 2020).

The role of STAT and IFN signalling in pathogen defence has been discussed previously in this chapter and Figure 7.25 has shown the antagonist role of IFN production by ORF7b and other proteins.

As can be expected with viruses, mutations can occur and a SARS-CoV-2 variant with a 382-nucleotide deletion was detected in a cluster of cases in Singapore between January-February 2020. The deletion truncates ORF7b and removes the ORF8 transcription-regulatory sequence and has not been detected after March 2020 (Young *et al.*, 2020).

In terms of ligands, FunFOLD3 predicted chlorophyll A, which is a metalloporphyrin with Mg^{2+} at its core. Many organometallic compounds derived from chlorophyll A or B are approved for human consumption, for example, sodium copper chlorophyllin is promoted for its use an antibacterial and anti-viral agent the Mg^{2+} has been replaced with Cu^{2+}(Clark &

Taylor-Robinson, 2020). Zinc pheophorbide *a* (ZnPh), a chlorophyll derivative for which Zn^{2+} is replaced with natural occurring Mg^{2+} and are harnessed to produce a cytotoxic effect in the treatment of cancer cells (Clark & Taylor-Robinson, 2020). Zn^{2+} ions attached to this tetrapyrrole derivative can pass through cell membranes as it is water-soluble but no localisation of ZnPh has been shown in mitochondria, therefore it is unlikely to impair the function of mitochondria in healthy cells (Clark & Taylor-Robinson, 2020). Free Zn^{2+} appears to promote an antiviral effect and demonstrates accumulation of Zn^{2+} in human lung tissue and zinc supplementation and using ionophore such as pyrithione effectively impairs RNA replication by human coronaviruses leading to improved treatment outcomes (Zhang & Liu, 2020). ZnPh is non-toxic to humans, and there is a strong likelihood that this compound could act as a carrier molecule for Zn^{2+} to trigger anti-viral response that impairs SARS-CoV-2 replication and this provide an novel therapeutic option for COVID-19 (Clark & Taylor-Robinson, 2020).

The role/function of each CASP Commons protein, based on the FunFOLD3 ligand and GO term prediction or the templates which had biologically relevant ligands are given below:

1. Nsp2 (C1901) – serves as a hydrolase, potentially a serine hydrolase to contribute to the viruses cell cycle.
2. Nsp4 (C1902) – most common predicted template is for transport. However, a template related to blood clotting is particularly interesting due to the a porphyrin (chlorophyll A) being predicted.
3. Nsp6 (C1903) – limited GO term predictions for this target, molecular function suggests role in catalytic activity or a biological process of metabolic process.
4. PL-PRO (C1904) – based on the predicted ligand (GDP) and the templates containing this ligand, most likely has a role in signalling/lipid-binding. This would also compliment the amino acid predictions.

5. ORF3a (C1905) – role in binding and based on the predicted GO terms for cellular component is a plasma membrane or an integral component of membrane.
6. Membrane protein (C1906) – a transmembrane protein
7. ORF6a (C1907) – based on both the predicted ligand (haeme) and the GO predicted terms for clotting involvement and iron ion binding has a role in blood clot formation.
8. ORF8 (C1908) – no GO terms related to molecular function where predicted. However, GO terms related to cell adhesion and innate immune response and inflammatory response. Therefore, this protein could potentially bind to a host cell and as a result of this binding or adhesion causes an immune response and this in turn causes the production of an inflammatory response.
9. ORF10 (C1909) – a hydrolase, in particular a viral hydrolase or oxidoreductase.
10. ORF7b (C1910) – A porphyrin was predicted (chlorophyll A) so may have a role in blood clotting or blood coagulation, at this stage it is a speculation. However, there is data to support it. GO predicted terms around ATP, suggest a role in energy-carrying, so potentially involved in reactions within a host cell.

In comparison, the roles of the different proteins as per information in literature are outlined below and are listed in CASP Commons target order (refer to Table 7.1 for further details and other proteins/domains). For completion, the proteins which were not part of CASP Commons have also been included:

1. The spike protein (a glycoprotein) mediates the attachment of the virus to the host cell by the angiotensin-converting enzyme (ACE2) receptor on host cell surface (Yoshimoto, 2020).
2. Non-structural polyproteins are expressed by ORF1ab and consists of 16 non-structural polyproteins
3. Nsp2 (C1901) and binds to two host proteins: prohibitin 1 and prohibitin 2 (PHB1 and PHB2)(Yoshimoto, 2020). PHB1 and PHB2 proteins are known to play roles in cell

cycle progression, cell migration, cellular differentiation, apoptosis and mitochondrial biogenesis (Yoshimoto, 2020). Downstream effects of this interaction inhibit the production of interferon and proinflammatory cytokine IL-6 (UniProt Consortium, 2019).

4. Nsp4 (C1902) interacts with Nsp3 and possibly host proteins to confer a role related to membrane rearrangement (Yoshimoto, 2020).
5. Nsp6 (C1903) generates autophagosomes from the ER. Autophagosomes facilitate assembly of replicase proteins (Yoshimoto, 2020). Additionally, nsp6 limited autophagosomes/lysosome expansion, which in turn prevents autophagosomes from delivering viral components from degradation in lysosomes (Yoshimoto, 2020).
6. PL-PRO (C1904) attenuates host antiviral IFN pathways (Shin *et al.*, 2020).
7. ORF3a (C1905) mediates trafficking of spike protein (Michel *et al.*, 2020).
8. Membrane protein (C1906) is an integral membrane protein that plays an important role in viral assembly (Yoshimoto, 2020). The M protein interacts with nucleocapsid to encapsulate the RNA genome. M protein is expressed by ORF5 (Yoshimoto, 2020).
9. ORF6 (C1907) interacts with nsp8, nsp8 is related to promoting RNA polymerase activity (Yoshimoto, 2020) with downstream activities on interferon signalling (Yoshimoto, 2020).
10. ORF8 (C1908) interacts with a variety of host proteins and causes downstream suppression of IFN (Zinzula, 2020).
11. ORF10 (C1909) interacts with ubiquitin ligases (Gordon *et al.*, 2020).
12. ORF7b (C1910) is localised in the Golgi complex (Yoshimoto, 2020) and suppresses IFN signalling (Xia, Cao, Xie, Zhang, Chen, Wang, Vineet D. Menachery, *et al.*, 2020).

To date, there is only one drug which has been approved for the treatment of SARS-CoV-2, remdesivir, an inhibitor of the viral RNA-dependent, RNA polymerase (Beigel *et al.*, 2020) and is shown below in Figure 7.26

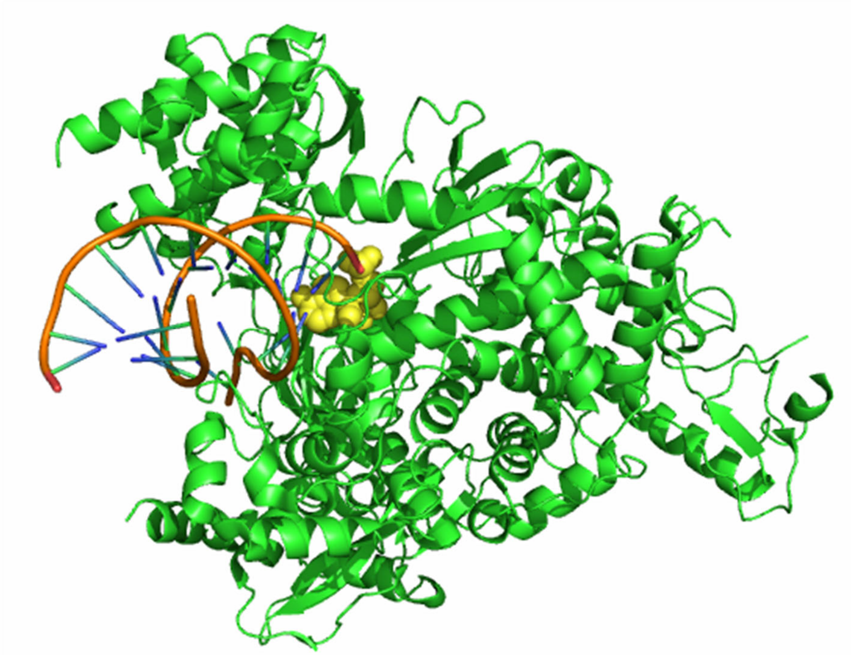


Figure 7.36. PyMOL generated image of SARS-CoV-2 (PDB ID 7bv2) and F86
PDB structure for SARS-CoV-2 (PDB ID 7bv2) shown as cartoon and coloured green, the ligand F86 (remedesivir) is shown as sphere and coloured yellow. RNA is also shown with the ligand F86 within RNA showing the inhibition of activity.

In conclusion, CASP Commons brought together the protein prediction community to understand the role and function of the various SARS-CoV-2 proteins and domains. Results from CASP11, CASP12 and CASP13 showed that FunFOLD3 was able to predict ligands and ligand-binding site residues, with varying results and the benchmarking in CAFA3 showed that GO terms, although not specific enough can be used in infer the role and function of a protein. Therefore, FunFOLD3 was used as the method of choice to determine a) if there is a role for FunFOLD3 outside the CASP competitions and in the application of novel proteins and b) aid in the understanding of the proteins and domains that make the SARS-CoV-2 virus. Stand-alone FunFOLD3 predicted ligands for six of the 32 targets released for prediction by CASP Commons. Ligands together with templates could potentially provide insight in the potential function of each protein such as signalling protein, cell adhesion proteins and immune system. Taken together the ligands and template functions could be used to explore potential drug targeting strategies and ultimately contribute to the understanding of a novel virus.

Chapter 8: Synopsis and Next Directions

8.1 Introduction

The aims of this thesis were to objectively measure the performance of FunFOLD3, in structure-function prediction using targets from the Critical Assessment of protein Structure Prediction (CASP11, 12 and 13) experiments. An additional aim was to evaluate the GO term annotations using FunFOLDQ, in the Critical Assessment of Functional Annotation (CAFA3) experiment. Each of these experiments provided an independent accuracy benchmarks, which could be used to determine strengths and weaknesses of the methods and they provided an indication where improvements could be made. Protein-ligand docking was next investigated as the method of choice to refine FunFOLD3 predictions and thereby improve the identification of likely ligand-binding residues. Thus, chapters 3, 4 and 5 can be thought of assessing the state-of-the-art and determination of the known. Chapter 6 is "doing it better" and chapter 7 is going into the unknown and a highly relevant application of FunFOLD3.

The main findings of each of the different algorithms is provided below:

1. **FunFOLD3:** The main findings with FunFOLD3 is the inconsistency with the prediction of ligand-binding site residues. As shown in Chapter 3, there is a variation in the MCC and BDT scores obtained across the protein targets. This is directly related to variation of the templates for the ligands. For example, if a specific ligand is predicted and is related to only a few templates (e.g. PLP and aminotransferase) then the MCC and BDT scores are likely to be on the higher end of the scale (>0.70). If the ligand is ubiquitous (e.g. metal ion) then it is likely to bind to a large number of proteins and therefore not specific. Thereby, resulting in lower MCC and BDT scores
2. **IntFOLD:** The main finding with IntFOLD was the clear strength in predicting holo vs apo for ligands and ligand-binding site residues. The main difference between IntFOLD and FunFOLD is the utilisation of templates by FunFOLD3. Templates with

similar folds will bind similar ligands and this aids in the prediction of ligands and ligand-binding residues by FunFOLD3. As holo structures contain the ligand in crystal, there is no need for templates, however with apo structures where the ligand isn't present, templates could be useful in identification of ligands. As FunFOLD3, is the function prediction aspect and utilises this, this limitation has been addressed

3. **FunFOLDQ:** The main finding from FunFOLDQ was the prediction of GO terms were not specific enough, despite being related to the actual GO terms. FunFOLDQ is the only algorithm of the two previously described which focuses solely on the prediction of GO terms to provide insights into a protein's function. This algorithm raised the question of "*are GO terms enough to provide insight into a protein's function?"* This thesis has demonstrated that both ligands and GO terms are valuable to answer questions related to function, particularly with the utilisation in the SARS-CoV-2 virus, which was an example of going into the unknown
4. **FunFOLD3-D:** This was the method developed in the thesis based on the outcomes of FunFOLD3. This method utilised an extra step with docking via AutoDock Vina and had novelty with the inclusion of four different grid box calculation methods. The mian findings were 1) ligand-binding residues can be improved or, impaired depending on the original prediction. Better predictions (MCC or BDT scores of >0.70) are hard to improve further 2) no standard grid box calculation "fits all". The next stage, will be to publish the early findings from this methodology in a book chapter (refer to Appendix 1, Arvinas Book Chapter).

This synopsis will provide an overview of the different aspects of this project and discuss how they interlink with one another. Additionally, this synopsis will highlight the potential contribution made by the application of FunFOLD3 towards functional insights on the less understood SARS-CoV-2 proteins, and finally, the future directions for FunFOLD developments will be discussed.

8.2 Prediction of ligands and ligand-binding residues by FunFOLD3

The determination of ligand-binding sites is an important aspect into providing insights into the elucidation of protein functions, additionally the specific identification of the ligand binding site residues in a protein is also important, because substrate specificity of an enzyme is determined by finer details of the binding site, for example side chain orientation and physiochemical properties (Schwede *et al.*, 2009). Thus, highlighting the need for both ligands and the interacting residues to be predicted.

The FunFOLD server was developed by the McGuffin group and uses an automatic approach for identification of clusters of putative ligands and selection of interacting residues (Roche, Tetchner and McGuffin, 2011). Previous versions of FunFOLD have been benchmarked in Critical Assessment of protein Structure Predictions since CASP9 and they have featured among the top ranked methods. Despite the good performance of FunFOLD, relative to competing methods, the accuracy of predicted ligand-binding sites varies greatly, as has been seen in Chapter 3 and also the accuracy of GO term prediction as has been shown in Chapter 5. Furthermore, a limitation of earlier versions of FunFOLD was that only one ligand-binding site could be predicted per target, even if the observed target has multiple ligand-binding sites, as was seen with CASP11 target T0845 (Figure S.10). A further consideration is the quality of the 3D model for which the ligand-binding prediction is based. It could be assumed, that the better the 3D model quality the better the outcome of the ligand predictions. This general trend was seen in CASP11 and shown in Figure 3.2. However, in CASP12, as shown in Figure 3.3., the better 3D models, did not necessarily always lead to better BDT and MCC scores. In fact, the poorer scoring 3D models, as shown by the TM-score scored the best BDT and MCC scores (e.g. T0916). CASP12 target T0912 was a case in point and it highlighted the need to focus on the ligand-binding space; T0912 was modelled correctly, but the true ligand-binding site, may not have been well modelled, as it was incorrectly predicted to be in a different region of the protein. Thus, these results pointed towards two possible areas for the improvement of prediction accuracy 1) to improve the model around the ligand-binding site, or 2) to improve the ligand-pose by docking. Option two was later explored in this thesis (Chapter 6 and section 6.4).

8.3 Prediction of function using GO terms by FunFOLDQ

Chapter 4 of this thesis aimed to explore higher throughput methods for protein function prediction and utilised GO terms which provides a classification of function based on the three categories - molecular function, biological process and cellular component (Ashburner *et al.*, 2000). Correct prediction of a biologically-relevant ligand might not provide this level of detail and the function of a protein can cover a plethora of aspects related to its role and is not well defined (Wrzeszczynski *et al.*, 2003). One definition of function is "*a complex phenomenon that is associated with many mutually over-lapping levels: biochemical, cellular, organism-mediated, developmental and physiological"*(Wrzeszczynski *et al.*, 2003). These overlapping levels are intertwined in complex ways (Wrzeszczynski *et al.*, 2003) and can explain why GO term prediction, can serve to better understand the role of protein prediction.

The results of FunFOLDQ in CAFA3 showed how complex and nuanced the prediction of GO terms can be, despite the requirement for absolute GO term predictions in CAFA3. GO terms can be thought of as layers, and, as knowledge is gained about the function of a protein, the layer becomes deeper. CAFA3 required the prediction of the latest layer, whereas in some instances FunFOLDQ predicted a GO term, which was lower down the annotation, but was nevertheless related to the latest term when looking at ancestry of terms on a hierarchical graph, and as a result the prediction was deemed incorrect. The "top layer" can be viewed as the most recent knowledge and compliments previous knowledge as opposed to deeming this knowledge incorrect, which as per CAFA3 this would be the case.

The difficulty in predicting GO terms was highlighted in CASP6, were the organisers observed the prediction of EC numbers and GO terms were not suitable in assessing predictions. The “problem” with CASP6 was availability of newer GO terms being identified with targets after the end of the CASP competition (López *et al.*, 2007). GO terms and relationships are being updated constantly so that new functions, error corrections and amended definitions can be included (Friedberg, 2006). Given the dynamic nature of GO term prediction and the better performance of FunFOLD3 in a blind competition, refinement of FunFOLD3 and ligand-binding site prediction was favoured. However, the recent global pandemic (which occurred during the last year of this PhD project) highlighted both the importance of predicting ligands for the investigation of potential drug targets, but also the importance of GO term prediction for understanding the plethora of functions of the SARS-CoV-2 virus.

8.4 The development of FunFOLD3-D

In Chapter 6, the focus was on the improvement of ligand binding site residues using AutoDock Vina. The rationale for using docking was that whilst FunFOLD3 is able to potentially provide the general area for the ligand, docking may improve the rotation of the ligand within the ligand-binding space and therefore improve the subsequent ligand-binding residue prediction accuracy. As was seen in Chapter 6, this was carried out with somewhat varying degrees of success, however unlike COACH-D (Wu *et al.*, 2018) in which the mean MCCs were very similar to COACH (0.66 versus 0.67), in comparison for FunFOLD3-D, the mean MCC score improved from 0.25 to 0.34 and the mean BDT score improved from 0.33 to 0.35 thereby, FunFOLD3-D was shown to improve the ligand-binding for some CASP targets and is therefore a potential improvement to the method, which is worth considering further. Note the MCC and BDT scores were compared against targets which were docked which was a total of 21 targets. It is worth noting, that whilst the methods are tested on different data sets, this nevertheless served to show that there is potential for improvements to be made.

A further improvement to FunFOLD3-D may be a consensus method for the ligand-binding residues. Consensus methods are used for functional prediction (Xie & Hwang, 2012) (e.g. DoGSite (Volkamer *et al.*, 2010)) and it would seem plausible that a similar approach could be applied for FunFOLD3-D where the most commonly predicted residues across the grid box calculations are selected as the key ligand-binding residues. However, a small scale test was attempted for targets (T0819, T0849 and T0916) and no improvements were made versus the top-scoring grid box calculation. An additional aspect that could potentially be explored is finer grained alternative grid box thresholds, e.g. could 12% or 15% be more suitable.

The template-based modelling approach for prediction of protein structure is a really powerful approach when proper templates can be found (Lance, Deane and Wood, 2010). However, in instances where there are limited templates, could the role of docking be to improve the location of ligand-binding sites? FunFOLD3-D development and testing explored three aspects of docking (i) can an incorrect ligand be rotated to a portion of protein which contains the correct ligand-binding space (ii) can a correct ligand be docked to improve the ligand-binding residues and (iii) can an already good prediction be improved even further. Based on our analysis, we found that a good predictions of >0.7 could not be improved further, but differing ligands could be rotated to a portion of the correct ligand-binding space and for matching ligands, and the ligand-binding residues could potentially be improved further. It is therefore worth considering if docking can be used in future FunFOLD pipelines, as a way of improving the modelling specifically around the ligand-binding space, by potentially providing a better rotation of the ligand.

Template based modelling generally consist of, four steps: 1) detect template, 2) align sequence onto template, 3) build model and 4) refine model (Haddad, Adam and Heger, 2020). Docking could be added as an extra stage when investigating functional prediction from sequence via structure. Thus, the procedure would become: 1) detect template, 2) align sequence, 3) build model, 4) refine model 5) predict ligand binding site and 6) refine ligand-binding space using docking. One of the limitations with FunFOLD3-D, is the utilisation in novel protein prediction. Another limitation, is there is currently no scoring method to predict which one of the nine models might have the best ligand-binding site predictions, thereby a quality scoring method could be developed to overcome this, similar to the ModFOLD element on the IntFOLD server.

8.5 The novel SARS-CoV-2 virus and the contribution of FunFOLD3

The recent global pandemic provided an additional opportunity for the structure and function prediction community to come together to share knowledge in order to produce 3D models for the domains of the SARS-CoV-2 virus, in addition to just competing against one another in the biennial CASP competition. Additionally, for the first time in this thesis the different aspects of FunFOLD3 have come together to aid in the understanding of this novel virus. For example, and where possible, ligands and similarity to other ligands were used in order to improve understanding of the clinical pathology of the virus. This was shown with the predictions for porphyrin binding sites (e.g. haeme and the similar structure of chlorophyll A). SARS-CoV-2 has shown to have effects on the haematopoietic system and blood clots which develop as a result of SARS-CoV-2 infection have been widely reported and described in literature (Biswas *et al.*, 2021). Understanding the putative porphyrin binding sites in SARS-CoV-2 proteins could potentially provide an insight into the less well understood mechanism f the virus. The second aspect of FunFOLD3, which has proved to be insightful are the predicted templates. FunFOLD3 uses template-based modelling and works on the concept that, ligand containing templates from the PDB with the same folds, may contain similar binding sites. For nsp4 (C1902), the FunFOLD3 component of IntFOLD6 predicted porphyrin binding and of the templates which were deemed to have biologically relevant ligands had blood clotting as a role/function associated with this target. Therefore, whilst the ligand itself, has not provided the clearest indication of the function of the protein, thinking more broadly around ligand similarity and the templates and their role has potentially provided interesting information around the role/function. The final aspect of FunFOLD3 is GO term prediction and for SARS-CoV-2, there were some GO term predictions which has supported the available literature on the virus. For example, for ORF8 (C1908) there were predicted GO terms for biological process around innate immune response and literature information has suggested ORF8 interacts with a variety of host proteins and causes suppression of IFN, a key cytokine involved in immune response (Zhang et al., 2020). In summary, the timely analysis of SARS-CoV-2 proteins served as a useful case study for the application of FunFOLD3, bringing everything together (e.g. ligand-binding site prediction and GO term prediction) It also demonstrated how FunFOLD3 can contribute to the wider community and despite the difficulties of benchmarking using GO terms, the GO term prediction by FunFOLD3 provided some insights into the roles of the lesser understood SARS-CoV-2 proteins.

Overall, there are two principle ways to determine the functions of a protein, by either predicting interacting partners such as ligands or by prediction of GO terms. The need for two different aspects to determine function complimented the understanding of this impactful novel virus.

8.6 Summary of Future Directions

In order to enhance the benchmarking of the FunFOLD3 results and develop a new FunFOLD pipeline, which can be incorporated into the IntFOLD6 server, the next step will be to investigate the AutoDock Vina method further, as docking was shown to have promise in improving ligand-binding residues as shown by BDT and MCC scores. One aspect is to explore alternative grid box thresholds, to determine if there is a “one size fit all” or if the size of the ligand should determine the grid box threshold. Additionally, AutoDock Vina produces up to nine different models and each model has to be assessed in order to determine if there has been an improvement in MCC or BDT scores. One method to assess the quality of each model could be to incorporate FunFOLDQA to act a benchmarking tool to select the most improved docking models. Incorporation of refinement restricted to the binding site location could be an additional direction for this project. With ligand-binding site prediction, refinement of the entire structure might not be necessary and therefore refinement focusing on the binding-site location and then docking around the refined binding site.

Other methods for protein-ligand docking could be explored such as PLANTS (Korb, Stützle and Exner, 2006), which provides a scoring function to find a low energy ligand conformation. This method could prove helpful as it could provide a way of scoring the models, in the form of a quality check, as opposed to manually checking the AutoDock Vina outputs. An additional aspect to explore is the prediction of the folding of the protein models particularly within flexible loops, which appears to be a universal problem across all of the predicted protein models, especially because flexible loops are known to be important in ligand-interactions.

A more longer-term aspirational goal of FunFOLD is the utilisation of structure-function prediction in playing an important role in understanding the role of proteins in disease. For example, docking was used to optimise drug candidates by examining and modelling molecular interactions between ligands and macromolecules (Kapetanovic, 2008). An example being the designing of a selective ER_{β} agonist, ERB-041 by the Wyeth group (Kapetanovic, 2008). The first stage in optimisation of the potential drug candidate was understanding the ligands binding to ER$_\beta$ receptor and identification of the binding site. Following various substitutions and computational modelling, a selective ER$_\beta$ agonist, ERB-041 with similar affinity but more than 200-fold greater selectivity for ER$_\beta$ than that of 17β-estradiol was identified.

Drug discovery and development is a time and resource consuming processes and therefore there is an ever growing effort to apply computational power in order to streamline drug discovery, design, development and optimisation (Kapetanovic, 2008). Based on the above example it is clear that structure-function prediction and docking, could be methods which can be used to streamline drug discovery in several ways, by either understating the proteins involved in diseases by elucidating their function *in silico* or by docking within the ligand-binding pockets to identify potential drug targets.

This thesis has identified the potential of FunFOLD3 in the prediction of ligands and ligand-binding site residues. Thinking of the bigger picture for FunFOLD3 could be utilisation in high-throughput ligand docking. Molecular docking is central to rational drug design (Souza *et al.*, 2021) and structure-based drug design has been extensively used by pharmaceutical companies and academia research groups to reduce both the time and cost for the discovery of new drugs. The docking aspect of FunFOLD3-D could be utilised in the docking of potential drug candidates, especially with the next step of this methodology is to become fully automated and without manual curation, and with this improvement there is potential for high-throughput.

Another application of FunFOLD3 could be in plastic. Plastics (e.g. polyethylene terephthalate (PET)) are widely used for various applications (Almeida *et al.*, 2019). Improper plastic waste management and difficulty in recycling has meant plastic waste has become an environmental issue. In the past decade, a number of bacterial enzymes capable of degrading PET have been identified (Almeida *et al.*, 2019) and have been identified from bacterium (Almeida *et al.*, 2019). Almeida et al., 2019, utilised protein structure analysis with SWISS-MODEL, and molecular docking with AutoDock Vina in order to order to investigate a class of PETase-like enzymes. AutoDock Vina was used to analyse the likelihood of an enzymes capacity to bind plastics as substrates. There are similarities between FunFOLD3 and SWISS-MODEL, as both methods make use of templates and FunFOLD3-D, utilised AutoDock Vina, therefore this research by Almeida et al., 2019 highlights the potential application of FunFOLD3-D, not only in bacterium but in a highly topical and relevant environmental problem. This therefore demonstrates why one should care about ligand and ligand-binding site prediction.

References

References

10th Community Wide Experiment on the Critical Assessment of techniques for protein structure prediction (2012). Available at: https://predictioncenter.org/casp10/index.cgi?page=format#FN.

Alberts, Bruce; Johnson, Alexander; Lewis, Julian; Raff, Martin; Roberts, Keith; Walter, P. (2002) *Molecular Biology of the Cell*. 4th edn. New York: Garland Sciences.

Alborzi, S.Z., Devignes, M.-D. and Ritchie, D.W. (2017) 'ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains.', *BMC bioinformatics*, 18(1), p. 107. doi:10.1186/s12859-017-1519-x.

Alexandrov, N.N. and Gō, N. (1994) 'Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins', *Protein Science*, 3(6), pp. 866–875. doi:10.1002/pro.5560030601.

Allen, W.J., Balius, T.E., Mukherjee, S., Brozell, S.R., Moustakas, D.T., Lang, P.T., Case, D.A., Kuntz, I.D. and Rizzo, R.C. (2015) 'DOCK 6: Impact of new features and current docking performance.', *Journal of computational chemistry*, 36(15), pp. 1132–56. doi:10.1002/jcc.23905.

Almeida, E.L., Carrillo Rincón, A.F., Jackson, S.A. and Dobson, A.D.W. (2019) 'In silico Screening and Heterologous Expression of a Polyethylene Terephthalate Hydrolase (PETase)-Like Enzyme (SM14est) With Polycaprolactone (PCL)-Degrading Activity, From the Marine Sponge-Derived Strain Streptomyces sp. SM14', *Frontiers in Microbiology*, 10, p. 2187. doi:10.3389/fmicb.2019.02187.

AlQuraishi, M. (2019) 'AlphaFold at CASP13.', *Bioinformatics (Oxford, England)*, 35(22), pp. 4862–4865. doi:10.1093/bioinformatics/btz422.

AlQuraishi, M. (2021) 'Protein-structure prediction revolutionized', *Nature*, 596(7873), pp. 487–488. doi:10.1038/d41586-021-02265-4.

Altman, R.B. (2016) *Functional Assessment of CASP predictions*. Available at: http://predictioncenter.org/casp12/doc/presentations/CASP12_function_altman.pdf.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) 'Basic local alignment search tool.', *Journal of molecular biology*, 215(3), pp. 403–10. doi:10.1016/S0022-2836(05)80360-2.

Anderson, K.A., Madsen, A.S., Olsen, C.A. and Hirschey, M.D. (2017) 'Metabolic control by sirtuins and other enzymes that sense NAD+, NADH, or their ratio.', *Biochimica et biophysica acta. Bioenergetics*, 1858(12), pp. 991–998. doi:10.1016/j.bbabio.2017.09.005.

Anfinsen, C.B. (1973) 'Principles that Govern the Folding of Protein Chains', *Science*, 181(4096), pp. 223 LP – 230. doi:10.1126/science.181.4096.223.

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S. and Ciccozzi, M. (2020) 'COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis', *Journal of Medical Virology*, 92(6), pp. 584–588. doi:10.1002/jmv.25719.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., *et al.* (2000) 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.', *Nature genetics*, 25(1), pp. 25–9. doi:10.1038/75556.

Astuti, I. and Ysrafil (2020) 'Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response', *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4), pp. 407–412. doi:10.1016/j.dsx.2020.04.020.

Atasever, S., Aydın, Z., Erbay, H. and Sabzekar, M. (2019) 'Sample Reduction Strategies for Protein Secondary Structure Prediction', *Applied Sciences*, 9(20), p. 4429. doi:10.3390/app9204429.

Báez-Santos, Y.M., St John, S.E. and Mesecar, A.D. (2015) 'The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds.', *Antiviral research*, 115, pp. 21–38. doi:10.1016/j.antiviral.2014.12.015.

Bai, Z., Cao, Y., Liu, W. and Li, J. (2021) 'The SARS-CoV-2 Nucleocapsid Protein and Its Role in Viral Structure, Biological Functions, and a Potential Target for Drug or Vaccine Mitigation', *Viruses*, 13(6), p. 1115. doi:10.3390/v13061115.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L.G., *et al.* (2021) 'UniProt: the universal protein knowledgebase in 2021', *Nucleic Acids Research*, 49(D1), pp. D480–D489. doi:10.1093/nar/gkaa1100.

Baú, D., Martin, A.J.M., Mooney, C., Vullo, A., Walsh, I. and Pollastri, G. (2006) 'Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins.', *BMC bioinformatics*, 7, p. 402. doi:10.1186/1471-2105-7-402.

Bauer, R., Wilson, J.J., Philominathan, S.T.L., Davis, D., Matsushita, O. and Sakon, J. (2013) 'Structural comparison of ColH and ColG collagen-binding domains from Clostridium histolyticum.', *Journal of bacteriology*, 195(2), pp. 318–27. doi:10.1128/JB.00010-12.

Beigel, J.H., Tomashek, K.M., Dodd, L.E., Mehta, A.K., Zingman, B.S., Kalil, A.C., Hohmann, E., Chu, H.Y., Luetkemeyer, A., Kline, S., Lopez de Castilla, D., Finberg, R.W., Dierberg, K., Tapson, V., Hsieh, L., Patterson, T.F., Paredes, R., Sweeney, D.A., Short, W.R., *et al.* (2020) 'Remdesivir for the Treatment of Covid-19 — Final Report', *New England Journal of Medicine*, 383(19), pp. 1813–1826. doi:10.1056/NEJMoa2007764.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2013) 'GenBank.', *Nucleic acids research*, 41(Database issue), pp. D36-42. doi:10.1093/nar/gks1195.

Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J. and Carlson, H.A. (2007) 'Binding MOAD, a high-quality protein ligand database', *Nucleic Acids Research*, 36(Database), pp. D674–D678. doi:10.1093/nar/gkm911.

Bera, I. and Ray, S. (2009) 'A study of interface roughness of heteromeric obligate and non-obligate protein-protein complexes.', *Bioinformation*, 4(5), pp. 210–5. doi:10.6026/97320630004210.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R. and Ben-Tal, N. (2004) 'ConSeq: the identification of functionally and structurally important residues in protein sequences.', *Bioinformatics (Oxford, England)*, 20(8), pp. 1322–4. doi:10.1093/bioinformatics/bth070.

Berke, G. (1995) 'The CTL's kiss of death', *Cell*, 81(1), pp. 9–12. doi:10.1016/0092-8674(95)90365-8.

Berman, H., Henrick, K. and Nakamura, H. (2003) 'Announcing the worldwide Protein Data Bank.', *Nature structural biology*, 10(12), p. 980. doi:10.1038/nsb1203-980.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28(1), pp. 235–242. doi:10.1093/nar/28.1.235.

Berntsson, O., Diensthuber, R.P., Panman, M.R., Björling, A., Gustavsson, E., Hoernke, M., Hughes, A.J., Henry, L., Niebling, S., Takala, H., Ihalainen, J.A., Newby, G., Kerruth, S., Heberle, J., Liebi, M., Menzel, A., Henning, R., Kosheleva, I., Möglich, A., *et al.* (2017) 'Sequential conformational transitions and α-helical supercoiling regulate a sensor histidine kinase.', *Nature communications*, 8(1), p. 284. doi:10.1038/s41467-017-00300-5.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. and Schwede, T. (2017) 'Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology', *Scientific Reports*, 7(1), p. 10480. doi:10.1038/s41598-017-09654-8.

Biegert, A. and Söding, J. (2008) 'De novo identification of highly diverged protein repeats by probabilistic consistency', *Bioinformatics*, 24(6), pp. 807–814. doi:10.1093/bioinformatics/btn039.

Bisson, S.A., Page, A.-L. and Ganem, D. (2009) 'A Kaposi's sarcoma-associated herpesvirus protein that forms inhibitory complexes with type I interferon receptor subunits, Jak and STAT proteins, and blocks interferon-mediated signal transduction.', *Journal of* *virology*, 83(10), pp. 5056–66. doi:10.1128/JVI.02516-08.

Biswas, S., Thakur, V., Kaur, P., Khan, A., Kulshrestha, S. and Kumar, P. (2021) ‘Blood clots in COVID-19 patients: Simplifying the curious mystery.’, *Medical hypotheses*, 146, p. 110371. doi:10.1016/j.mehy.2020.110371.

Boson, B., Legros, V., Zhou, B., Siret, E., Mathieu, C., Cosset, F.-L., Lavillette, D. and Denolly, S. (2020) ‘The SARS-CoV-2 envelope and membrane proteins modulate maturation and retention of the spike protein, allowing assembly of virus-like particles.’, *The Journal of biological chemistry*, 296, p. 100111. doi:10.1074/jbc.RA120.016175.

Bourgeois, D. and Royant, A. (2005) ‘Advances in kinetic protein crystallography’, *Current Opinion in Structural Biology*, 15(5), pp. 538–547. doi:https://doi.org/10.1016/j.sbi.2005.08.002.

Boyle, A.L. (2018) ‘3 - Applications of de novo designed peptides’, in Koutsopoulos, S. (ed.) *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*. Woodhead Publishing, pp. 51–86. doi:https://doi.org/10.1016/B978-0-08-100736-5.00003-X.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M.S., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E.M. and Baker, D. (2003) ‘Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation’, *Proteins: Structure, Function, and Bioinformatics*, 53(S6), pp. 457–468. doi:https://doi.org/10.1002/prot.10552.

Brown, B.L., Kardon, J.R., Sauer, R.T. and Baker, T.A. (2018) ‘Structure of the Mitochondrial Aminolevulinic Acid Synthase, a Key Heme Biosynthetic Enzyme.’, *Structure (London, England : 1993)*, 26(4), pp. 580-589.e4. doi:10.1016/j.str.2018.02.012.

Brünger, A.T. (1997) ‘X-ray crystallography and NMR reveal complementary views of structure and dynamics’, *Nature structural biology*, 4 Suppl, p. 862—865. Available at: http://europepmc.org/abstract/MED/9377160.

Brylinski, M. and Feinstein, W.P. (2013) ‘eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands’, *Journal of Computer-Aided Molecular Design*, 27(6), pp. 551–567. doi:10.1007/s10822-013-9663-5.

Buchan, D.W.A. and Jones, D.T. (2019) ‘The PSIPRED Protein Analysis Workbench: 20 years on’, *Nucleic Acids Research*, 47(W1), pp. W402–W407. doi:10.1093/nar/gkz297.

Buchko, G.W., Pulavarti, S.V.S.R.K., Ovchinnikov, V., Shaw, E.A., Rettie, S.A., Myler, P.J., Karplus, M., Szyperski, T., Baker, D. and Bahl, C.D. (2018) ‘Cytosolic expression, solution structures, and molecular dynamics simulation of genetically encodable disulfide-rich de novo designed peptides.’, *Protein science : a publication of the Protein Society*, 27(9), pp. 1611–1623. doi:10.1002/pro.3453.

Buchwald, P. (2010a) ‘Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations.’, *IUBMB life*, 62(10), pp. 724–31. doi:10.1002/iub.383.

Buchwald, P. (2010b) ‘Small-molecule protein-protein interaction inhibitors: Therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations’, *IUBMB Life*, 62(10), pp. 724–731. doi:10.1002/iub.383.

Buenavista, M.T., Roche, D.B. and McGuffin, L.J. (2012) ‘Improvement of 3D protein models using multiple templates guided by single-template model quality assessment’, *Bioinformatics*, 28(14), pp. 1851–1857. doi:10.1093/bioinformatics/bts292.

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H. and Velankar, S. (2017) ‘Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.’, *Methods in molecular biology (Clifton, N.J.)*, 1607, pp. 627–641. doi:10.1007/978-1-4939-7000-1_26.

Butler, A., Helliwell, M. V., Zhang, Y., Hancox, J.C. and Dempsey, C.E. (2020) ‘An Update on the Structure of hERG’, *Frontiers in Pharmacology*, 10. doi:10.3389/fphar.2019.01572.

Cai, L., Bai, H., Mahairaki, V., Gao, Y., He, C., Wen, Y., Jin, Y.-C., Wang, Y., Pan, R.L., Qasba, A., Ye, Z. and Cheng, L. (2018) ‘A Universal Approach to Correct Various HBB Gene

Mutations in Human Stem Cells for Gene Therapy of Beta-Thalassemia and Sickle Cell Disease.', *Stem cells translational medicine*, 7(1), pp. 87–97. doi:10.1002/sctm.17-0066.

Cala, O., Guilliere, F. and Krimm, I. (2013) 'NMR-based analysis of protein-ligand interactions', *Analytical and bioanalytical chemistry*, 406. doi:10.1007/s00216-013-6931-0.

Cameron, K.O., Bhattacharya, S.K. and Loomis, A.K. (2014) 'Small molecule ghrelin receptor inverse agonists and antagonists.', *Journal of medicinal chemistry*, 57(21), pp. 8671–91. doi:10.1021/jm5003183.

Cao, H., Walton, J.D., Brumm, P. and Phillips, G.N. (2020) 'Crystal Structure of α-Xylosidase from Aspergillus niger in Complex with a Hydrolyzed Xyloglucan Product and New Insights in Accurately Predicting Substrate Specificities of GH31 Family Glycosidases.', *ACS sustainable chemistry & engineering*, 8(6), pp. 2540–2547. doi:10.1021/acssuschemeng.9b07073.

Cao, R., Wang, Z. and Cheng, J. (2014) 'Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment.', *BMC structural biology*, 14, p. 13. doi:10.1186/1472-6807-14-13.

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M. and Funkhouser, T.A. (2009) 'Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.', *PLoS Computational Biology*. Edited by T. Lengauer, 5(12), p. e1000585. doi:10.1371/journal.pcbi.1000585.

Carta, G., Murru, E., Banni, S. and Manca, C. (2017) 'Palmitic Acid: Physiological Role, Metabolism and Nutritional Implications.', *Frontiers in physiology*, 8, p. 902. doi:10.3389/fphys.2017.00902.

Cary, G.A., Yoon, S.H., Garmendia Torres, C., Wang, K., Hays, M., Ludlow, C., Goodlett, D.R. and Dudley, A.M. (2014) 'Identification and characterization of a drug-sensitive strain enables puromycin-based translational assays in Saccharomyces cerevisiae', *Yeast*, 31(5), pp. 167–178. doi:10.1002/yea.3007.

Casteel, D.E., Smith-Nguyen, E. V, Sankaran, B., Roh, S.H., Pilz, R.B. and Kim, C. (2010) 'A crystal structure of the cyclic GMP-dependent protein kinase I{beta} dimerization/docking domain reveals molecular details of isoform-specific anchoring.', *The Journal of biological chemistry*, 285(43), pp. 32684–8. doi:10.1074/jbc.C110.161430.

Chan, J.F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K.K.-W., Yuan, S. and Yuen, K.-Y. (2020) 'Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan', *Emerging Microbes & Infections*, 9(1), pp. 221–236. doi:10.1080/22221751.2020.1719902.

Chan, J.F.-W., Yuan, S., Kok, K.-H., To, K.K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.C.-Y., Poon, R.W.-S., Tsoi, H.-W., Lo, S.K.-F., Chan, K.-H., Poon, V.K.-M., Chan, W.-M., Ip, J.D., Cai, J.-P., Cheng, V.C.-C., Chen, H., *et al.* (2020) 'A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.', *Lancet (London, England)*, 395(10223), pp. 514–523. doi:10.1016/S0140-6736(20)30154-9.

Chanfreau, G.F. (no date) 'Zinc'ing down RNA polymerase I.', *Transcription*, 4(5), pp. 217–20. doi:10.4161/trns.26594.

Chao, K.L., Kulakova, L. and Herzberg, O. (2017) 'Gene polymorphism linked to increased asthma and IBD risk alters gasdermin-B structure, a sulfatide and phosphoinositide binding protein.', *Proceedings of the National Academy of Sciences of the United States of America*, 114(7), pp. E1128–E1137. doi:10.1073/pnas.1616783114.

Chen, J., Malone, B., Llewellyn, E., Grasso, M., Shelton, P.M.M., Olinares, P.D.B., Maruthi, K., Eng, E.T., Vatandaslar, H., Chait, B.T., Kapoor, T.M., Darst, S.A. and Campbell, E.A. (2020) 'Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex', *Cell*, 182(6), pp. 1560-1573.e13. doi:10.1016/j.cell.2020.07.033.

Chen, X., Cao, R. and Zhong, W. (2019) 'Host Calcium Channels and Pumps in Viral Infections.', *Cells*, 9(1). doi:10.3390/cells9010094.

Chen, Y., Li, Z., Hu, S., Zhang, J., Wu, J., Shao, N., Bo, X., Ni, M. and Ying, X. (2017) 'Gut metagenomes of type 2 diabetic patients have characteristic single-nucleotide polymorphism distribution in Bacteroides coprocola.', *Microbiome*, 5(1), p. 15. doi:10.1186/s40168-017-0232-3.

Chothia, C. (1992) 'One thousand families for the molecular biologist', *Nature*, 357(6379), pp. 543–544. doi:10.1038/357543a0.

Clark, N.F. and Taylor-Robinson, A.W. (2020) 'COVID-19 Therapy: Could a Chlorophyll Derivative Promote Cellular Accumulation of Zn(2+) Ions to Inhibit SARS-CoV-2 RNA Synthesis?', *Frontiers in plant science*, 11, p. 1270. doi:10.3389/fpls.2020.01270.

Clevers, H., Loh, K.M. and Nusse, R. (2014) 'An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control', *Science*, 346(6205), p. 1248012. doi:10.1126/science.1248012.

Consortium, T.U. (2017) *UniProt*. Available at: http://www.uniprot.org/uniprot/A4D126 (Accessed: 9 August 2017).

Cormio, L., De Siati, M., Lorusso, F., Selvaggio, O., Mirabella, L., Sanguedolce, F. and Carrieri, G. (2011) 'Oral L-citrulline supplementation improves erection hardness in men with mild erectile dysfunction.', *Urology*, 77(1), pp. 119–22. doi:10.1016/j.urology.2010.08.028.

Cornillez-Ty, C.T., Liao, L., Yates, J.R., Kuhn, P. and Buchmeier, M.J. (2009) 'Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling', *Journal of Virology*, 83(19), pp. 10314–10318. doi:10.1128/JVI.00842-09.

Cottam, E.M., Maier, H.J., Manifava, M., Vaux, L.C., Chandra-Schoenfelder, P., Gerner, W., Britton, P., Ktistakis, N.T. and Wileman, T. (2011) 'Coronavirus nsp6 proteins generate autophagosomes from the endoplasmic reticulum via an omegasome intermediate.', *Autophagy*, 7(11), pp. 1335–47. doi:10.4161/auto.7.11.16642.

Couturier, M., Roussel, A., Rosengren, A., Leone, P., Stålbrand, H. and Berrin, J.-G. (2013) 'Structural and biochemical analyses of glycoside hydrolase families 5 and 26 β-(1,4)-mannanases from Podospora anserina reveal differences upon manno-oligosaccharide catalysis.', *The Journal of biological chemistry*, 288(20), pp. 14624–35. doi:10.1074/jbc.M113.459438.

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B. and Tramontano, A. (2009) 'Evaluation of template-based models in CASP8 with standard measures.', *Proteins*, 77 Suppl 9, pp. 18–28. doi:10.1002/prot.22561.

Cross, G.H., Reeves, A.A., Brand, S., Popplewell, J.F., Peel, L.L., Swann, M.J. and Freeman, N.J. (2003) 'A new quantitative optical biosensor for protein characterisation', *Biosensors and Bioelectronics*, 19(4), pp. 383–390. doi:https://doi.org/10.1016/S0956-5663(03)00203-3.

Cui, Y., Dong, Q., Hong, D. and Wang, X. (2019) 'Predicting protein-ligand binding residues with deep convolutional neural networks', *BMC Bioinformatics*, 20(1), p. 93. doi:10.1186/s12859-019-2672-1.

Cutting, G.R. (2015) 'Cystic fibrosis genetics: from molecular understanding to clinical application.', *Nature reviews. Genetics*, 16(1), pp. 45–56. doi:10.1038/nrg3849.

Czjzek, M. and Michel, G. (2017) 'Innovating glycoside hydrolase activity on a same structural scaffold.', *Proceedings of the National Academy of Sciences of the United States of America*, 114(19), pp. 4857–4859. doi:10.1073/pnas.1704802114.

Dai, T., Liu, Q., Gao, J., Cao, Z. and Zhu, R. (2011) 'A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information', *BMC Bioinformatics*, 12(14), p. S9. doi:10.1186/1471-2105-12-S14-S9.

Danielson, P.B. (2002) 'The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans.', *Current drug metabolism*, 3(6), pp. 561–97. doi:10.2174/1389200023337054.

Danishuddin, M. and Khan, A.U. (2015) 'Structure based virtual screening to discover putative drug candidates: necessary considerations and successful case studies.', *Methods (San Diego, Calif.)*, 71, pp. 135–45. doi:10.1016/j.ymeth.2014.10.019.

Danley DE (2008) ‘Crystallization to obtain protein-ligand complexes for structure-aided drug design’, *Acta crystallographica. Section D, Biological crystallography*, 62, pp. 569–75. doi:10.1107/S0907444906012601.

Dauter, M. and Dauter, Z. (2011) ‘Deprotonated imidodiphosphate in AMPPNP-containing protein structures.’, *Acta crystallographica. Section D, Biological crystallography*, 67(Pt 12), pp. 1073–5. doi:10.1107/S0907444911046105.

Davies, Jonathan P; Almasy Katherine M; McDonald, Eli F; Plate, L. (2020) ‘Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependenciese’, *bioRxiv* [Preprint]. doi:10.1101/2020.07.13.201517.

Deng, H., Jia, Y. and Zhang, Y. (2018) ‘Protein structure prediction.’, *International journal of modern physics. B*, 32(18). doi:10.1142/S021797921840009X.

Deng, X., Hackbart, M., Mettelman, R.C., O’Brien, A., Mielech, A.M., Yi, G., Kao, C.C. and Baker, S.C. (2017) ‘Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages.’, *Proceedings of the National Academy of Sciences of the United States of America*, 114(21), pp. E4251–E4260. doi:10.1073/pnas.1618310114.

Dessailly, B.H., Lensink, M.F., Orengo, C.A. and Wodak, S.J. (2007) ‘LigASite a database of biologically relevant binding sites in proteins with known apo-structures’, *Nucleic Acids Research*, 36(Database), pp. D667–D673. doi:10.1093/nar/gkm839.

Deutzmann, R. (2004) ‘Structural characterization of proteins and peptides.’, *Methods in molecular medicine*, 94, pp. 269–97. doi:10.1385/1-59259-679-7:269.

Dill, K.A., Ozkan, S.B., Shell, M.S. and Weikl, T.R. (2008) ‘The protein folding problem’, *Annual review of biophysics*, 37, pp. 289–316. doi:10.1146/annurev.biophys.37.092707.153558.

Dobson, C.M. (2003) ‘Protein folding and misfolding’, *Nature*, 426(6968), pp. 884–890. doi:10.1038/nature02261.

Dorokhov, Y.L., Sheshukova, E. V., Kosobokova, E.N., Shindyapina, A. V., Kosorukov, V.S. and Komarova, T. V. (2016) ‘Functional role of carbohydrate residues in human immunoglobulin G and therapeutic monoclonal antibodies’, *Biochemistry (Moscow)*, 81(8), pp. 835–857. doi:10.1134/S0006297916080058.

Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) ‘JPred4: a protein secondary structure prediction server’, *Nucleic Acids Research*, 43(W1), pp. W389–W394. doi:10.1093/nar/gkv332.

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L. and Liu, S.-Q. (2016) ‘Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods.’, *International journal of molecular sciences*, 17(2). doi:10.3390/ijms17020144.

Dubey, B.N., Agustoni, E., Böhm, R., Kaczmarczyk, A., Mangia, F., von Arx, C., Jenal, U., Hiller, S., Plaza-Menacho, I. and Schirmer, T. (2020) ‘Hybrid histidine kinase activation by cyclic di-GMP-mediated domain liberation.’, *Proceedings of the National Academy of Sciences of the United States of America*, 117(2), pp. 1000–1008. doi:10.1073/pnas.1911427117.

Dukka, B.K. (2013) ‘Structure-based Methods for Computational Protein Functional Site Prediction.’, *Computational and structural biotechnology journal*, 8, p. e201308005. doi:10.5936/csbj.201308005.

Dunne, M., Denyes, J.M., Arndt, H., Loessner, M.J., Leiman, P.G. and Klumpp, J. (2018) ‘Salmonella Phage S16 Tail Fiber Adhesin Features a Rare Polyglycine Rich Domain for Host Recognition.’, *Structure (London, England : 1993)*, 26(12), pp. 1573-1582.e4. doi:10.1016/j.str.2018.07.017.

Dutta, P., Basu, S. and Kundu, M. (2017) ‘Assessment of semantic similarity between proteins using information content and topological properties of the Gene Ontology graph.’, *IEEE/ACM transactions on computational biology and bioinformatics* [Preprint]. doi:10.1109/TCBB.2017.2689762.

Eddy, S.R. (2004) ‘Where did the BLOSUM62 alignment score matrix come from?’, *Nature Biotechnology*, 22(8), pp. 1035–1036. doi:10.1038/nbt0804-1035.

Ekici, O.D., Paetzel, M. and Dalbey, R.E. (2008) ‘Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration.’, *Protein science : a publication of the Protein Society*, 17(12), pp. 2023–37. doi:10.1110/ps.035436.108.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E. and Finn, R.D. (2019) ‘The Pfam protein families database in 2019’, *Nucleic Acids Research*, 47(D1), pp. D427–D432. doi:10.1093/nar/gky995.

Elia, A.E.H., Wang, D.C., Willis, N.A., Boardman, A.P., Hajdu, I., Adeyemi, R.O., Lowry, E., Gygi, S.P., Scully, R. and Elledge, S.J. (2015) ‘RFWD3-Dependent Ubiquitination of RPA Regulates Repair at Stalled Replication Forks.’, *Molecular cell*, 60(2), pp. 280–93. doi:10.1016/j.molcel.2015.09.011.

Engelking, L.R. (2015) ‘Chapter 4 - Protein Structure’, in Engelking, L.R. (ed.) *Textbook of Veterinary Physiological Chemistry (Third Edition)*. Third Edit. Boston: Academic Press, pp. 18–25. doi:https://doi.org/10.1016/B978-0-12-391909-0.50004-9.

Englander, S.W. and Mayne, L. (2014) ‘The nature of protein folding pathways’, *Proceedings of the National Academy of Sciences*, 111(45), pp. 15873 LP – 15880. doi:10.1073/pnas.1411798111.

Esnouf, R.M., Hamer, R., Sussman, J.L., Silman, I., Trudgian, D., Yang, Z.-R. and Prilusky, J. (2006) ‘Honing the in silico toolkit for detecting protein disorder.’, *Acta crystallographica. Section D, Biological crystallography*, 62(Pt 10), pp. 1260–6. doi:10.1107/S0907444906033580.

Ewing, T.J., Makino, S., Skillman, A.G. and Kuntz, I.D. (2001) ‘DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases.’, *Journal of computer-aided molecular design*, 15(5), pp. 411–28. doi:10.1023/a:1011115820450.

Farhadi, T. (2018) ‘Advances in protein tertiary structure prediction’, *Biomedical and Biotechnology Research Journal (BBRJ)*, 2(1), p. 20. doi:10.4103/bbrj.bbrj_94_17.

Feinstein, W.P. and Brylinski, M. (2015) ‘Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets.’, *Journal of cheminformatics*, 7, p. 18. doi:10.1186/s13321-015-0067-5.

Feller, S.M. and Lewitzky, M. (2012) ‘What’s in a loop?’, *Cell communication and signaling : CCS*, 10(1), p. 31. doi:10.1186/1478-811X-10-31.

Feng, Y., He, X., Yang, Y., Chao, D., Lazarus, L.H. and Xia, Y. (2012) ‘Current research on opioid receptor function.’, *Current drug targets*, 13(2), pp. 230–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22204322.

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) ‘Ligand Depot: a data warehouse for ligands bound to macromolecules.’, *Bioinformatics (Oxford, England)*, 20(13), pp. 2153–5. doi:10.1093/bioinformatics/bth214.

Fischer, J.D., Mayer, C.E. and Söding, J. (2008) ‘Prediction of protein functional residues from sequence by probability density estimation.’, *Bioinformatics (Oxford, England)*, 24(5), pp. 613–20. doi:10.1093/bioinformatics/btm626.

Fisher, J.F. and Mobashery, S. (2016) ‘β-Lactam Resistance Mechanisms: Gram-Positive Bacteria and Mycobacterium tuberculosis.’, *Cold Spring Harbor perspectives in medicine*, 6(5). doi:10.1101/cshperspect.a025221.

Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S. and Olson, A.J. (2016) ‘Computational protein-ligand docking and virtual drug screening with the AutoDock suite.’, *Nature protocols*, 11(5), pp. 905–19. doi:10.1038/nprot.2016.051.

Friedberg, I. (2006) ‘Automated protein function prediction--the genomic challenge’, *Briefings in Bioinformatics*, 7(3), pp. 225–242. doi:10.1093/bib/bbl004.

Frieman, M., Yount, B., Heise, M., Kopecky-Bromberg, S.A., Palese, P. and Baric, R.S. (2007) ‘Severe acute respiratory syndrome coronavirus ORF6 antagonizes STAT1 function by sequestering nuclear import factors on the rough endoplasmic reticulum/Golgi

membrane.', *Journal of Virology*, 81(18), pp. 9812–24. doi:10.1128/JVI.01012-07.
Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E., Francis, P. and Shenkin, P.S. (2004) 'Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.', *Journal of medicinal chemistry*, 47(7), pp. 1739–49. doi:10.1021/jm0306430.
Frolova, E.I., Fayzulin, R.Z., Cook, S.H., Griffin, D.E., Rice, C.M. and Frolov, I. (2002) 'Roles of nonstructural protein nsP2 and Alpha/Beta interferons in determining the outcome of Sindbis virus infection.', *Journal of virology*, 76(22), pp. 11254–64. doi:10.1128/jvi.76.22.11254-11264.2002.
Gallo Cassarino, T., Bordoli, L. and Schwede, T. (2014) 'Assessment of ligand binding site predictions in CASP10.', *Proteins*, 82 Suppl 2, pp. 154–63. doi:10.1002/prot.24495.
Gao, J., Zhang, Q., Liu, M., Zhu, L., Wu, D., Cao, Z. and Zhu, R. (2016) 'bSiteFinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming', *Journal of Cheminformatics*, 8(1), p. 38. doi:10.1186/s13321-016-0149-z.
Gao, X., Qin, B., Chen, P., Zhu, K., Hou, P., Wojdyla, J.A., Wang, M. and Cui, S. (2020) 'Crystal structure of SARS-CoV-2 papain-like protease', *Acta Pharmaceutica Sinica B* [Preprint]. doi:10.1016/j.apsb.2020.08.014.
Gay, S.C., Roberts, A.G. and Halpert, J.R. (2010) 'Structural features of cytochromes P450 and ligands that affect drug metabolism as revealed by X-ray crystallography and NMR.', *Future medicinal chemistry*, 2(9), pp. 1451–68. doi:10.4155/fmc.10.229.
GenBank release notes (2017). Available at: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (Accessed: 1 September 2017).
Gene Ontology Consortium (2015) 'Gene Ontology Consortium: going forward.', *Nucleic acids research*, 43(Database issue), pp. D1049-56. doi:10.1093/nar/gku1179.
Gene Ontology Consortium (no date). Available at: http://www.geneontology.org (Accessed: 1 May 2017).
Geng, H., Liu, Y.-M., Chan, W.-S., Lo, A.W.-I., Au, D.M.-Y., Waye, M.M.-Y. and Ho, Y.-Y. (2005) 'The putative protein 6 of the severe acute respiratory syndrome-associated coronavirus: expression and functional characterization.', *FEBS letters*, 579(30), pp. 6763–8. doi:10.1016/j.febslet.2005.11.007.
Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R. and Whalen, K.L. (2015) 'Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks.', *Biochimica et biophysica acta*, 1854(8), pp. 1019–37. doi:10.1016/j.bbapap.2015.04.015.
Ghersi, D. and Sanchez, R. (2009) 'EasyMIFs and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures', *Bioinformatics*, 25(23), pp. 3185–3186. doi:10.1093/bioinformatics/btp562.
Ghersi, D. and Sanchez, R. (2011) 'Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures', *Journal of Structural and Functional Genomics*, 12(2), pp. 109–117. doi:10.1007/s10969-011-9110-6.
Giri, R., Bhardwaj, T., Shegane, M., Gehi, B.R., Kumar, P., Gadhave, K., Oldfield, C.J. and Uversky, V.N. (2020) 'Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses', *Cellular and Molecular Life Sciences* [Preprint]. doi:10.1007/s00018-020-03603-x.
Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) 'ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information.', *Bioinformatics (Oxford, England)*, 19(1), pp. 163–4. doi:10.1093/bioinformatics/19.1.163.
Glitsch, H.G. (2001) 'Electrophysiology of the sodium-potassium-ATPase in cardiac cells.', *Physiological reviews*, 81(4), pp. 1791–826. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11581502.

Gong, Q., Ning, W. and Tian, W. (2016) ‘GoFDR: A sequence alignment based method for predicting protein functions.’, *Methods (San Diego, Calif.)*, 93, pp. 3–14. doi:10.1016/j.ymeth.2015.08.009.

Gonzalez, F.J. and Gelboin, H. V (1992) ‘Human cytochromes P450: evolution and cDNA-directed expression.’, *Environmental health perspectives*, 98, pp. 81–5. doi:10.1289/ehp.929881.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O’Meara, M.J., Rezelj, V. V., Guo, J.Z., Swaney, D.L., Tummino, T.A., Hüttenhain, R., Kaake, R.M., Richards, A.L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., *et al.* (2020) ‘A SARS-CoV-2 protein interaction map reveals targets for drug repurposing’, *Nature*, 583(7816), pp. 459–468. doi:10.1038/s41586-020-2286-9.

Guarienti, M., Giacopuzzi, E., Gianoncelli, A., Sigala, S., Spano, P., Pecorelli, S., Pani, L. and Memo, M. (2015) ‘Computational and functional analysis of biopharmaceutical drugs in zebrafish: Erythropoietin as a test model.’, *Pharmacological research*, 102, pp. 12–21. doi:10.1016/j.phrs.2015.09.004.

Gucinski, G.C., Michalska, K., Garza-Sánchez, F., Eschenfeldt, W.H., Stols, L., Nguyen, J.Y., Goulding, C.W., Joachimiak, A. and Hayes, C.S. (2019) ‘Convergent Evolution of the Barnase/EndoU/Colicin/RelE (BECR) Fold in Antibacterial tRNase Toxins.’, *Structure (London, England : 1993)*, 27(11), pp. 1660-1674.e5. doi:10.1016/j.str.2019.08.010.

Guerrero, A., Guiho, R., Herranz, N., Uren, A., Withers, D.J., Martínez-Barbera, J.P., Tietze, L.F. and Gil, J. (2020) ‘Galactose-modified duocarmycin prodrugs as senolytics.’, *Aging cell*, 19(4), p. e13133. doi:10.1111/acel.13133.

Guex, N. and Peitsch, M.C. (1997) ‘SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.’, *Electrophoresis*, 18(15), pp. 2714–23. doi:10.1002/elps.1150181505.

Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) ‘Fpocket: An open source platform for ligand pocket detection’, *BMC Bioinformatics*, 10(1), p. 168. doi:10.1186/1471-2105-10-168.

Gunalan, V., Mirazimi, A. and Tan, Y.-J. (2011) ‘A putative diacidic motif in the SARS-CoV ORF6 protein influences its subcellular localization and suppression of expression of co-transfected expression constructs.’, *BMC research notes*, 4, p. 446. doi:10.1186/1756-0500-4-446.

Gundersen, G.W., Jones, M.R., Rouillard, A.D., Kou, Y., Monteiro, C.D., Feldmann, A.S., Hu, K.S. and Ma’ayan, A. (2015) ‘GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions.’, *Bioinformatics (Oxford, England)*, 31(18), pp. 3060–2. doi:10.1093/bioinformatics/btv297.

Haber, E. and Anfinsen, C.B. (1962) ‘Side-chain Interactions Governing the Pairing of Half-cystine Residues in Ribonuclease’, *Journal of Biological Chemistry*, 237(6), pp. 1839–1844. doi:10.1016/S0021-9258(19)73945-3.

Haddad, Y., Adam, V. and Heger, Z. (2020) ‘Ten quick tips for homology modeling of high-resolution protein 3D structures’, *PLOS Computational Biology*. Edited by F. Ouellette, 16(4), p. e1007449. doi:10.1371/journal.pcbi.1007449.

Hafez, I.M. and Cullis, P.R. (2000) ‘Cholesteryl hemisuccinate exhibits pH sensitive polymorphic phase behavior’, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1463(1), pp. 107–114. doi:10.1016/S0005-2736(99)00186-8.

Haidar, C. and Jeha, S. (2011) ‘Drug interactions in childhood cancer’, *The Lancet Oncology*, 12(1), pp. 92–99. doi:10.1016/S1470-2045(10)70105-4.

Haines, R.J., Pendleton, L.C. and Eichler, D.C. (2011) ‘Argininosuccinate synthase: at the center of arginine metabolism.’, *International journal of biochemistry and molecular biology*, 2(1), pp. 8–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21494411.

Halgren, T.A. (2009) ‘Identifying and Characterizing Binding Sites and Assessing Druggability’, *Journal of Chemical Information and Modeling*, 49(2), pp. 377–389. doi:10.1021/ci800324m.

Hammes, G.G. (2002) ‘Multiple conformational changes in enzyme catalysis.’, *Biochemistry*, 41(26), pp. 8221–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12081470.

Harries, M. and Smith, I. (2002) ‘The development and clinical use of trastuzumab (Herceptin).’, *Endocrine-related cancer*, 9(2), pp. 75–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12121832.

Hartl, F.U., Bracher, A. and Hayer-Hartl, M. (2011) ‘Molecular chaperones in protein folding and proteostasis’, *Nature*, 475(7356), pp. 324–332. doi:10.1038/nature10317.

Hasnain, S.S. (2004) ‘Synchrotron techniques for metalloproteins and human disease in post genome era.’, *Journal of synchrotron radiation*, 11(Pt 1), pp. 7–11. doi:10.1107/s0909049503024166.

Hassan, S.S., Attrish, D., Ghosh, S., Choudhury, P.P., Uversky, V.N., Uhal, B.D., Lundstrom, K., Rezaei, N., Aljabali, A.A.A., Seyran, M., Pizzol, D., Adadi, P., Abd El-Aziz, T.M., Soares, A., Kandimalla, R., Tambuwala, M., Lal, A., Azad, G.K., Sherchan, S.P., *et al.* (2020) ‘Notable sequence homology of the ORF10 protein introspects the architecture of SARS-COV-2’, *bioRxiv*, p. 2020.09.06.284976. doi:10.1101/2020.09.06.284976.

He, J., Hu, H.-J., Harrison, R., Tai, P.C. and Pan, Y. (2006) ‘Rule generation for protein secondary structure prediction with support vector machines and decision tree’, *IEEE Transactions on NanoBioscience*, 5(1), pp. 46–53. doi:10.1109/TNB.2005.864021.

Hendrickson, W.A. (1999) ‘Maturation of MAD phasing for the determination of macromolecular structures’, *Journal of Synchrotron Radiation*, 6(4), pp. 845–851. doi:10.1107/S0909049599007591.

Hendry, G.A. and Jones, O.T. (1980) ‘Haems and chlorophylls: comparison of function and formation.’, *Journal of medical genetics*, 17(1), pp. 1–14. doi:10.1136/jmg.17.1.1.

Henikoff, S. and Henikoff, J.G. (1992) ‘Amino acid substitution matrices from protein blocks.’, *Proceedings of the National Academy of Sciences*, 89(22), pp. 10915–10919. doi:10.1073/pnas.89.22.10915.

Henzler-Wildman, K. and Kern, D. (2007) ‘Dynamic personalities of proteins’, *Nature*, 450(7172), pp. 964–972. doi:10.1038/nature06522.

Heo, L., Shin, W.-H., Lee, M.S. and Seok, C. (2014) ‘GalaxySite: ligand-binding-site prediction by using molecular docking’, *Nucleic Acids Research*, 42(W1), pp. W210–W214. doi:10.1093/nar/gku321.

Hergueta-Redondo, M., Sarrió, D., Molina-Crespo, Á., Megias, D., Mota, A., Rojo-Sebastian, A., García-Sanz, P., Morales, S., Abril, S., Cano, A., Peinado, H. and Moreno-Bueno, G. (2014) ‘Gasdermin-B Promotes Invasion and Metastasis in Breast Cancer Cells’, *PLoS ONE*. Edited by H. Wanjin, 9(3), p. e90099. doi:10.1371/journal.pone.0090099.

Hillen, H.S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D. and Cramer, P. (2020) ‘Structure of replicating SARS-CoV-2 polymerase’, *Nature*, 584(7819), pp. 154–156. doi:10.1038/s41586-020-2368-8.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M.A., Drosten, C. and Pöhlmann, S. (2020) ‘SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.’, *Cell*, 181(2), pp. 271-280.e8. doi:10.1016/j.cell.2020.02.052.

Hu, P., Bader, G., Wigle, D.A. and Emili, A. (2007) ‘Computational prediction of cancer-gene function’, *Nature Reviews Cancer*, 7(1), pp. 23–34. doi:10.1038/nrc2036.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., *et al.* (2020) ‘Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China’, *The Lancet*, 395(10223), pp. 497–506. doi:10.1016/S0140-6736(20)30183-5.

Huang, L., Liao, T.-W., Wang, J., Ha, T. and Lilley, D.M.J. (2020) ‘Crystal structure and ligand-induced folding of the SAM/SAH riboswitch.’, *Nucleic acids research*, 48(13), pp. 7545–7556. doi:10.1093/nar/gkaa493.

Hur, Y.-S., Shin, K.-H., Kim, S., Nam, K.H., Lee, M.-S., Chun, J.-Y. and Cheon, C.-I. (2009) ‘Overexpression of GmAKR1, a stress-induced aldo/keto reductase from soybean, retards nodule development.', *Molecules and cells*, 27(2), pp. 217–23. doi:10.1007/s10059-009-0027-x.

Ivanov, A.A., Khuri, F.R. and Fu, H. (2013) 'Targeting protein-protein interactions as an anticancer strategy.', *Trends in pharmacological sciences*, 34(7), pp. 393–400. doi:10.1016/j.tips.2013.04.007.

Jeffery, C.J. (2005) 'Mass spectrometry and the search for moonlighting proteins.', *Mass spectrometry reviews*, 24(6), pp. 772–82. doi:10.1002/mas.20041.

Jian, J.-W., Elumalai, P., Pitti, T., Wu, C.Y., Tsai, K.-C., Chang, J.-Y., Peng, H.-P. and Yang, A.-S. (2016) 'Predicting Ligand Binding Sites on Protein Surfaces by 3-Dimensional Probability Density Distributions of Interacting Atoms', *PLOS ONE*. Edited by Y. Zhang, 11(8), p. e0160315. doi:10.1371/journal.pone.0160315.

Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D'Andrea, D., Lepore, R., Funk, C.S., Kahanda, I., Verspoor, K.M., Ben-Hur, A., Koo, D.C.E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, S.M.E., Martelli, P.L., Profiti, G., *et al.* (2016) 'An expanded evaluation of protein function prediction methods shows an improvement in accuracy', *Genome Biology*, 17(1), p. 184. doi:10.1186/s13059-016-1037-6.

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A.S. and De Fabritiis, G. (2017) 'DeepSite: protein-binding site predictor using 3D-convolutional neural networks', *Bioinformatics*. Edited by A. Valencia, 33(19), pp. 3036–3042. doi:10.1093/bioinformatics/btx350.

Johnson, P.M., Gucinski, G.C., Garza-Sánchez, F., Wong, T., Hung, L.-W., Hayes, C.S. and Goulding, C.W. (2016) 'Functional Diversity of Cytotoxic tRNase/Immunity Protein Complexes from Burkholderia pseudomallei', *Journal of Biological Chemistry*, 291(37), pp. 19387–19400. doi:10.1074/jbc.M116.736074.

Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (1997) 'Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen', *Journal of Molecular Biology*, 267(3), pp. 727–748. doi:10.1006/jmbi.1996.0897.

Jones, S. and Thornton, J.M. (1996) 'Principles of protein-protein interactions.', *Proceedings of the National Academy of Sciences of the United States of America*, 93(1), pp. 13–20. doi:10.1073/pnas.93.1.13.

Jorgensen, I. and Miao, E.A. (2015) 'Pyroptotic cell death defends against intracellular pathogens.', *Immunological reviews*, 265(1), pp. 130–42. doi:10.1111/imr.12287.

Kahanda, I., Funk, C.S., Ullah, F., Verspoor, K.M. and Ben-Hur, A. (2015) 'A close look at protein function prediction evaluation protocols.', *GigaScience*, 4, p. 41. doi:10.1186/s13742-015-0082-5.

Kang, S., Brown, H.M. and Hwang, S. (2018) 'Direct Antiviral Mechanisms of Interferon-Gamma', *Immune Network*, 18(5). doi:10.4110/in.2018.18.e33.

Kaore, Shilpa N. Kaore, N.M. (2014) *Biomarkers in Toxicology*. 1st Editio. Edited by R.C. Gupta. doi:10.1016/C2012-0-01373-7.

Kapetanovic, I.M. (2008) 'Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach.', *Chemico-biological interactions*, 171(2), pp. 165–76. doi:10.1016/j.cbi.2006.12.006.

Kashiwagi, K., Ito, T. and Yokoyama, S. (2014) 'Crystal structure of the eukaryotic translation initiation factor 2A from Schizosaccharomyces pombe.', *Journal of structural and functional genomics*, 15(3), pp. 125–30. doi:10.1007/s10969-014-9177-y.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*, 10(6), pp. 845–858. doi:10.1038/nprot.2015.053.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) 'A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis', *Nature*, 181(4610), pp. 662–666. doi:10.1038/181662a0.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N. and Chang, H. (2020) 'The Architecture of SARS-CoV-2 Transcriptome', *Cell*, 181(4), pp. 914-921.e10.

doi:10.1016/j.cell.2020.04.011.
Kim, D.E., Chivian, D. and Baker, D. (2004) ‘Protein structure prediction and analysis using the Robetta server.’, *Nucleic acids research*, 32(Web Server issue), pp. W526-31. doi:10.1093/nar/gkh468.
Kim, Y., Jedrzejczak, R., Maltseva, N.I., Wilamowski, M., Endres, M., Godzik, A., Michalska, K. and Joachimiak, A. (2020) ‘Crystal structure of Nsp15 endoribonuclease <scp>NendoU</scp> from <scp>SARS-CoV</scp> -2’, *Protein Science*, 29(7), pp. 1596–1605. doi:10.1002/pro.3873.
Kitahara, M. (2005) ‘Bacteroides plebeius sp. nov. and Bacteroides coprocola sp. nov., isolated from human faeces’, *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 55(5), pp. 2143–2147. doi:10.1099/ijs.0.63788-0.
Kiyotani, K., Toyoshima, Y., Nemoto, K. and Nakamura, Y. (2020) ‘Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2’, *Journal of Human Genetics*, 65(7), pp. 569–575. doi:10.1038/s10038-020-0771-5.
Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.-A., Bick, M.J., Bauer, A., Liu, G., Ishida, Y., Boykov, A., Estep, R.D., Kleinfelter, S., Nørgård-Solano, T., Wei, L., Players, F., Montelione, G.T., DiMaio, F., Popović, Z., Khatib, F., *et al.* (2019) ‘De novo protein design by citizen scientists.’, *Nature*, 570(7761), pp. 390–394. doi:10.1038/s41586-019-1274-4.
Konc, J. and Janežič, D. (2014) ‘ProBiS-ligands: a web server for prediction of ligands by examination of protein binding sites.’, *Nucleic acids research*, 42(Web Server issue), pp. W215-20. doi:10.1093/nar/gku460.
Konkolova, E., Klima, M., Nencka, R. and Boura, E. (2020) ‘Structural analysis of the putative SARS-CoV-2 primase complex.’, *Journal of structural biology*, 211(2), p. 107548. doi:10.1016/j.jsb.2020.107548.
Koonin, E. V and Tatusov, R.L. (1994) ‘Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search.’, *Journal of molecular biology*, 244(1), pp. 125–32. doi:10.1006/jmbi.1994.1711.
Kopecky-Bromberg, S.A., Martínez-Sobrido, L., Frieman, M., Baric, R.A. and Palese, P. (2007) ‘Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists.’, *Journal of virology*, 81(2), pp. 548–57. doi:10.1128/JVI.01782-06.
Korb, O., Stützle, T. and Exner, T.E. (2006) ‘PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design’, in, pp. 247–258. doi:10.1007/11839088_22.
Koskinen, P., Törönen, P., Nokso-Koivisto, J. and Holm, L. (2015) ‘PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment.’, *Bioinformatics (Oxford, England)*, 31(10), pp. 1544–52. doi:10.1093/bioinformatics/btu851.
Koval, A. V, Vlasov, P., Shichkova, P., Khunderyakova, S., Markov, Y., Panchenko, J., Volodina, A., Kondrashov, F.A. and Katanaev, V.L. (2014) ‘Anti-leprosy drug clofazimine inhibits growth of triple-negative breast cancer cells via inhibition of canonical Wnt signaling.’, *Biochemical pharmacology*, 87(4), pp. 571–8. doi:10.1016/j.bcp.2013.12.007.
Krivák, R. and Hoksza, D. (2018) ‘P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure.’, *Journal of cheminformatics*, 10(1), p. 39. doi:10.1186/s13321-018-0285-8.
Kumar, P., Gunalan, V., Liu, B., Chow, V.T.K.K., Druce, J., Birch, C., Catton, M., Fielding, B.C., Tan, Y.-J. and Lal, S.K. (2007) ‘The nonstructural protein 8 (nsp8) of the SARS coronavirus interacts with its ORF6 accessory protein.’, *Virology*, 366(2), pp. 293–303. doi:10.1016/j.virol.2007.04.029.
Kwong, K. and Carr, M.J. (2015) ‘Voltage-gated sodium channels.’, *Current opinion in pharmacology*, 22, pp. 131–9. doi:10.1016/j.coph.2015.04.007.
La, D., Kong, M., Hoffman, W., Choi, Y.I. and Kihara, D. (2013) ‘Predicting permanent and transient protein-protein interfaces.’, *Proteins*, 81(5), pp. 805–18. doi:10.1002/prot.24235.
Lahoz, E.G., de Haro, M.A.L. and Esponda, P. (1991) ‘Use of puromycin N-acetyltransferase

(PAC) as a new reporter gene in transgenic animals', *Nucleic Acids Research*, 19(12), pp. 3465–3465. doi:10.1093/nar/19.12.3465.

Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. and Hsueh, P.-R. (2020) 'Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges', *International Journal of Antimicrobial Agents*, 55(3), p. 105924. doi:10.1016/j.ijantimicag.2020.105924.

Lan, J., Yang, Q., Zhou, M., Xu, R., Zhou, C., Wang, J. and Zheng, H. (2016) 'A meta-analysis of association between glutathione S-transferase gene polymorphism and osteosarcoma chemosensitivity in Chinese population.', *Journal of cancer research and therapeutics*, 12(Supplement), pp. 64–67. doi:10.4103/0973-1482.191634.

Lance, B.K., Deane, C.M. and Wood, G.R. (2010) 'Exploring the potential of template-based modelling', *Bioinformatics*, 26(15), pp. 1849–1856. doi:10.1093/bioinformatics/btq294.

Laver, D.R., Lenz, G.K.E. and Dulhunty, A.F. (2001) 'Phosphate ion channels in sarcoplasmic reticulum of rabbit skeletal muscle', *The Journal of Physiology*, 535(3), pp. 715–728. doi:10.1111/j.1469-7793.2001.t01-1-00715.x.

Leano, J.B., Batarni, S., Eriksen, J., Juge, N., Pak, J.E., Kimura-Someya, T., Robles-Colmenares, Y., Moriyama, Y., Stroud, R.M. and Edwards, R.H. (2019) 'Structures suggest a mechanism for energy coupling by a family of organic anion transporters.', *PLoS biology*, 17(5), p. e3000260. doi:10.1371/journal.pbio.3000260.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K.W., Renfrew, P.D., Smith, C.A., Sheffler, W., Davis, I.W., Cooper, S., Treuille, A., Mandell, D.J., Richter, F., Ban, Y.-E.A., Fleishman, S.J., Corn, J.E., Kim, D.E., *et al.* (2011) 'Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules', in Johnson, M.L. and Brand, L. (eds) *Computer Methods, Part C*. Academic Press (Methods in Enzymology), pp. 545–574. doi:https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

Lee, H.S. and Im, W. (2013) 'Ligand Binding Site Detection by Local Structure Alignment and Its Performance Complementarity', *Journal of Chemical Information and Modeling*, 53(9), pp. 2462–2470. doi:10.1021/ci4003602.

Lei, X., Dong, X., Ma, R., Wang, W., Xiao, X., Tian, Z., Wang, C., Wang, Y., Li, L., Ren, L., Guo, F., Zhao, Z., Zhou, Z., Xiang, Z. and Wang, J. (2020) 'Activation and evasion of type I interferon responses by SARS-CoV-2.', *Nature communications*, 11(1), p. 3810. doi:10.1038/s41467-020-17665-9.

Lepore, R., Kryshtafovych, A., Alahuhta, M., Veraszto, H.A., Bomble, Y.J., Bufton, J.C., Bullock, A.N., Caba, C., Cao, H., Davies, O.R., Desfosses, A., Dunne, M., Fidelis, K., Goulding, C.W., Gurusaran, M., Gutsche, I., Harding, C.J., Hartmann, M.D., Hayes, C.S., *et al.* (2019) 'Target highlights in CASP13: Experimental target structures through the eyes of their authors', *Proteins: Structure, Function, and Bioinformatics*, 87(12), pp. 1037–1057. doi:10.1002/prot.25805.

Li, C., Iosef, C., Jia, C.Y.H., Gkourasas, T., Han, V.K.M. and Shun-Cheng Li, S. (2003) 'Disease-causing SAP mutants are defective in ligand binding and protein folding.', *Biochemistry*, 42(50), pp. 14885–92. doi:10.1021/bi034798l.

Li, J.-Y., Liao, C.-H., Wang, Q., Tan, Y.-J., Luo, R., Qiu, Y. and Ge, X.-Y. (2020) 'The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway.', *Virus research*, 286, p. 198074. doi:10.1016/j.virusres.2020.198074.

Li, W.-J., Li, D.-F., Hu, Y.-L., Zhang, X.-E., Bi, L.-J. and Wang, D.-C. (2013) 'Crystal structure of L,D-transpeptidase LdtMt2 in complex with meropenem reveals the mechanism of carbapenem against Mycobacterium tuberculosis.', *Cell research*, 23(5), pp. 728–31. doi:10.1038/cr.2013.53.

Liekniņa, I., Kalniņš, G., Akopjana, I., Bogans, J., Šišovs, M., Jansons, J., Rūmnieks, J. and Tārs, K. (2019) 'Production and characterization of novel ssRNA bacteriophage virus-like particles from metagenomic sequencing data.', *Journal of nanobiotechnology*, 17(1), p. 61. doi:10.1186/s12951-019-0497-8.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) ‘Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings’, *Advanced Drug Delivery Reviews*, 23(1), pp. 3–25. doi:https://doi.org/10.1016/S0169-409X(96)00423-1.

Littler, D.R., Gully, B.S., Colson, R.N. and Rossjohn, J. (2020) ‘Crystal Structure of the SARS-CoV-2 Non-structural Protein 9, Nsp9’, *iScience*, 23(7), p. 101258. doi:10.1016/j.isci.2020.101258.

Liu, T., Ish-Shalom, S., Torng, W., Lafita, A., Bock, C., Mort, M., Cooper, D.N., Bliven, S., Capitani, G., Mooney, S.D. and Altman, R.B. (2018) ‘Biological and functional relevance of CASP predictions.’, *Proteins*, 86 Suppl 1, pp. 374–386. doi:10.1002/prot.25396.

Liu, T., Lin, Y., Wen, X., Jorissen, R.N. and Gilson, M.K. (2007) ‘BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.’, *Nucleic acids research*, 35(Database issue), pp. D198-201. doi:10.1093/nar/gkl999.

Liu, W.L.H. (2020a) ‘COVID-19:Attacks the 1-Beta Chain of Hemoglobin and Captures the Porphyrin to Inhibit Human Heme Metabolism’, *ChemRxiv* [Preprint].

Liu, W.L.H. (2020b) ‘COVID-19 Disease: ORF8 and surface glycoprotein inhibit heme metabolism by binding to porphyrin’, *ChemRxiv* [Preprint]. doi:https://doi.org/10.26434/chemrxiv.11938173.v1.

Lodish, Harvey; Berk, Arnold; Zipursky, S Lawerence; Matsudaira, Paul; Baltimore, David; Darnell, J. (2000) ‘Section 3.3 Functional Design of Proteins’, in *Molecular Cell Biology 4th Edition*. 4th edn. New York, pp. 854–918. Available at: https://www.ncbi.nlm.nih.gov/books/NBK21475/.

Londoño, O.M., Tancredi, P., Rivas, P., Muraca, D., Socolovsky, L.M. and Knobel, M. (2018) ‘Small-Angle X-Ray Scattering to Analyze the Morphological Properties of Nanoparticulated Systems’, in *Handbook of Materials Characterization*. Cham: Springer International Publishing, pp. 37–75. doi:10.1007/978-3-319-92955-2_2.

López, G., Ezkurdia, I. and Tress, M.L. (2009) ‘Assessment of ligand binding residue predictions in CASP8.’, *Proteins*, 77 Suppl 9, pp. 138–46. doi:10.1002/prot.22557.

López, G., Rojas, A., Tress, M. and Valencia, A. (2007) ‘Assessment of predictions submitted for the CASP7 function prediction category.’, *Proteins*, 69 Suppl 8, pp. 165–74. doi:10.1002/prot.21651.

Lopez, G., Valencia, A. and Tress, M. (2007) ‘FireDB--a database of functionally important residues from proteins of known structure.’, *Nucleic acids research*, 35(Database issue), pp. D219-23. doi:10.1093/nar/gkl897.

Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R. and Shi, J. (2020) ‘Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials’, *Signal Transduction and Targeted Therapy*, 5(1), p. 213. doi:10.1038/s41392-020-00315-3.

Magnan, C.N. and Baldi, P. (2014) ‘SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity.’, *Bioinformatics (Oxford, England)*, 30(18), pp. 2592–7. doi:10.1093/bioinformatics/btu352.

Magrane, M. and UniProt Consortium (2011) ‘UniProt Knowledgebase: a hub of integrated protein data.’, *Database : the journal of biological databases and curation*, 2011, p. bar009. doi:10.1093/database/bar009.

Majumdar, P. and Niyogi, S. (2020) ‘ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection’, *Epidemiology and Infection*, 148, p. e262. doi:10.1017/S0950268820002599.

Malinverni, J.C. and Silhavy, T.J. (2009) ‘An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane.’, *Proceedings of the National Academy of Sciences of the United States of America*, 106(19), pp. 8009–14. doi:10.1073/pnas.0903229106.

Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. and Evans, R.M. (1995) ‘The nuclear receptor superfamily: the second decade.', *Cell*, 83(6), pp. 835–9. doi:10.1016/0092-8674(95)90199-x.

Marienhagen, J., Sandalova, T., Sahm, H., Eggeling, L. and Schneider, G. (2008) 'Insights into the structural basis of substrate recognition by histidinol-phosphate aminotransferase from Corynebacterium glutamicum.', *Acta crystallographica. Section D, Biological crystallography*, 64(Pt 6), pp. 675–85. doi:10.1107/S0907444908009438.

Mateo, M., Generous, A., Sinn, P.L. and Cattaneo, R. (2015) 'Connections matter--how viruses use cell–cell adhesion components.', *Journal of cell science*, 128(3), pp. 431–9. doi:10.1242/jcs.159400.

Matthews, B.W. (1975) 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme.', *Biochimica et biophysica acta*, 405(2), pp. 442–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1180967.

Mayr, P. and Nidetzky, B. (2002) 'Catalytic reaction profile for NADH-dependent reduction of aromatic aldehydes by xylose reductase from Candida tenuis.', *The Biochemical journal*, 366(Pt 3), pp. 889–99. doi:10.1042/BJ20020080.

McClain, C.B. and Vabret, N. (2020) 'SARS-CoV-2: the many pros of targeting PLpro', *Signal Transduction and Targeted Therapy*, 5(1), p. 223. doi:10.1038/s41392-020-00335-z.

McGuffin, LJ; Adiyaman, R. (2021) 'ReFOLD3: refinement of 3D protein models with gradual restraints based on predicted local quality and residue contacts', *Nucleic acids research* [Preprint].

McGuffin, LJ; Aldowsari, F; Alharbi, S.A.R. (2021) 'ModFOLD8: accurate global and local quality estimates for 3D protein models', *Nucleic acids research* [Preprint].

McGuffin, L.J., Atkins, J.D., Salehe, B.R., Shuid, A.N. and Roche, D.B. (2015) 'IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences.', *Nucleic acids research*, 43(W1), pp. W169-73. doi:10.1093/nar/gkv236.

McGuffin, L.J. and Roche, D.B. (2010) 'Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments.', *Bioinformatics (Oxford, England)*, 26(2), pp. 182–8. doi:10.1093/bioinformatics/btp629.

McGuffin, L.J. and Roche, D.B. (2011) 'Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method.', *Proteins*, 79 Suppl 1, pp. 137–46. doi:10.1002/prot.23120.

McGuffin, L.J.L.J., Adiyaman, R., Maghrabi, A.H.A.A.H.A., Shuid, A.N.A.N., Brackenridge, D.A.D.A., Nealon, J.O.J.O. and Philomina, L.S.L.S. (2019) 'IntFOLD: an integrated web resource for high performance protein structure and function prediction', *Nucleic Acids Research*, 47(W1), pp. W408–W413. doi:10.1093/nar/gkz322.

Menéndez-Conejero, R., Nguyen, T.H., Singh, A.K., Condezo, G.N., Marschang, R.E., van Raaij, M.J. and San Martín, C. (2017) 'Structure of a Reptilian Adenovirus Reveals a Phage Tailspike Fold Stabilizing a Vertebrate Virus Capsid.', *Structure (London, England : 1993)*, 25(10), pp. 1562-1573.e5. doi:10.1016/j.str.2017.08.007.

Meyer, B. and Peters, T. (2003) 'NMR Spectroscopy Techniques for Screening and Identifying Ligand Binding to Protein Receptors', *Angewandte Chemie International Edition*, 42(8), pp. 864–890. doi:10.1002/anie.200390233.

Mezei, M. (2003) 'A new method for mapping macromolecular topography.', *Journal of molecular graphics & modelling*, 21(5), pp. 463–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12543141.

Miao, G., Zhao, H., Li, Y., Ji, M., Chen, Y., Shi, Y., Bi, Y., Wang, P. and Zhang, H. (2021) 'ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation.', *Developmental cell*, 56(4), pp. 427-442.e5. doi:10.1016/j.devcel.2020.12.010.

Michel, C.J., Mayer, C., Poch, O. and Thompson, J.D. (2020) 'Characterization of accessory genes in coronavirus genomes', *Virology Journal*, 17(1), p. 131. doi:10.1186/s12985-020-01402-1.

Miorin, L., Kehrer, T., Sanchez-Aparicio, M.T., Zhang, K., Cohen, P., Patel, R.S., Cupic, A., Makio, T., Mei, M., Moreno, E., Danziger, O., White, K.M., Rathnasinghe, R., Uccellini, M., Gao, S., Aydillo, T., Mena, I., Yin, X., Martin-Sancho, L., *et al.* (2020) 'SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling', *Proceedings of the National Academy of Sciences*, 117(45), pp. 28344 LP – 28354. doi:10.1073/pnas.2016650117.

Mir-Sanchis, I., Pigli, Y.Z. and Rice, P.A. (2018) 'Crystal Structure of an Unusual Single-Stranded DNA-Binding Protein Encoded by Staphylococcal Cassette Chromosome Elements.', *Structure (London, England : 1993)*, 26(8), pp. 1144-1150.e3. doi:10.1016/j.str.2018.05.016.

Mittermaier, A. and Kay, L.E. (2006) 'New Tools Provide New Insights in NMR Studies of Protein Dynamics', *Science*, 312(5771), pp. 224 LP – 228. doi:10.1126/science.1124964.

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A. and Kryshtafovych, A. (2016) 'New encouraging developments in contact prediction: Assessment of the CASP11 results.', *Proteins*, 84 Suppl 1, pp. 131–44. doi:10.1002/prot.24943.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. and Olson, A.J. (2009) 'AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.', *Journal of computational chemistry*, 30(16), pp. 2785–91. doi:10.1002/jcc.21256.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. and Tramontano, A. (2007) 'Critical assessment of methods of protein structure prediction-Round VII.', *Proteins*, 69 Suppl 8, pp. 3–9. doi:10.1002/prot.21767.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Tramontano, A. (2016) 'Critical assessment of methods of protein structure prediction: Progress and new directions in round XI.', *Proteins*, 84 Suppl 1, pp. 4–14. doi:10.1002/prot.25064.

Moult, J., Fidelis, K., Rost, B., Hubbard, T. and Tramontano, A. (2005) 'Critical assessment of methods of protein structure prediction (CASP)--round 6.', *Proteins*, 61 Suppl 7, pp. 3–7. doi:10.1002/prot.20716.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001) 'Critical assessment of methods of protein structure prediction (CASP): round IV.', *Proteins*, Suppl 5, pp. 2–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11835476.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) 'Critical assessment of methods of protein structure prediction (CASP)-round V.', *Proteins*, 53 Suppl 6, pp. 334–9. doi:10.1002/prot.10556.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. (1997) 'Critical assessment of methods of protein structure prediction (CASP): round II.', *Proteins*, Suppl 1, pp. 2–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9485489.

Moult, J., Hubbard, T. and Fidelis, K. (1999) 'Critical assessment of methods of protein structure prediction (CASP): Round III', *Proteins: Structure, Function, and Bioinformatics*, 37(S3), pp. 2–6.

Moult, J., Pedersen, J.T., Judson, R. and Fidelis, K. (1995) 'A large-scale experiment to assess protein structure prediction methods.', *Proteins*, 23(3), pp. ii–v. doi:10.1002/prot.340230303.

Mousavizadeh, L. and Ghasemi, S. (2020) 'Genotype and phenotype of COVID-19: Their roles in pathogenesis', *Journal of Microbiology, Immunology and Infection* [Preprint]. doi:10.1016/j.jmii.2020.03.022.

Moutevelis, E. and Woolfson, D.N. (2009) 'A Periodic Table of Coiled-Coil Protein Structures', *Journal of Molecular Biology*, 385(3), pp. 726–732. doi:10.1016/j.jmb.2008.11.028.

Na, I., Catena, D., Kong, M.J., Ferreira, G.C. and Uversky, V.N. (2018) 'Anti-Correlation between the Dynamics of the Active Site Loop and C-Terminal Tail in Relation to the Homodimer Asymmetry of the Mouse Erythroid 5-Aminolevulinate Synthase.', *International journal of molecular sciences*, 19(7). doi:10.3390/ijms19071899.

Neal-McKinney, J.M. and Konkel, M.E. (2012) 'The Campylobacter jejuni CiaC virulence protein is secreted from the flagellum and delivered to the cytosol of host cells.', *Frontiers in*

cellular and infection microbiology, 2, p. 31. doi:10.3389/fcimb.2012.00031.
Needleman, S.B. and Wunsch, C.D. (1970) ‘A general method applicable to the search for similarities in the amino acid sequence of two proteins.’, *Journal of molecular biology*, 48(3), pp. 443–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5420325.
Nelson, D.R., Koymans, L., Kamataki, T., Stegeman, J.J., Feyereisen, R., Waxman, D.J., Waterman, M.R., Gotoh, O., Coon, M.J., Estabrook, R.W., Gunsalus, I.C. and Nebert, D.W. (1996) ‘P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature.’, *Pharmacogenetics*, 6(1), pp. 1–42. doi:10.1097/00008571-199602000-00002.
Nemoto, W. and Toh, H. (2012) ‘Functional region prediction with a set of appropriate homologous sequences--an index for sequence selection by integrating structure and sequence information with spatial statistics.’, *BMC structural biology*, 12, p. 11. doi:10.1186/1472-6807-12-11.
Ng, L.F., Kaur, P., Bunnag, N., Suresh, J., Sung, I.C.H., Tan, Q.H., Gruber, J. and Tolwinski, N.S. (2019) ‘WNT Signaling in Disease.’, *Cells*, 8(8). doi:10.3390/cells8080826.
Ngan, C.-H., Hall, D.R., Zerbe, B., Grove, L.E., Kozakov, D. and Vajda, S. (2012) ‘FTSite: high accuracy detection of ligand binding sites on unbound protein structures’, *Bioinformatics*, 28(2), pp. 286–287. doi:10.1093/bioinformatics/btr651.
Niu, W., Shu, Q., Chen, Z., Mathews, S., Di Cera, E. and Frieden, C. (2010) ‘The role of Zn2+ on the structure and stability of murine adenosine deaminase.’, *The journal of physical chemistry. B*, 114(49), pp. 16156–65. doi:10.1021/jp106041v.
Nooren, I.M.A. and Thornton, J.M. (2003) ‘Diversity of protein-protein interactions.’, *The EMBO journal*, 22(14), pp. 3486–92. doi:10.1093/emboj/cdg359.
Nopoulos, P.C. (2016) ‘Huntington disease: a single-gene degenerative disorder of the striatum.’, *Dialogues in clinical neuroscience*, 18(1), pp. 91–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27069383.
O’Connor, C.M. (2021) ‘BLOSUM62 scoring matrix for amino acid substitutions’. Boston College. Available at: https://bio.libretexts.org/@go/page/17548.
Ogrizek, M., Konc, J., Bren, U., Hodoščck, M. and Janežič, D. (2016) ‘Role of magnesium ions in the reaction mechanism at the interface between Tm1631 protein and its DNA ligand.’, *Chemistry Central journal*, 10, p. 41. doi:10.1186/s13065-016-0188-6.
Oh, M., Joo, K. and Lee, J. (2009) ‘Protein-binding site prediction based on three-dimensional protein modeling.’, *Proteins*, 77 Suppl 9, pp. 152–6. doi:10.1002/prot.22572.
Okuyama, M., Yoshida, T., Hondoh, H., Mori, H., Yao, M. and Kimura, A. (2014) ‘Catalytic role of the calcium ion in GH97 inverting glycoside hydrolase.’, *FEBS letters*, 588(17), pp. 3213–7. doi:10.1016/j.febslet.2014.07.002.
Oostra, M., te Lintelo, E.G., Deijs, M., Verheije, M.H., Rottier, P.J.M. and de Haan, C.A.M. (2007) ‘Localization and membrane topology of coronavirus nonstructural protein 4: involvement of the early secretory pathway in replication.’, *Journal of virology*, 81(22), pp. 12323–36. doi:10.1128/JVI.01506-07.
Ouellette, R.J. and Rawn, J.D. (2015) ‘14 - Amino Acids, Peptides, and Proteins’, in Ouellette, R.J. and Rawn, J.D. (eds) *Principles of Organic Chemistry*. Boston: Elsevier, pp. 371–396. doi:https://doi.org/10.1016/B978-0-12-802444-7.00014-8.
Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F. and Baker, D. (2018) ‘Protein structure prediction using Rosetta in CASP12.’, *Proteins*, 86 Suppl 1, pp. 113–121. doi:10.1002/prot.25390.
Pagadala, N.S., Syed, K. and Tuszynski, J. (2017) ‘Software for molecular docking: a review.’, *Biophysical reviews*, 9(2), pp. 91–102. doi:10.1007/s12551-016-0247-1.
Palm-Espling, M.E., Niemiec, M.S. and Wittung-Stafshede, P. (2012) ‘Role of metal in folding and stability of copper proteins in vitro.’, *Biochimica et biophysica acta*, 1823(9), pp. 1594–603. doi:10.1016/j.bbamcr.2012.01.013.
Papadopoulos, V., Fan, J. and Zirkin, B. (2017) ‘Translocator protein (18 kDa): an update on its function in steroidogenesis.’, *Journal of neuroendocrinology* [Preprint].

doi:10.1111/jne.12500.

Park, J., Matralis, A.N., Berghuis, A.M. and Tsantrizos, Y.S. (2014) ‘Human isoprenoid synthase enzymes as therapeutic targets.’, *Frontiers in chemistry*, 2, p. 50. doi:10.3389/fchem.2014.00050.

Pasternak, G.W. and Pan, Y.-X. (2013) ‘Mu opioids and their receptors: evolution of a concept.’, *Pharmacological reviews*, 65(4), pp. 1257–317. doi:10.1124/pr.112.007138.

Patel, M. and Shah, H. (2013) ‘Protein Secondary Structure Prediction Using Support Vector Machines (SVMs)’, in *2013 International Conference on Machine Intelligence and Research Advancement*, pp. 594–598. doi:10.1109/ICMIRA.2013.124.

Pauling, L., Corey, R.B. and Branson, H.R. (1951) ‘The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.’, *Proceedings of the National Academy of Sciences of the United States of America*, 37(4), pp. 205–11. doi:10.1073/pnas.37.4.205.

Pelton, J.T. and McLean, L.R. (2000) ‘Spectroscopic Methods for Analysis of Protein Secondary Structure’, *Analytical Biochemistry*, 277(2), pp. 167–176. doi:10.1006/abio.1999.4320.

Peng, Q., Peng, R., Yuan, B., Zhao, J., Wang, M., Wang, X., Wang, Q., Sun, Y., Fan, Z., Qi, J., Gao, G.F. and Shi, Y. (2020) ‘Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2.’, *Cell reports*, 31(11), p. 107774. doi:10.1016/j.celrep.2020.107774.

Pertsemlidis, A. and Fondon, J.W. (2001) ‘Having a BLAST with bioinformatics (and avoiding BLASTphemy).’, *Genome biology*, 2(10), p. REVIEWS2002. doi:10.1186/gb-2001-2-10-reviews2002.

Pham, T., Perry, J.L., Dosey, T.L., Delcour, A.H. and Hyser, J.M. (2017) ‘The Rotavirus NSP4 Viroporin Domain is a Calcium-conducting Ion Channel.’, *Scientific reports*, 7, p. 43487. doi:10.1038/srep43487.

Pinoli, P., Chicco, D. and Masseroli, M. (2015) ‘Computational algorithms to predict Gene Ontology annotations.’, *BMC bioinformatics*, 16 Suppl 6, p. S4. doi:10.1186/1471-2105-16-S6-S4.

Piovesan, D., Giollo, M., Ferrari, C. and Tosatto, S.C.E. (2015) ‘Protein function prediction using guilty by association from interaction networks.’, *Amino acids*, 47(12), pp. 2583–92. doi:10.1007/s00726-015-2049-3.

Piovesan, D., Giollo, M., Leonardi, E., Ferrari, C. and Tosatto, S.C.E.E. (2015) ‘INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity’, *Nucleic acids research*, 43(W1), pp. W134–W140. doi:10.1093/nar/gkv523.

Pirovano, W. and Heringa, J. (2010) ‘Protein Secondary Structure Prediction’, in Carugo, O. and Eisenhaber, F. (eds) *Data Mining Techniques for the Life Sciences*. Totowa, NJ: Humana Press, pp. 327–348. doi:10.1007/978-1-60327-241-4_19.

Polakis, P. (2000) ‘Wnt signaling and cancer.’, *Genes & development*, 14(15), pp. 1837–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10921899.

Pollastri, G. and McLysaght, A. (2005) ‘Porter: a new, accurate server for protein secondary structure prediction’, *Bioinformatics*, 21(8), pp. 1719–1720. doi:10.1093/bioinformatics/bti203.

Powers, D. (2020) ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’, *ArXiv*, abs/2010.1.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A. and Finn, R.D. (2012) ‘The Pfam protein families database.’, *Nucleic acids research*, 40(Database issue), pp. D290-301. doi:10.1093/nar/gkr1065.

Punta, M. and Ofran, Y. (2008) ‘The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function.’, *PLoS computational biology*, 4(10), p. e1000160. doi:10.1371/journal.pcbi.1000160.

QuickGO (no date). Available at: http://www.ebi.ac.uk/QuickGO (Accessed: 6 May 2017).

Quinten, T.A. and Kuhn, A. (2012) 'Membrane Interaction of the Portal Protein gp20 of Bacteriophage T4', *Journal of Virology*, 86(20), pp. 11107–11114. doi:10.1128/JVI.01284-12.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo, S., Souza, M.L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., *et al.* (2013) 'A large-scale evaluation of computational protein function prediction.', *Nature methods*, 10(3), pp. 221–7. doi:10.1038/nmeth.2340.

Rang, H.P, Ritter, J.M, Flower, R.J, Henderson, G. (2015) *Rang and Dale's pharmacology*. Eighth. London: Elsevier Churchill Livingstone.

Rangwala, H. and Karypis, G. (2010) 'Introduction to Protein Structure Prediction', in *Introduction to Protein Structure Prediction*. John Wiley & Sons, Ltd, pp. 1–13. doi:https://doi.org/10.1002/9780470882207.ch1.

Rarey, M, Kramer, B., Lengauer, T. and Klebe, G. (1996) 'A fast flexible docking method using an incremental construction algorithm.', *Journal of molecular biology*, 261(3), pp. 470–89. doi:10.1006/jmbi.1996.0477.

Rarey, Matthias, Kramer, B., Lengauer, T. and Klebe, G. (1996) 'A Fast Flexible Docking Method using an Incremental Construction Algorithm', *Journal of molecular biology*, 261(3), pp. 470–89. doi:10.1006/jmbi.1996.0477.

Rasmussen, K.K., Palencia, A., Varming, A.K., El-Wali, H., Boeri Erba, E., Blackledge, M., Hammer, K., Herrmann, T., Kilstrup, M., Lo Leggio, L. and Jensen, M.R. (2020) 'Revealing the mechanism of repressor inactivation during switching of a temperate bacteriophage', *Proceedings of the National Academy of Sciences*, 117(34), pp. 20576–20585. doi:10.1073/pnas.2005218117.

Ratnasari, D., Nazir, F., Toresano, L.O.H.Z., Pawiro, S.A. and Soejoko, D.S. (2016) 'The correlation between effective renal plasma flow (ERPF) and glomerular filtration rate (GFR) with renal scintigraphy 99m Tc-DTPA study', *Journal of Physics: Conference Series*, 694, p. 012062. doi:10.1088/1742-6596/694/1/012062.

Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2012) 'HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment', *Nature methods*, 9(2), pp. 173–5. doi:10.1038/nmeth.1818.

Ren, Y., Shu, T., Wu, D., Mu, J., Wang, C., Huang, M., Han, Y., Zhang, X.-Y., Zhou, W., Qiu, Y. and Zhou, X. (2020) 'The ORF3a protein of SARS-CoV-2 induces apoptosis in cells', *Cellular & Molecular Immunology*, 17(8), pp. 881–883. doi:10.1038/s41423-020-0485-9.

Ren, Z., Bourgeois, D., Helliwell, J.R., Moffat, K., Šrajer, V. and Stoddard, B.L. (1999) 'Laue crystallography: coming of age', *Journal of Synchrotron Radiation*, 6(4), pp. 891–917. doi:10.1107/S0909049599006366.

Rocco, A.G., Mollica, L., Ricchiuto, P., Baptista, A.M., Gianazza, E. and Eberini, I. (2008) 'Characterization of the Protein Unfolding Processes Induced by Urea and Temperature', *Biophysical Journal*, 94(6), pp. 2241–2251. doi:https://doi.org/10.1529/biophysj.107.115535.

Roche, D.B. McGuffin, L.. (2015) 'In silico identification and characterization of protein-ligand 611 binding sites, methods in molecular biology', in Stoddard, B. Baker, D. (ed.) *Structure based and computational design of 612 ligand binding proteins*. Humana Press.

Roche, D.B., Brackenridge, D.A. and McGuffin, L.J. (2015) 'Proteins and their interacting partners: An introduction to protein-ligand binding site prediction methods', *International Journal of Molecular Sciences*, 16(12). doi:10.3390/ijms161226202.

Roche, D.B., Buenavista, M.T., Harwell, M. and Mcguffin, L.J. (2012) 'PREDICTING PROTEIN STRUCTURES AND STRUCTURAL ANNOTATION OF PROTEOMES Affiliation', in Roberts, G.C.. (ed.) *Encyclopedia of biophysics*. Springer: Berling and Heidelberg, p. 469.

Roche, D.B., Buenavista, M.T. and McGuffin, L.J. (2012) 'FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions.', *PloS one*, 7(5), p. e38219. doi:10.1371/journal.pone.0038219.

Roche, D.B., Buenavista, M.T. and McGuffin, L.J. (2013) ‘The FunFOLD2 server for the prediction of protein-ligand interactions.’, *Nucleic acids research*, 41(Web Server issue), pp. W303-7. doi:10.1093/nar/gkt498.

Roche, D.B., Buenavista, M.T., Tetchner, S.J. and McGuffin, L.J. (2011) ‘The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction.’, *Nucleic acids research*, 39(Web Server issue), pp. W171-6. doi:10.1093/nar/gkr184.

Roche, D.B. and McGuffin, L.J. (2016) ‘In silico Identification and Characterization of Protein-Ligand Binding Sites.’, *Methods in molecular biology (Clifton, N.J.)*, 1414, pp. 1–21. doi:10.1007/978-1-4939-3569-7_1.

Roche, D.B., Tetchner, S.J. and McGuffin, L.J. (2010) ‘The binding site distance test score: a robust method for the assessment of predicted protein binding sites.’, *Bioinformatics*, 26(22), pp. 2920–2921. doi:10.1093/bioinformatics/btq543.

Roche, D.B., Tetchner, S.J. and McGuffin, L.J. (2011) ‘FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins.’, *BMC bioinformatics*, 12, p. 160. doi:10.1186/1471-2105-12-160.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. and Baker, D. (2004) ‘Protein Structure Prediction Using Rosetta’, in, pp. 66–93. doi:10.1016/S0076-6879(04)83004-0.

Romano, M., Ruggiero, A., Squeglia, F., Maga, G. and Berisio, R. (2020) ‘A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping.’, *Cells*, 9(5). doi:10.3390/cells9051267.

Rosas-Lemus, M., Minasov, G., Shuvalova, L., Inniss, N.L., Kiryukhina, O., Brunzelle, J. and Satchell, K.J.F. (2020) ‘High-resolution structures of the SARS-CoV-2 2’-O-methyltransferase reveal strategies for structure-based inhibitor design.’, *Science signaling*, 13(651). doi:10.1126/scisignal.abe1202.

Routledge, P., Vale, J.A., Bateman, D.N., Johnston, G.D., Jones, A., Judd, A., Thomas, S., Volans, G., Prescott, L.F. and Proudfoot, A. (1998) ‘Paracetamol (acetaminophen) poisoning. No need to change current guidelines to accident departments.’, *BMJ (Clinical research ed.)*, 317(7173), pp. 1609–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9848898.

Rovira, P., Kurz-Besson, C., Hernàndez, P., Coûteaux, M.-M. and Vallejo, V.R. (2008) ‘Searching for an indicator of N evolution during organic matter decomposition based on amino acids behaviour: a study on litter layers of pine forests’, *Plant and Soil*, 307(1–2), pp. 149–166. doi:10.1007/s11104-008-9592-6.

Rowinsky, E.K. and Donehower, R.C. (1991) ‘The clinical pharmacology and use of antimicrotubule agents in cancer chemotherapeutics.’, *Pharmacology & therapeutics*, 52(1), pp. 35–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1687171.

Roy, A., Yang, J. and Zhang, Y. (2012) ‘COFACTOR: an accurate comparative algorithm for structure-based protein function annotation’, *Nucleic Acids Research*, 40(W1), pp. W471–W477. doi:10.1093/nar/gks372.

Roy, A. and Zhang, Y. (2012) ‘Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement.’, *Structure (London, England : 1993)*, 20(6), pp. 987–97. doi:10.1016/j.str.2012.03.009.

Rubin, I. and Yarden, Y. (2001) ‘The basic biology of HER2.’, *Annals of oncology : official journal of the European Society for Medical Oncology*, 12 Suppl 1, pp. S3-8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11521719.

Ruhe, Z.C., Nguyen, J.Y., Xiong, J., Koskiniemi, S., Beck, C.M., Perkins, B.R., Low, D.A. and Hayes, C.S. (2017) ‘CdiA Effectors Use Modular Receptor-Binding Domains To Recognize Target Bacteria’, *mBio*. Edited by M.T. Laub, 8(2). doi:10.1128/mBio.00290-17.

Rydberg, P. and Olsen, L. (2012) ‘Predicting drug metabolism by cytochrome P450 2C9: comparison with the 2D6 and 3A4 isoforms.’, *ChemMedChem*, 7(7), pp. 1202–9. doi:10.1002/cmdc.201200160.

Ryu, W.-S. (2017) ‘Chapter 2 - Virus Structure’, in Ryu, W.-S.B.T.-M.V. of H.P.V. (ed.). Boston: Academic Press, pp. 21–29. doi:https://doi.org/10.1016/B978-0-12-800838-6.00002-3.

Saber-Ayad, M., Saleh, M.A. and Abu-Gharbieh, E. (2020) ‘The Rationale for Potential Pharmacotherapy of COVID-19’, *Pharmaceuticals*, 13(5), p. 96. doi:10.3390/ph13050096.

Sahraeian, S.M., Luo, K.R. and Brenner, S.E. (2015) ‘SIFTER search: a web server for accurate phylogeny-based protein function prediction.’, *Nucleic acids research*, 43(W1), pp. W141-7. doi:10.1093/nar/gkv461.

Saibil, H. (2013) ‘Chaperone machines for protein folding, unfolding and disaggregation’, *Nature Reviews Molecular Cell Biology*, 14(10), pp. 630–642. doi:10.1038/nrm3658.

Salentin, S., Schreiber, S., Haupt, V.J., Adasme, M.F. and Schroeder, M. (2015) ‘PLIP: fully automated protein-ligand interaction profiler.’, *Nucleic acids research*, 43(W1), pp. W443-7. doi:10.1093/nar/gkv315.

San Juan, I., Bruzzone, C., Bizkarguenaga, M., Bernardo-Seisdedos, G., Laín, A., Gil-Redondo, R., Diercks, T., Gil-Martínez, J., Urquiza, P., Arana, E., Seco, M., García de Vicuña, A., Embade, N., Mato, J.M. and Millet, O. (2020) ‘Abnormal concentration of porphyrins in serum from COVID-19 patients’, *British Journal of Haematology*, 190(5). doi:10.1111/bjh.17060.

Sánchez, R. and Sali, A. (1998) ‘Large-scale protein structure modeling of the Saccharomyces cerevisiae genome.’, *Proceedings of the National Academy of Sciences of the United States of America*, 95(23), pp. 13597–602. doi:10.1073/pnas.95.23.13597.

Sanger, F. and Tuppy, H. (1951) ‘The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates.’, *The Biochemical journal*, 49(4), pp. 463–81. doi:10.1042/bj0490463.

Santerre, M., Arjona, S.P., Allen, C.N., Shcherbik, N. and Sawaya, B.E. (2020) ‘Why do SARS-CoV-2 NSPs rush to the ER?’, *Journal of Neurology* [Preprint]. doi:10.1007/s00415-020-10197-8.

Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I. and Overington, J.P. (2017) ‘A comprehensive map of molecular drug targets.’, *Nature reviews. Drug discovery*, 16(1), pp. 19–34. doi:10.1038/nrd.2016.230.

Sanvictores, Terrence; Farci, F. (2020) ‘Biochemistry, Primary Protein Structure’, in *StatPearls*. Florida: StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK564343/.

Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. and Karsch-Mizrachi, I. (2019) ‘GenBank’, *Nucleic Acids Research*, 47(D1), pp. D94–D99. doi:10.1093/nar/gky989.

Scanlan, E., Ardill, L., Whelan, M.V.X., Shortt, C., Nally, J.E., Bourke, B. and Ó Cróinín, T. (2017) ‘Relaxation of DNA supercoiling leads to increased invasion of epithelial cells and protein secretion by Campylobacter jejuni.’, *Molecular microbiology*, 104(1), pp. 92–104. doi:10.1111/mmi.13614.

Schaecher, S.R., Mackenzie, J.M. and Pekosz, A. (2007) ‘The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles.’, *Journal of virology*, 81(2), pp. 718–31. doi:10.1128/JVI.01691-06.

Schlichting, I. (2005) ‘X-Ray Crystallography of Protein–Ligand Interactions’, in *Protein-Ligand Interactions*. New Jersey: Humana Press, pp. 155–166. doi:10.1385/1-59259-912-5:155.

Schmidt, T., Haas, J., Gallo Cassarino, T. and Schwede, T. (2011) ‘Assessment of ligand-binding residue predictions in CASP9.’, *Proteins*, 79 Suppl 1, pp. 126–36. doi:10.1002/prot.23174.

Schroeder, H.W. and Cavacini, L. (2010) ‘Structure and function of immunoglobulins.’, *The Journal of allergy and clinical immunology*, 125(2 Suppl 2), pp. S41-52. doi:10.1016/j.jaci.2009.09.046.

Schubert, K., Karousis, E.D., Jomaa, A., Scaiola, A., Echeverria, B., Gurzeler, L.-A., Leibundgut, M., Thiel, V., Mühlemann, O. and Ban, N. (2020) 'SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation', *Nature Structural & Molecular Biology*, 27(10), pp. 959–966. doi:10.1038/s41594-020-0511-8.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N. V, Dunbrack, R.L., Fidelis, K., Fiser, A., Godzik, A., Huang, Y.J., Humblet, C., Jacobson, M.P., Joachimiak, A., Krystek, S.R., *et al.* (2009) 'Outcome of a workshop on applications of protein models in biomedical research.', *Structure (London, England : 1993)*, 17(2), pp. 151–9. doi:10.1016/j.str.2008.12.014.

Sedgwick, P. (2015) 'How to read a receiver operating characteristic curve.', *BMJ (Clinical research ed.)*, 350, p. h2464. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25956305.

Seeliger, D. and de Groot, B.L. (2010) 'Conformational transitions upon ligand binding: holo-structure prediction from apo conformations.', *PLoS computational biology*, 6(1), p. e1000634. doi:10.1371/journal.pcbi.1000634.

Seif, F., Aazami, H., Khoshmirsafa, M., Kamali, M., Mohsenzadegan, M., Pornour, M. and Mansouri, D. (2020) 'JAK Inhibition as a New Treatment Strategy for Patients with COVID-19', *International Archives of Allergy and Immunology*, 181(6), pp. 467–475. doi:10.1159/000508247.

Seifried, A., Schultz, J. and Gohla, A. (2013) 'Human HAD phosphatases: structure, mechanism, and roles in health and disease.', *The FEBS journal*, 280(2), pp. 549–71. doi:10.1111/j.1742-4658.2012.08633.x.

Serrano-Posada, H., Valderrama, B., Stojanoff, V. and Rudiño-Piñera, E. (2011) 'Thermostable multicopper oxidase from Thermus thermophilus HB27: crystallization and preliminary X-ray diffraction analysis of apo and holo forms.', *Acta crystallographica. Section F, Structural biology and crystallization communications*, 67(Pt 12), pp. 1595–8. doi:10.1107/S174430911103805X.

Sh Lebedeva, N., A Gubarev, Y., O Koifman, M. and I Koifman, O. (2020) 'The Application of Porphyrins and Their Analogues for Inactivation of Viruses.', *Molecules (Basel, Switzerland)*, 25(19). doi:10.3390/molecules25194368.

Shahiduzzaman, M. and Coombs, K.M. (2012) 'Activity based protein profiling to detect serine hydrolase alterations in virus infected cells', *Frontiers in Microbiology*, 3. doi:10.3389/fmicb.2012.00308.

Shalmali, N., Ali, M.R. and Bawa, S. (2018) 'Imidazole: An Essential Edifice for the Identification of New Lead Compounds and Drug Development.', *Mini reviews in medicinal chemistry*, 18(2), pp. 142–163. doi:10.2174/1389557517666170228113656.

Shangary, S. and Wang, S. (2009) 'Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy.', *Annual review of pharmacology and toxicology*, 49, pp. 223–41. doi:10.1146/annurev.pharmtox.48.113006.094723.

Shi, J., Zhao, Y., Wang, K., Shi, X., Wang, Y., Huang, H., Zhuang, Y., Cai, T., Wang, F. and Shao, F. (2015) 'Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death.', *Nature*, 526(7575), pp. 660–5. doi:10.1038/nature15514.

Shi, L. and Tu, B.P. (2015) 'Acetyl-CoA and the regulation of metabolism: mechanisms and consequences.', *Current opinion in cell biology*, 33, pp. 125–31. doi:10.1016/j.ceb.2015.02.003.

Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., Schulz, L., Widera, M., Mehdipour, A.R., Tascher, G., Geurink, P.P., Wilhelm, A., van der Heden van Noort, G.J., Ovaa, H., Müller, S., Knobeloch, K.-P., Rajalingam, K., Schulman, B.A., Cinatl, J., *et al.* (2020) 'Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity', *Nature*, 587(7835), pp. 657–662. doi:10.1038/s41586-020-2601-5.

Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. (2000) 'MaxSub: an automated measure for the assessment of protein structure prediction quality', *Bioinformatics*, 16(9), pp. 776–785. doi:10.1093/bioinformatics/16.9.776.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ‘ROCR: visualizing classifier performance in R.’, *Bioinformatics (Oxford, England)*, 21(20), pp. 3940–1. doi:10.1093/bioinformatics/bti623.

Sippl, M.J., Lackner, P., Domingues, F.S., Prlić, A., Malik, R., Andreeva, A. and Wiederstein, M. (2001) ‘Assessment of the CASP4 fold recognition category.’, *Proteins*, Suppl 5, pp. 55–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11835482.

Siu, K.-L., Yuen, K.-S., Castaño-Rodriguez, C., Ye, Z.-W., Yeung, M.-L., Fung, S.-Y., Yuan, S., Chan, C.-P., Yuen, K.-Y., Enjuanes, L. and Jin, D.-Y. (2019) ‘Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC.’, *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 33(8), pp. 8865–8877. doi:10.1096/fj.201802418R.

Siu, Y.L., Teoh, K.T., Lo, J., Chan, C.M., Kien, F., Escriou, N., Tsao, S.W., Nicholls, J.M., Altmeyer, R., Peiris, J.S.M., Bruzzone, R. and Nal, B. (2008) ‘The M, E, and N Structural Proteins of the Severe Acute Respiratory Syndrome Coronavirus Are Required for Efficient Assembly, Trafficking, and Release of Virus-Like Particles’, *Journal of Virology*, 82(22), pp. 11318–11330. doi:10.1128/JVI.01052-08.

Skipper, L. (2005) ‘PROTEINS | Overview☆’, in Worsfold, P., Poole, C., Townshend, A., and Miró, M. (eds) *Encyclopedia of Analytical Science (Third Edition)*. Third Edit. Oxford: Academic Press, pp. 412–419. doi:https://doi.org/10.1016/B978-0-08-101983-2.00493-X.

Skolnick, J. and Brylinski, M. (2009) ‘FINDSITE: a combined evolution/structure-based approach to protein function prediction.’, *Briefings in bioinformatics*, 10(4), pp. 378–91. doi:10.1093/bib/bbp017.

Smallman, R.E. and Ngan, A.H.W. (2014) ‘Chapter 5 - Characterization and Analysis’, in Smallman, R.E. and Ngan, A.H.W.B.T.-M.P.M. (Eighth E. (eds). Oxford: Butterworth-Heinemann, pp. 159–250. doi:https://doi.org/10.1016/B978-0-08-098204-5.00005-5.

Söding, J. (2005) ‘Protein homology detection by HMM–HMM comparison’, *Bioinformatics*, 21(7), pp. 951–960. doi:10.1093/bioinformatics/bti125.

Souza, P.C.T., Limongelli, V., Wu, S., Marrink, S.J. and Monticelli, L. (2021) ‘Perspectives on High-Throughput Ligand/Protein Docking With Martini MD Simulations’, *Frontiers in Molecular Biosciences*, 8. doi:10.3389/fmolb.2021.657222.

Sparks, J.L., Kumar, R., Singh, M., Wold, M.S., Pandita, T.K. and Burgers, P.M. (2012) ‘Human exonuclease 5 is a novel sliding exonuclease required for genome stability.’, *The Journal of biological chemistry*, 287(51), pp. 42773–83. doi:10.1074/jbc.M112.422444.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J. and Söding, J. (2019) ‘HH-suite3 for fast remote homology detection and deep protein annotation’, *BMC Bioinformatics*, 20(1), p. 473. doi:10.1186/s12859-019-3019-7.

Stoker, H.S. (Howard S. (2016) *General, organic, and biological chemistry / H. Stephen Stoker.* Seventh ed, *General, organic, and biological chemistry*. Seventh ed. Boston, MA: Cengage Learning.

Su, Y.C.F., Anderson, D.E., Young, B.E., Linster, M., Zhu, F., Jayakumar, J., Zhuang, Y., Kalimuddin, S., Low, J.G.H., Tan, C.W., Chia, W.N., Mak, T.M., Octavia, S., Chavatte, J.-M., Lee, R.T.C., Pada, S., Tan, S.Y., Sun, L., Yan, G.Z., *et al.* (2020) ‘Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2.’, *mBio*, 11(4). doi:10.1128/mBio.01610-20.

Tan, Yong-Jun; Schneider Theresa; Shukla, Prakash K; Chandrasekharan Mahesh B; Aravind, L; Zhang, D. (2020) ‘Unification of the M/ORF3-related proteins points to a diversified role for ion conductance in pathogenesis of coronaviruses and other nidoviruses’, *bioRxiv* [Preprint]. doi:10.1101/2020.11.10.377366.

Tau, G. and Rothman, P. (1999) ‘Biologic functions of the IFN-gamma receptors.’, *Allergy*, 54(12), pp. 1233–51. doi:10.1034/j.1398-9995.1999.00099.x.

Taylor, R.M., Whitehouse, C.J. and Caldecott, K.W. (2000) ‘The DNA ligase III zinc finger stimulates binding to DNA secondary structure and promotes end joining.’, *Nucleic acids* *research*, 28(18), pp. 3558–63. doi:10.1093/nar/28.18.3558.

The Protein Prediction Center (no date).

Thomas, S. (2020) ‘The Structure of the Membrane Protein of SARS-CoV-2 Resembles the Sugar Transporter SemiSWEET.’, *Pathogens & immunity*, 5(1), pp. 342–363. doi:10.20411/pai.v5i1.377.

Thong, S., Ercan, B., Torta, F., Fong, Z.Y., Wong, H.Y.A., Wenk, M.R. and Chng, S.-S. (2016) ‘Defining key roles for auxiliary proteins in an ABC transporter that maintains bacterial outer membrane lipid asymmetry.’, *eLife*, 5. doi:10.7554/eLife.19042.

Tivol, W.F., Briegel, A. and Jensen, G.J. (2008) ‘An Improved Cryogen for Plunge Freezing’, *Microscopy and Microanalysis*, 14(5), pp. 375–379. doi:10.1017/S1431927608080781.

Trott, O. and Olson, A.J. (2010) ‘AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.’, *Journal of computational chemistry*, 31(2), pp. 455–61. doi:10.1002/jcc.21334.

Tseng, Y.-T., Wang, S.-M., Huang, K.-J., Lee, A.I.-R., Chiang, C.-C. and Wang, C.-T. (2010) ‘Self-assembly of Severe Acute Respiratory Syndrome Coronavirus Membrane Protein’, *Journal of Biological Chemistry*, 285(17), pp. 12862–12872. doi:10.1074/jbc.M109.030270.

Tsujikawa, H., Sato, K., Wei, C., Saad, G., Sumikoshi, K., Nakamura, S., Terada, T. and Shimizu, K. (2016) ‘Development of a protein–ligand-binding site prediction method based on interaction energy and sequence conservation’, *Journal of Structural and Functional Genomics*, 17(2), pp. 39–49. doi:10.1007/s10969-016-9204-2.

UniProt Consortium (2019) ‘UniProt: a worldwide hub of protein knowledge.’, *Nucleic acids research*, 47(D1), pp. D506–D515. doi:10.1093/nar/gky1049.

Uversky, V.N. (2011) ‘Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes’, *Chem. Soc. Rev.*, 40(3), pp. 1623–1634. doi:10.1039/C0CS00057D.

Valdés, J., Pedroso, I., Quatrini, R., Dodson, R.J., Tettelin, H., Blake, R., Eisen, J.A. and Holmes, D.S. (2008) ‘Acidithiobacillus ferrooxidans metabolism: from genome sequence to industrial applications.’, *BMC genomics*, 9, p. 597. doi:10.1186/1471-2164-9-597.

Venkatachalam, C.M., Jiang, X., Oldfield, T. and Waldman, M. (2003) ‘LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.’, *Journal of molecular graphics & modelling*, 21(4), pp. 289–307. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12479928.

Viet Hung, L., Caprari, S., Bizai, M., Toti, D. and Polticelli, F. (2015) ‘LIBRA: LIgand Binding site Recognition Application’, *Bioinformatics*, p. btv489. doi:10.1093/bioinformatics/btv489.

Villarroya-Beltri, C., Guerra, S. and Sánchez-Madrid, F. (2017) ‘ISGylation - a key to lock the cell gates for preventing the spread of threats.’, *Journal of cell science*, 130(18), pp. 2961–2969. doi:10.1242/jcs.205468.

Viswanathan, T., Misra, A., Chan, S.-H., Qi, S., Dai, N., Arya, S., Martinez-Sobrido, L. and Gupta, Y.K. (2021) ‘A metal ion orients SARS-CoV-2 mRNA to ensure accurate 2′-O methylation of its first nucleotide’, *Nature Communications*, 12(1), p. 3287. doi:10.1038/s41467-021-23594-y.

Volkamer, A., Griewel, A., Grombacher, T. and Rarey, M. (2010) ‘Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets’, *Journal of Chemical Information and Modeling*, 50(11), pp. 2041–2052. doi:10.1021/ci100241y.

de Vries, S.J., van Dijk, M. and Bonvin, A.M.J.J. (2010) ‘The HADDOCK web server for data-driven biomolecular docking’, *Nature Protocols*, 5(5), pp. 883–897. doi:10.1038/nprot.2010.32.

Wang, H. (2015) ‘Cryo-electron microscopy for structural biology: current status and future perspectives’, *Science China Life Sciences*, 58(8), pp. 750–756. doi:10.1007/s11427-015-4851-2.

Wang, R., Fang, X., Lu, Y. and Wang, S. (2004) ‘The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.’, *Journal of medicinal chemistry*, 47(12), pp. 2977–80. doi:10.1021/jm030580l.

Wang, T., Mori, H., Zhang, C., Kurokawa, K., Xing, X.-H. and Yamada, T. (2015) 'DomSign: a top-down annotation pipeline to enlarge enzyme space in the protein universe.', *BMC bioinformatics*, 16, p. 96. doi:10.1186/s12859-015-0499-y.

Wass, M.N., Kelley, L.A. and Sternberg, M.J.E. (2010) '3DLigandSite: predicting ligand-binding sites using similar structures.', *Nucleic acids research*, 38(Web Server issue), pp. W469-73. doi:10.1093/nar/gkq406.

Webb, B. and Sali, A. (2016) 'Comparative Protein Structure Modeling Using MODELLER', *Current Protocols in Bioinformatics*, 54(1). doi:10.1002/cpbi.3.

Webb, B., Viswanath, S., Bonomi, M., Pellarin, R., Greenberg, C.H., Saltzberg, D. and Sali, A. (2018) 'Integrative structure modeling with the Integrative Modeling Platform', *Protein Science*, 27(1), pp. 245–258. doi:10.1002/pro.3311.

Welch, W.J. (2004) 'Role of quality control pathways in human diseases involving protein misfolding', *Seminars in Cell & Developmental Biology*, 15(1), pp. 31–38. doi:10.1016/j.semcdb.2003.12.011.

wenzhong, liu; hualan, L. (2020) 'COVID-19:Attacks the 1-Beta Chain of Hemoglobin and Captures the Porphyrin to Inhibit Human Heme Metabolism', *ChemRxiv* [Preprint]. doi:doi.org/10.26434/chemrxiv.11938173.v9.

Westerink, W.M.A., Stevenson, J.C.R. and Schoonen, W.G.E.J. (2008) 'Pharmacologic profiling of human and rat cytochrome P450 1A1 and 1A2 induction and competition.', *Archives of toxicology*, 82(12), pp. 909–21. doi:10.1007/s00204-008-0317-7.

Willett, P. (2014) 'The Calculation of Molecular Structural Similarity: Principles and Practice.', *Molecular informatics*, 33(6–7), pp. 403–13. doi:10.1002/minf.201400024.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., *et al.* (2018) 'DrugBank 5.0: a major update to the DrugBank database for 2018.', *Nucleic acids research*, 46(D1), pp. D1074–D1082. doi:10.1093/nar/gkx1037.

Wolf-Yadlin, A., Kumar, N., Zhang, Y., Hautaniemi, S., Zaman, M., Kim, H.-D., Grantcharova, V., Lauffenburger, D.A. and White, F.M. (2006) 'Effects of HER2 overexpression on cell signaling networks governing proliferation and migration.', *Molecular systems biology*, 2, p. 54. doi:10.1038/msb4100094.

World Health Organisation (2021). Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Wrzeszczynski, K.O., Ofran, Y., Rost, B., Nair, R. and Liu, J. (2003) 'Automatic prediction of protein function', *Cellular and Molecular Life Sciences (CMLS)*, 60(12), pp. 2637–2650. doi:10.1007/s00018-003-3114-8.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C. and Zhang, Y.-Z. (2020) 'A new coronavirus associated with human respiratory disease in China', *Nature*, 579(7798), pp. 265–269. doi:10.1038/s41586-020-2008-3.

Wu, G. (2009) 'Amino acids: metabolism, functions, and nutrition', *Amino Acids*, 37(1), pp. 1–17. doi:10.1007/s00726-009-0269-0.

Wu, Q., Peng, Z., Zhang, Y. and Yang, J. (2018) 'COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking.', *Nucleic acids research*, 46(W1), pp. W438–W442. doi:10.1093/nar/gky439.

Xia, H., Cao, Z., Xie, X., Zhang, X., Chen, J.Y.-C., Wang, H., Menachery, Vineet D, Rajsbaum, R. and Shi, P.-Y. (2020) 'Evasion of Type I Interferon by SARS-CoV-2.', *Cell reports*, 33(1), p. 108234. doi:10.1016/j.celrep.2020.108234.

Xia, H., Cao, Z., Xie, X., Zhang, X., Chen, J.Y.-C., Wang, H., Menachery, Vineet D., Rajsbaum, R. and Shi, P.-Y. (2020) 'Evasion of Type I Interferon by SARS-CoV-2', *Cell Reports*, 33(1), p. 108234. doi:10.1016/j.celrep.2020.108234.

Xiao, W., Wang, R.-S., Handy, D.E. and Loscalzo, J. (2018) 'NAD(H) and NADP(H) Redox

Couples and Cellular Energy Metabolism.', *Antioxidants & redox signaling*, 28(3), pp. 251–272. doi:10.1089/ars.2017.7216.

Xie, G., Bonner, C.A. and Jensen, R.A. (2000) 'Cyclohexadienyl dehydrogenase from Pseudomonas stutzeri exemplifies a widespread type of tyrosine-pathway dehydrogenase in the TyrA protein family.', *Comparative biochemistry and physiology. Toxicology & pharmacology : CBP*, 125(1), pp. 65–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11790331.

Xie, Z.-R. and Hwang, M. (2012) 'Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles', *Bioinformatics*, 28(12), pp. 1579–1585. doi:10.1093/bioinformatics/bts182.

Xing, S., Wallmeroth, N., Berendzen, K.W. and Grefen, C. (2016) 'Techniques for the Analysis of Protein-Protein Interactions in Vivo.', *Plant physiology*, 171(2), pp. 727–58. doi:10.1104/pp.16.00470.

Xu, D. and Zhang, Y. (2012) 'Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.', *Proteins*, 80(7), pp. 1715–35. doi:10.1002/prot.24065.

Xu, J. and Zhang, Y. (2010) 'How significant is a protein structure similarity with TM-score = 0.5?', *Bioinformatics (Oxford, England)*, 26(7), pp. 889–95. doi:10.1093/bioinformatics/btq066.

Yamanishi, M., Kinoshita, K., Fukuoka, M., Saito, T., Tanokuchi, A., Ikeda, Y., Obayashi, H., Mori, K., Shibata, N., Tobimatsu, T. and Toraya, T. (2012) 'Redesign of coenzyme B(12) dependent diol dehydratase to be resistant to the mechanism-based inactivation by glycerol and act on longer chain 1,2-diols.', *The FEBS journal*, 279(5), pp. 793–804. doi:10.1111/j.1742-4658.2012.08470.x.

Yang, J., Roy, A. and Zhang, Y. (2012) 'BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions', *Nucleic Acids Research*, 41(D1), pp. D1096–D1103. doi:10.1093/nar/gks966.

Yang, J., Roy, A. and Zhang, Y. (2013) 'BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions', *Nucleic Acids Research*, 41(D1), pp. D1096–D1103. doi:10.1093/nar/gks966.

Yang, Jianyi, Roy, A. and Zhang, Y. (2013) 'Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment.', *Bioinformatics*, 29(20), pp. 2588–2595. doi:10.1093/bioinformatics/btt447.

Yang, J. and Zhang, Y. (2015) 'I-TASSER server: new development for protein structure and function predictions.', *Nucleic acids research*, 43(W1), pp. W174-81. doi:10.1093/nar/gkv342.

Yoshimoto, F.K. (2020) 'The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19', *The Protein Journal*, 39(3), pp. 198–216. doi:10.1007/s10930-020-09901-4.

Young, B.E., Fong, S.-W., Chan, Y.-H., Mak, T.-M., Ang, L.W., Anderson, D.E., Lee, C.Y.-P., Amrun, S.N., Lee, B., Goh, Y.S., Su, Y.C.F., Wei, W.E., Kalimuddin, S., Chai, L.Y.A., Pada, S., Tan, S.Y., Sun, L., Parthasarathy, P., Chen, Y.Y.C., *et al.* (2020) 'Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study', *The Lancet*, 396(10251), pp. 603–611. doi:10.1016/S0140-6736(20)31757-8.

Yu, G., Zhu, H. and Domeniconi, C. (2015) 'Predicting protein functions using incomplete hierarchical labels.', *BMC bioinformatics*, 16, p. 1. doi:10.1186/s12859-014-0430-y.

Yu, J., Zhou, Y., Tanaka, I. and Yao, M. (2010) 'Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere', *Bioinformatics*, 26(1), pp. 46–52. doi:10.1093/bioinformatics/btp599.

Yu, L., Wang, L. and Chen, S. (2010) 'Endogenous toll-like receptor ligands and their biological significance.', *Journal of cellular and molecular medicine*, 14(11), pp. 2592–603. doi:10.1111/j.1582-4934.2010.01127.x.

Zasowski, E.J., Rybak, J.M. and Rybak, M.J. (2015) 'The β-Lactams Strike Back: Ceftazidime-Avibactam.', *Pharmacotherapy*, 35(8), pp. 755–70. doi:10.1002/phar.1622.

Zemla, A., Venclovas, ?eslovas, Moult, J. and Fidelis, K. (2001) 'Processing and evaluation of predictions in CASP4', *Proteins: Structure, Function, and Genetics*, 45(S5), pp. 13–21. doi:10.1002/prot.10052.

Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. (1999) 'Processing and analysis of CASP3 protein structure predictions.', *Proteins*, Suppl 3, pp. 22–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10526349.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. and Hilgenfeld, R. (2020) 'Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors.', *Science (New York, N.Y.)*, 368(6489), pp. 409–412. doi:10.1126/science.abb3405.

Zhang, L. and Liu, Y. (2020) 'Potential interventions for novel coronavirus in China: A systematic review', *Journal of medical virology*. 2020/03/03, 92(5), pp. 479–490. doi:10.1002/jmv.25707.

Zhang, W., Bell, E.W., Yin, M. and Zhang, Y. (2020) 'EDock: blind protein–ligand docking by replica-exchange monte carlo simulation', *Journal of Cheminformatics*, 12(1), p. 37. doi:10.1186/s13321-020-00440-9.

Zhang, Y. (2007) 'Template-based modeling and free modeling by I-TASSER in CASP7', *Proteins: Structure, Function, and Bioinformatics*, 69(S8), pp. 108–117. doi:10.1002/prot.21702.

Zhang, Y. and Skolnick, J. (2005) 'TM-align: a protein structure alignment algorithm based on the TM-score.', *Nucleic acids research*, 33(7), pp. 2302–9. doi:10.1093/nar/gki524.

Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., Yang, T., Yu, F., Liu, J., Liu, B., Song, Z., Chen, J., Pan, T., Zhang, X., Li, Y., Li, R., Huang, W., Xiao, F. and Zhang, H. (2020) 'The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I', *bioRxiv*, p. 2020.05.24.111823. doi:10.1101/2020.05.24.111823.

Zhang, Z., Li, Y., Lin, B., Schroeder, M. and Huang, B. (2011) 'Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction', *Bioinformatics*, 27(15), pp. 2083–2088. doi:10.1093/bioinformatics/btr331.

Zhao, B., Lei, L., Vassylyev, D.G., Lin, X., Cane, D.E., Kelly, S.L., Yuan, H., Lamb, D.C. and Waterman, M.R. (2009) 'Crystal structure of albaflavenone monooxygenase containing a moonlighting terpene synthase active site.', *The Journal of biological chemistry*, 284(52), pp. 36711–9. doi:10.1074/jbc.M109.064683.

Zhao, J., Cao, Y. and Zhang, L. (2020) 'Exploring the computational methods for protein-ligand binding site prediction', *Computational and Structural Biotechnology Journal*, 18, pp. 417–426. doi:https://doi.org/10.1016/j.csbj.2020.02.008.

Zhao, J., Falcón, A., Zhou, H., Netland, J., Enjuanes, L., Pérez Breña, P. and Perlman, S. (2009) 'Severe acute respiratory syndrome coronavirus protein 6 is required for optimal replication.', *Journal of virology*, 83(5), pp. 2368–73. doi:10.1128/JVI.02371-08.

Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S.M. and Zhang, Y. (2019) 'Deep-learning contact-map guided protein structure prediction in CASP13', *Proteins*. 2019/08/14, 87(12), pp. 1149–1164. doi:10.1002/prot.25792.

Zheng, W., Zhang, C., Wuyun, Q., Pearce, R., Li, Y. and Zhang, Y. (2019) 'LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins', *Nucleic Acids Research*, 47(W1), pp. W429–W436. doi:10.1093/nar/gkz384.

Zhou, H., Ferraro, D., Zhao, J., Hussain, S., Shao, J., Trujillo, J., Netland, J., Gallagher, T. and Perlman, S. (2010) 'The N-terminal region of severe acute respiratory syndrome coronavirus protein 6 induces membrane rearrangement and enhances virus replication.', *Journal of virology*, 84(7), pp. 3542–51. doi:10.1128/JVI.02570-09.

Zhou, H. and Skolnick, J. (2007) 'Ab Initio Protein Structure Prediction Using Chunk-TASSER', *Biophysical Journal*, 93(5), pp. 1510–1518. doi:10.1529/biophysj.107.109959.

Zhou, H. and Zhou, Y. (2005) ‘Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments.’, *Proteins*, 58(2), pp. 321–8. doi:10.1002/prot.20308.

Zhou, Ziliang, Huang, C., Zhou, Zhechong, Huang, Z., Su, L., Kang, S., Chen, X., Chen, Q., He, S., Rong, X., Xiao, F., Chen, J. and Chen, S. (2021) ‘Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14+ monocytes’, *iScience*, 24(3), p. 102187. doi:10.1016/j.isci.2021.102187.

Zhu, H. and Pisabarro, M.T. (2011) ‘MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets’, *Bioinformatics*, 27(3), pp. 351–358. doi:10.1093/bioinformatics/btq672.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N. and Alva, V. (2018) ‘A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core’, *Journal of Molecular Biology*, 430(15), pp. 2237–2243. doi:https://doi.org/10.1016/j.jmb.2017.12.007.

Zinzula, L. (2020) ‘Lost in deletion: The enigmatic ORF8 protein of SARS-CoV-2’, *Biochemical and Biophysical Research Communications* [Preprint]. doi:https://doi.org/10.1016/j.bbrc.2020.10.045.

Zoll, W.L., Horton, L.E., Komar, A.A., Hensold, J.O. and Merrick, W.C. (2002) ‘Characterization of mammalian eIF2A and identification of the yeast homolog.’, *The Journal of biological chemistry*, 277(40), pp. 37079–87. doi:10.1074/jbc.M207109200.

Appendices

Appendix 1

Publications to date

1. Int J Mol Sci. 2015 Dec 15;16(12):29829-42. doi: 10.3390/ijms161226202.

Proteins and Their Interacting Partners: An Introduction to Protein-Ligand Binding Site Prediction Methods.

Roche DB[1,2], **Brackenridge DA[3]**, McGuffin LJ[4].

Author information

[1]Institut de Biologie Computationnelle, LIRMM, CNRS, Université de Montpellier, Montpellier 34095, France. daniel.roche@lirmm.fr.

[2]Centre de Recherche de Biochimie Macromoléculaire, CNRS-UMR 5237, Montpellier 34293, France. daniel.roche@lirmm.fr.

[3]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK. d.a.brackenridge@pgr.reading.ac.uk.

[4]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK. l.j.mcguffin@reading.ac.uk.

Abstract

Elucidating the biological and biochemical roles of proteins, and subsequently determining their interacting partners, can be difficult and time consuming using in vitro and/or in vivo methods, and consequently the majority of newly sequenced proteins will have unknown structures and functions. However, in silico methods for predicting protein-ligand binding sites and protein biochemical functions offer an alternative practical solution. The characterisation of protein-ligand binding sites is essential for investigating new functional roles, which can impact the major biological research spheres of health, food, and energy security. In this review we discuss the role in silico methods play in 3D modelling of protein-ligand binding sites, along with their role in predicting biochemical functionality. In addition, we describe in detail some of the key alternative in silico prediction approaches that are available, as well as discussing the Critical Assessment of Techniques for Protein Structure Prediction (CASP) and the Continuous Automated Model EvaluatiOn (CAMEO) projects, and their impact on developments in the field. Furthermore, we discuss the importance of protein function prediction methods for tackling 21st century problems.

2. Abstract B-161 at International Society for Computational Biology: ISCB July 2017

**FunFOLDQ: a fast automated method for the prediction of ligand binding site residues and
Gene Ontology terms**

Brackenridge DA[1], Roche DB[2,3] and McGuffin LJ[1]

[1]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK.
[2]Centre de Recherche en Biologie cellulaire de Montpellier, CNRS-UMR 5237, 34293, Montpellier, France
[3]Institut de Biologie Computationnelle, LIRMM, CNRS, Université de Montpellier, 34095, Montpellier, France.

Protein ligand binding site prediction methods aim to predict, from amino acid sequence, protein-ligand interactions, putative ligands and ligand binding site residues using either sequence information, structural information or a combination of both. In silico characterisation of protein-ligand interactions have become extremely important to help determine a protein functionality, as in vivo based functional elucidation is unable to keep pace with the current growth of sequence databases. Additionally, in vitro biochemical functional elucidation is time consuming, costly and may not be feasible for large scale analysis, such as drug discovery. Thus, in silico prediction of protein-ligand interactions need to be utilized to aid in functional elucidation.

Hence, we developed a structurally informed functional annotation pipeline, called FunFOLDQ, which predicts in silico protein-ligand interactions and Gene Ontology terms. FunFOLDQ, along with its previous implementations, have been ranked amongst the top methods in previous Critical Assessment of Techniques for Protein Structure Prediction (CASP) competitions, ranked 2nd for prediction of “Holo” binding sites in the recent CASP12 competition. We also recently competed in Critical Assessment of protein Function Annotation 3 (CAFA3) challenge. We will present our new methodology and benchmarking results. FunFOLDQ can be used to improve the functional annotation of protein domains, protein dark matter as well as the study of protein-ligand interactions in areas such as rational drug design.

Further output from this abstract is expected in the form a paper that will be published by the CAFA team

3. Abstract at CASP 13

Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement

Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement

L.J. McGuffin[1], R. Adiyaman[1], **D.A. Brakenridge[1]**, J.O. Nealon[1], L.S. Philomina[1] and A.N. Shuid[1,2]
1 - School of Biological Sciences, University of Reading, Reading, UK
2 - Infectomics Cluster, Advanced Medical and Dental Institute, University of Science Malaysia, Pulau Pinang, Malaysia
l.j.mcguffin@reading.ac.uk

For our manual predictions we used several components from our latest servers[1,2,3] (also see our IntFOLD5 and ModFOLD7 server abstracts). For our tertiary structure (TS) predictions we made use of the CASP hosted 3D server models, which we ranked using ModFOLD7_rank and then refined with the our new refinement method (ReFOLD2). For our quaternary structure predictions, we used a docking and template based approach (MultiFOLD) along with our newly developed quality assessment method (ModFOLDdock). Finally, clues from likely ligand binding sites (predicted with FunFOLD3), aided our manual evaluation of submitted models.

Methods

Tertiary structure predictions: The server models were ranked according the ModFOLD7_rank global quality scores (see our ModFOLD7 abstract). The top ranked initial model was then selected and submitted to the ReFOLD2 and MultiFOLD pipelines described below. For each model, the ModFOLD7 predicted per-residue error scores were added into the B-factor column for each set of atom records.

Refinement (ReFOLD2): For the refinement of 3D models of proteins we used a modified version of our automated ReFOLD method[3]. Our new refinement pipeline, ReFOLD2, consisted of three protocols that were similar to the original version. The first protocol used a rapid iterative strategy (i3Drefine[4]) and the second employed a more CPU/GPU intensive molecular dynamic simulation strategy (using NAMD[5]) to refine each starting model.

The major new step for ReFOLD version 2 was the modification of the second protocol, which included the introduction of molecular dynamics simulations that were guided by the per-residue accuracy scores obtained from ModFOLD7. The per-residue accuracy scores were used to identify the poorly predicted regions, which were then targeted for refinement to improve the overall model quality. A new restraint was applied by putting a threshold based on the per-residue accuracy scores (either 2, 3 or 5 Å) during the molecular dynamic simulation. For each starting model, the threshold was determined by considering the distribution of the per-residue accuracy scores.

Refined models generated from the first two protocols were then assessed and ranked using ModFOLD7_rank. The third protocol was a combination of the first 2 approaches, where the top ranked model from the 2nd protocol was then further refined using i3Drefine. Finally, all of the refined models generated by each of these protocols and the starting model were pooled and re-ranked again using ModFOLD7_rank and the final top 5 models were selected and submitted.

Quaternary structure predictions (MultiFOLD): The highest scoring models from the ReFOLD2 procedure, described above, were used to generate predicted quaternary structures using LZerD[6], MEGADOCK[7], FRODOCK[8], PatchDock[9] and ZDOCK[10] for dimeric complexes, and M-ZDOCK[11] and Multi-LZerD[12] for multimeric complexes. In addition to the docking strategy, a multimeric fold recognition approach was also deployed. The fold template lists (with PDB and chain IDs) generated by the IntFOLD server[1] were filtered using multimeric data extracted from PISA[13] for each template. Model assemblies were then constructed using TM-align[14] for structural superposition of tertiary models onto assemblies and PyMOL was used for visualisation and manual quality checking of the template generated models. The final predicted quaternary structures were then ranked for submission using the newly developed ModFOLDdock method described below. Furthermore, the information from our FunFOLD3 method (regarding the function and locations of putative bound ligands) along with visual inspection was used for some targets in order to manually filter the modelled complexes.

Quaternary structure model quality assessment (ModFOLDdock): The ModFOLDdock protocol uses a hybrid consensus approach for producing both global and local (interface residue) scores for predicted quaternary structures. The ModFOLDdock global score was taken as the mean score from four individual methods: ProQDock[15], QSscoreJury, DockQJury and ModFOLDIA. For each interacting pair of chains in a modelled complex, the ProQDock scores were simply taken and averaged to produce a global score for the complete assembly. For the QSscoreJury and DockQJury methods, pairwise comparisons were made for each quaternary structure model to every other model made for the target and then the mean QS[16] and DockQ[17] scores were calculated. The ModFOLDIA method also carries out structure based comparisons of alternative oligomer models and can produce both global and local/per-residue interface scores. The first stage of the ModFOLDIA method was to identify the interface residues in the model to be scored (defined as <= 5Å between the heavy atoms in different chains) and then obtain the minimum contact distance (D_{min}) for each contacting residue. The second stage was to locate the equivalent residues in all other models and then obtain the mean minimum distances of those residues in all other models ($MeanD_{min}$). The final IA score for each of the interface residues in the model was the absolute difference in the S_i from the mean S_i : $IA = 1-|S_i-MeanS_i|$, where $S_i = 1/(1+(D_{min}/20)^2)$ and $MeanS_i = 1/(1+(MeanD_{min}/20)^2)$. The global ModFOLDIA score for a model was then taken as the total interface score (sum of residue scores) normalised by the maximum of either the number of residues in the interface or the mean number of interface residues across all models for the same target.

Availability

Our software will be freely available after publication from:
http://www.reading.ac.uk/bioinf/downloads/
Server methods are available via:
http://www.reading.ac.uk/bioinf/

1. McGuffin,L.J., Atkins,J., Salehe,B.R., Shuid,A.N., Roche,D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.* **43**, W169-73.
2. Maghrabi, A.H.A. & McGuffin, L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. Nucleic Acids Research, 45, W416-W421, doi: 10.1093/nar/gkx332.
3. Shuid, A.N., Kempster, R., McGuffin, L.J. (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. Nucleic Acids Res. 45, W422-W428. doi: 10.1093/nar/gkx249.
4. Bhattacharya,D., Cheng,J. (2013) i3Drefine software for protein 3D structure refinement and its assessment in CASP10. *PLoS One*. **8**, e69648.
5. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L., Schulten,K.J. (2005) Scalable molecular dynamics with NAMD. Comput Chem. **26**, 1781-802.
6. Venkatraman,V., Yang,Y.D., Sael,L., Kihara,D. (2009) Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*. **10**, 407.
7. Ohue,M., Shimoda,T., Suzuki,S., Matsuzaki,Y., Ishida,T., Akiyama,Y. (2014) MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. *Bioinformatics*. **30**, 3281–3283.
8. Garzon,J.I., Lopéz-Blanco,J.R., Pons,C., Kovacs,J., Abagyan,R., Fernandez-Recio,J., Chacon,P. (2009). FRODOCK: a new approach for fast rotational protein–protein docking. *Bioinformatics*. **25**, 2544–2551.
9. Duhovny,D., Nussinov,R., Wolfson,H.J. (2002) Efficient unbound docking of rigid molecules, in: International Workshop on Algorithms in Bioinformatics. Springer, pp. 185–200.
10. Chen,R., Li,L., Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins*. **52**, 80–87.
11. Pierce,B., Tong,W., Weng,Z. (2005) M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics.* **21**, 1472–1478.
12. Esquivel-Rodríguez,J., Yang,Y.D., Kihara,D. (2012) Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins*. **80**, 1818-1833.
13. Krissinel,E., Henrick,K. (2007) Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797.
14. Zhang,Y., Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-9.
15. Basu, S., Wallner, B. (2016) Finding correct protein-protein docking models using ProQDock. Bioinformatics. **32**, i262-i270. doi: 10.1093/bioinformatics/btw257.
16. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. **7**, 10480. doi: 10.1038/s41598-017-09654-8.
17. Basu S1, Wallner B1,2. (2016) DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS One. **11**, e0161879. doi: 10.1371/journal.pone.0161879.

4. Nucleic Acids Res. 2019 May 2. pii: gkz322. doi: 10.1093/nar/gkz322. [Epub ahead of print]

IntFOLD: an integrated web resource for high performance protein structure and function prediction.

McGuffin LJ[1], Adiyaman R[1], Maghrabi AHA[1], Shuid AN[1,2], **Brackenridge DA[1]**, Nealon JO[1], Philomina LS[1].

Author information

[1] School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK.

[2] Infectomics cluster, Advanced Medical and Dental Institute, University of Science, Malaysia, Bertam, 13200, Kepala Batas, Pulau Pinang, Malaysia.

Abstract

The IntFOLD server provides a unified resource for the automated prediction of: protein tertiary structures with built-in estimates of model accuracy (EMA), protein structural domain boundaries, natively unstructured or disordered regions in proteins, and protein-ligand interactions. The component methods have been independently evaluated via the successive blind CASP experiments and the continual CAMEO benchmarking project. The IntFOLD server has established its ranking as one of the best performing publicly available servers, based on independent official evaluation metrics. Here, we describe significant updates to the server back end, where we have focused on performance improvements in tertiary structure predictions, in terms of global 3D model quality and accuracy self-estimates (ASE), which we achieve using our newly improved ModFOLD7_rank algorithm. We also report on various upgrades to the front end including: a streamlined submission process, enhanced visualization of models, new confidence scores for ranking, and links for accessing all annotated model data. Furthermore, we now include an option for users to submit selected models for further refinement via convenient push buttons. The IntFOLD server is freely available at: http://www.reading.ac.uk/bioinf/IntFOLD/.

Genome Biol. 2019 Nov 19;20(1):244. doi: 10.1186/s13059-019-1835-8.

The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens.

Zhou N[1,2], Jiang Y[3], Bergquist TR[4], Lee AJ[5], Kacsoh BZ[6,7], Crocker AW[8], Lewis KA[8], Georghiou G[9], Nguyen HN[1,10], Hamid MN[1,2], Davis L[2], Dogan T[11,12], Atalay V[13], Rifaioglu AS[13,14], Dalkıran A[13], Cetin Atalay R[15], Zhang C[16], Hurto RL[17], Freddolino PL[16,17], Zhang Y[16,17], Bhat P[18], Supek F[19,20], Fernández JM[21,22], Gemovic B[23], Perovic VR[23], Davidović RS[23], Sumonja N[23], Veljkovic N[23], Asgari E[24,25], Mofrad MRK[26], Profiti G[27,28], Savojardo C[27], Martelli PL[27], Casadio R[27], Boecker F[29], Schoof H[30], Kahanda I[31], Thurlby N[32], McHardy AC[33,34], Renaux A[35,36,37], Saidi R[12], Gough J[38], Freitas AA[39], Antczak M[40], Fabris F[39], Wass MN[40], Hou J[41,42], Cheng J[42], Wang Z[43], Romero AE[44], Paccanaro A[44], Yang H[45,46], Goldberg T[47], Zhao C[48,49,50], Holm L[51], Törönen P[51], Medlar AJ[51], Zosa E[52], Borukhov I[53], Novikov I[54], Wilkins A[55], Lichtarge O[55], Chi PH[56], Tseng WC[57], Linial M[58], Rose PW[59], Dessimoz C[60,61,62], Vidulin V[63], Dzeroski S[64,65], Sillitoe I[66], Das S[67], Lees JG[67,68], Jones DT[69,70], Wan C[71,69], Cozzetto D[71,69], Fa R[71,69], Torres M[44], Warwick Vesztrocy A[70,72], Rodriguez JM[73], Tress ML[74], Frasca M[75], Notaro M[75], Grossi G[75], Petrini A[75], Re M[75], Valentini G[75], Mesiti M[75,76], Roche DB[77], Reeb J[77], Ritchie DW[78], Aridhi S[78], Alborzi SZ[78,79], Devignes MD[78,80,79], Koo DCE[81], Bonneau R[82,83], Gligorijević V[84], Barot M[85], Fang H[86], Toppo S[87], Lavezzo E[87], Falda M[88], Berselli M[87], Tosatto SCE[89,90], Carraro M[90], Piovesan D[90], Ur Rehman H[91], Mao Q[92,93], Zhang S[92], Vucetic S[92], Black GS[94,95], Jo D[94,95], Suh E[94], Dayton JB[94,95], Larsen DJ[94,95], Omdahl AR[94,95], McGuffin LJ[96]**, Brackenridge DA[96]**, Babbitt PC[97,98], Yunes JM[99,98], Fontana P[100], Zhang F[101,102], Zhu S[103,104,105], You R[103,104,105], Zhang Z[103,105], Dai S[103,105], Yao S[103,104], Tian W[106,107], Cao R[108], Chandler C[108], Amezola M[108], Johnson D[108], Chang

JM[109], Liao WH[109], Liu YW[109], Pascarelli S[110], Frank Y[111], Hoehndorf R[112], Kulmanov M[112], Boudellioua I[113,114], Politano G[115], Di Carlo S[115], Benso A[115], Hakala K[116,117], Ginter F[116,118], Mehryary F[116,117], Kaewphan S[116,117,119], Björne J[120,121], Moen H[118], Tolvanen MEE[122], Salakoski T[120,121], Kihara D[123,124], Jain A[125], Šmuc T[126], Altenhoff A[127,128], Ben-Hur A[129], Rost B[47,130], Brenner SE[131], Orengo CA[67], Jeffery CJ[132], Bosco G[133], Hogan DA[6,8], Martin MJ[9], O'Donovan C[9], Mooney SD[4], Greene CS[134,135], Radivojac P[136], Friedberg I[137].

Abstract
BACKGROUND:
The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function.
RESULTS:
Here, we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in Candida albicans and Pseudomonas aureginosa genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility. We further performed targeted assays on selected genes in Drosophila melanogaster, which we suspected of being involved in long-term memory.
CONCLUSION:
We conclude that while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than the expectations set by baseline methods in C. albicans and D. melanogaster, it leaves considerable room and need for improvement. Finally, we report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.
KEYWORDS:
Biofilm; Community challenge; Critical assessment; Long-term memory; Protein function prediction

5. Under review – Arvinas Book Chapter

Proteins and their interacting partners: An introduction to protein-ligand binding site prediction methods with a focus on FunFOLD3

Danielle Allison Brackenridge[1] and Liam James McGuffin[2]

[1]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK.
d.a.brackenridge@pgr.reading.ac.uk
[2]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK.
l.j.mcguffin@reading.ac.uk

Abstract

Proteins are essential molecules with a diverse range of functions; elucidating their biological and biochemical roles from their interacting partners can be difficult and time consuming using *in vitro* and/or *in vivo* methods. Additionally, *in vivo* protein-ligand binding site elucidation is unable to keep pace with current growth in sequencing, leaving the majority of new sequences without known functions. Therefore, the development of new methods, which aim to predict the protein-ligand interactions and ligand-binding site residues directly from amino acid sequences, is becoming increasingly important. *In silico* prediction can utilise either sequence information, structural information or a combination of both. In this chapter, we will discuss the broad range of methods for ligand-binding site prediction from protein structure and we will describe our method FunFOLD3, for the prediction of protein-ligand interactions and ligand-binding sites based on template-based modelling. Additionally we will describe the step-by-step instructions on using the FunFOLD3 downloadable application, along with examples from the Critical Assessment of Techniques for Protein Structure Prediction (CASP) where FunFOLD3 has been used to aid ligand and ligand-binding site prediction. Finally, we will introduce our newer method, FunFOLD3-D, a version of FunFOLD3 which will aim to improve template based protein-ligand binding site prediction through the integration of docking, using AutoDock Vina.

Key words protein-ligand interactions, ligand-binding site prediction, Critical Assessment of Techniques for Protein Structure Prediction (CASP), protein structure prediction, template-based modelling, *in silico* prediction, FunFOLD3, docking

\

International Journal of
Molecular Sciences

MDPI

Review

Proteins and Their Interacting Partners: An Introduction to Protein–Ligand Binding Site Prediction Methods

Daniel Barry Roche [1,2,*], Danielle Allison Brackenridge [3] and Liam James McGuffin [3]

Received: 20 September 2015; Accepted: 10 December 2015; Published: 15 December 2015
Academic Editors: Tatyana Karabencheva-Christova and Christo Z. Christov

[1] Institut de Biologie Computationnelle, LIRMM, CNRS, Université de Montpellier, Montpellier 34095, France
[2] Centre de Recherche de Biochimie Macromoléculaire, CNRS-UMR 5237, Montpellier 34293, France
[3] School of Biological Sciences, University of Reading, Reading RG6 6AS, UK; d.a.brackenridge@pgr.reading.ac.uk (D.A.B.); l.j.mcguffin@reading.ac.uk (L.J.M.)
* Correspondence: daniel.roche@lirmm.fr or daniel.roche@crbm.cnrs.fr; Tel.: +33-4-34-35-9570; Fax: +33-4-34-35-9410

Abstract: Elucidating the biological and biochemical roles of proteins, and subsequently determining their interacting partners, can be difficult and time consuming using *in vitro* and/or *in vivo* methods, and consequently the majority of newly sequenced proteins will have unknown structures and functions. However, *in silico* methods for predicting protein–ligand binding sites and protein biochemical functions offer an alternative practical solution. The characterisation of protein–ligand binding sites is essential for investigating new functional roles, which can impact the major biological research spheres of health, food, and energy security. In this review we discuss the role *in silico* methods play in 3D modelling of protein–ligand binding sites, along with their role in predicting biochemical functionality. In addition, we describe in detail some of the key alternative *in silico* prediction approaches that are available, as well as discussing the Critical Assessment of Techniques for Protein Structure Prediction (CASP) and the Continuous Automated Model EvaluatiOn (CAMEO) projects, and their impact on developments in the field. Furthermore, we discuss the importance of protein function prediction methods for tackling 21st century problems.

Keywords: protein–ligand binding site prediction; protein function prediction; binding-site residue prediction; biochemical functional elucidation; sequence-based function prediction; structure-based function prediction; biological and biochemical role of enzymes; gene Ontology; enzyme commission numbers

1. Introduction

Proteins are essential molecules involved in a wide variety of essential intra- and inter-cellular activities. These activities include, but are not limited to: maintaining cellular defences, enzymatic catalysis, metabolism and catabolism, maintenance of the structural integrity of cells, and signalling within and between cells. Furthermore, protein–ligand interactions are essential for biochemical functionality and are implicated in all biochemical roles, in all kingdoms of life. Hence, studying protein–ligand binding sites and their associated residues, is an important step in the functional elucidation of proteins involved in these cellular processes [1–4].

Understanding protein–ligand interactions in the context of protein–ligand binding sites and ligand binding site residues is important for fully understanding cellular mechanisms, and is critical for understanding responses to drugs. Methods for the prediction of protein–ligand binding sites, which are detailed in the following section, can greatly enhance our understanding of the molecular mechanisms involved in many research spheres, helping us tackle numerous 21st century problems. The effects of protein–ligand binding are transient, but this knowledge can be exploited for the treatment of human and animal diseases, in addition to impacting food security research, examples of which are highlighted in Figure 1 and discussed in Section 6.

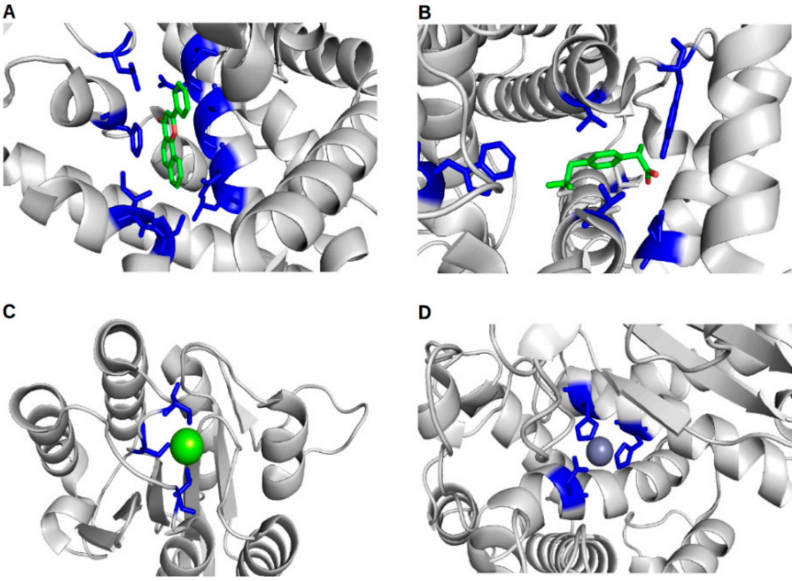


Figure 1. Examples of protein–ligand interactions, focusing on the ligand binding site. Proteins are shown in cartoon form and coloured in light grey, with binding site residues shown as blue sticks, and ligands shown as sticks or spheres coloured by element; (**A**) The Human cytochrome P450 1A1 protein (PDB ID 4i8v) bound to the drug *N*-Benzylformamide; (**B**) Cyclooxygenase-2 (PDB ID 4ph9) from *Mus musculus* bound to the drug Ibuprofen; (**C**) The *Plasmodium vivax* TRAP protein (PDB ID 4hqo, CASP ID T0686) bound to magnesium and; (**D**) The aminopeptidase N family protein Q5QTY1 (PDB ID 4fgm, CASP ID T0726) from *Idiomarina loihiensis* bound to zinc (a cofactor).

We begin by briefly highlighting some key protein–ligand interactions from a biomedical perspective. In Figure 1 we focus on four examples of proteins bound to diverse types of ligands, which are important in health and disease. This includes Cytochrome P450 bound to the drug *N*-Benzylformamide (Figure 1A—PDB ID 4i8v). The enzyme Cytochrome P450 has an essential role in the electron transfer chain, and is therefore ubiquitous in all kingdoms of life [5]. The human Cytochrome P450 (CYP1A1) is known to play a role in the biotransformation of polycyclic aromatic hydrocarbons into carcinogens [6]. In addition, CYP1A1 (PDB ID 4i8v) is responsible for the metabolism of theophylline [7], a drug used to provide symptomatic relief from asthma. Cyclooxygenase-2 from *Mus musculus*, which is involved in the biosynthesis of prostaglandins, is a target of non-steroidal anti-inflammatory drugs such as Ibuprofen (Figure 1B). The *Plasmodium vivax* TRAP protein, bound to magnesium, is involved in phosphate ester hydrolysis (Figure 1C). Finally, Figure 1D shows the protein–ligand binding site of the aminopeptidase N family protein Q5QTY1, from *Idiomarina loihiensis* bound to zinc (its cofactor), which can be used as a biomarker to detect kidney damage.

This review aims to provide an overview of the variety of different methodologies available for the prediction of protein–ligand binding sites and their associated binding site residues. Here we will focus on computational methods developed in the last six years, since the inclusion of the function prediction (FN) category in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition [8]. For methods developed before 2010, please refer to the review by Kaufmann and Karypis [9]. Furthermore, molecular docking methods are beyond the scope of this review, which have been recently reviewed by Yuriev *et al.* [10]. In this review, the term ligand is used to refer to molecules capable of binding to a protein, such as metal ions, small organic (e.g., ATP) and inorganic compounds (e.g., NH_4), peptides, and DNA/RNA; not large macromolecules such as proteins.

2. *In Silico* Methods for the Prediction of Protein–Ligand Binding Sites and Their Associated Binding Site Residues

In recent years, a large number of methods have been developed for the prediction of protein function and protein–ligand binding sites. In this review, we discuss methods for the prediction of protein–ligand binding sites and their associated binding site residues. These methods can be broadly divided into sequence-based methods and structure-based methods.

2.1. Sequence-Based Methods

Sequence-based methods that predict protein–ligand binding sites and their interacting ligand-binding site residues are those that use information from evolutionary conservation and/or sequence similarity of homologous proteins. These methods can be broadly categorised into methods that utilize machine learning (Multi-RELIEF [11], TargetS [12], LigandRF [13], and OMSL [14]), methods that utilize only position-specific scoring matrices or PSSMs (INTREPID [15], DISCERN [16], ConSurf [17], and ConFunc [18]) and graph-based methods such as Conditional Random Field (CRF) [19]. The advent of including machine learning-based strategies into sequence-based methods has resulted in improved method sensitivity. Machine learning is applied to PSSMs or multiple sequence alignment-based properties using various alternative strategies, examples of which will now be discussed.

Many of the sequence-based methods, such as Multi-RELIEF [11], deploy machine learning methods to directly interpret multiple sequence alignment profiles. Multi-RELIEF works by estimating the functional specificity of residues from a multiple sequence alignment using local conservation properties. This method uses a machine learning technique called RELIEF [20] for feature selection and weighting, using a binary classification to discriminate features from two classes. A residue's local specificity is determined by comparing the sequence with the closest homologue in each of the two classes (same class and opposite class), using global sequence identity to find the nearest neighbour sequence. If a residue has high local specificity to one pair of classes, it is labelled as relevant. Furthermore, global sequence similarity is considered while scoring each residue locally [11]. This results in the prediction of residues comprising a putative ligand binding site.

In contrast, LigandRFs [13] uses a random forest-based algorithm to predict protein–ligand binding site residues. LigandRFs extracts 544 amino acid properties from the AAindex database [21], which are then compared using the Matthews correlation coefficient. Each of the 544 properties are ranked in relation to the number of their related properties. The properties are filtered to remove all properties related to the top property; this removes redundant properties, which do not add any new information. This process is continued through the list until 34 properties remain. These properties relate to specific features crucial for determining putative binding site residues. The properties are then applied over a seven residue sliding window of a PSI-BLAST [22] profile. A 1 × 238 vector is used to represent the 34 amino acid properties for each seven residue window. A random forest is then utilized to learn the relationship between the large vector and the binding or non-binding residue properties [13].

TargetS [12] is another machine learning-based method, but in contrast to other methods, it utilizes secondary structure-based features in addition to sequence and PSSM-based features. Currently, TargetS can predict ligand-binding sites for proteins that bind to nucleotides, metal ions, DNA, and heme. The algorithm incorporates: protein conservation from a PSI-BLAST [22] PSSM searching SwissProt [23], secondary structure features determined from the PSIPRED algorithm [24], along with ligand-binding propensity of residues for each amino acid and each ligand category (nucleotides, metal ions, DNA, and heme). These properties are subsequently combined using a support vector machine (SVM) to predict ligand-binding site residues.

2.2. Structure-Based Methods

Structure-based methods are those that exploit information from 3D atomic coordinates (either predicted from sequence or derived from experiments). These methods either predict the location of the ligand binding site and/or the putative ligand binding site residues. Such methods can be further sub-categorised into: 1. Geometric-based methods (FINDSITE [25], LigDig [26], LISE [27], PatchSurfer2.0 [28], Surflex-PSIM [29], EvolutionaryTrace [30], PRANK [31], a Two-dimensional replica-exchange method [32], FMO-RESP [33], MapReduce approach [34], TIFP [35], ProGolem [36], a Chemogenomics approach [37], ProPose [38], FunFHMMer [39], mFASD [40], ProBis [41,42], and CavBase [43,44]); 2. Energetic methods (SITEHOUND [45], VISCANA [46], SiteComp [47], and FTMap [48]); 3. Miscellaneous methods, which use information from homology or template-based modelling (FunFOLD3 [3,4], COACH [49], COFACTOR [50], GalaxySite [51], GASS [52], VISM-CFA [53], and PLIP [54]), Surface accessibility based methods such as LigSite^{CSC} [55], in addition to Physicochemical properties exploited by Andersson and colleagues [56]. Examples of different methods from each sub-category are now described, in addition to their limitations.

2.2.1. Considerations When Employing Structure-Based Methods

Structure-based methods for prediction of protein–ligand binding sites have a number of limitations, including the following: 1. If a 3D model or experimental structure cannot be obtained, then it is not possible to make a prediction; in such cases the solution is to rely on purely sequence-based methods. 2. If templates with the same fold as the target protein that contain biologically relevant ligands cannot be detected, then it is not possible to make a prediction. 3. Most prediction servers, such as COACH [49] and FunFOLD [3,4,57], utilize in-house structure prediction pipelines to construct models for protein–ligand interaction predictions that may not always produce the best quality model for every target, which may result in over- and under-predicted protein–ligand binding sites. Nevertheless, despite these shortcomings, prediction methods are constantly under development and improvements can be gauged via the rigorous independent blind assessment scoring, described in Section 3.

2.2.2. Geometric Methods

FINDSITE [25] combines evolutionary and structural information to predict protein function, identifying binding pockets based on binding site similarity between homologous structures. This is undertaken by superposing templates onto the structure of interest and then finding sites where ligands overlap. These results are then used to determine putative binding pockets and then identifying the geometric centre of each pocket [25].

Similarly, LigDig [26] is another geometric method, but uses a ligand-centric approach, rather than the traditional protein-centric approach to detect ligand-binding pockets in proteins. LigDig utilizes a variety of information from ChEBI [58], PubChem, PDB [59], UniProt [23], and KEGG [60], combined via a graph-based network to locate similar ligands along with their potential binding partners. The method is available as a webserver and also uses text-based searches to find proteins that may bind to a particular ligand of interest [26]. This results in the prediction of putative protein–ligand binding sites.

In contrast to FINDSITE, LigDig, and the majority of geometric-based approaches, LISE [27] is an algorithm that utilizes a novel concept of binding site-enriched protein triangles in order to predict protein–ligand binding site locations. LISE uses ideas developed in a previous method, called MotifScore [61], that determined motifs in a protein–ligand interaction database, composed of 6276 protein–ligand structures. The motifs contain the interactions between three atoms of a protein and two atoms of a ligand. Thus, the three protein atoms of these motifs compose the “protein triangles”. An additional step is to encapsulate the protein into a 3D grid of 1 Å size steps. Each vertex in this grid is then labelled as occupied or empty (with a 2.7 Å distance cutoff). For each empty grid point, a grid point score is calculated, which equals the sum of the triangle scores. A large sphere of 11Å is then centred on each empty vertex, and for each sphere, a sphere score is calculated, which is based on the sum of the grid point scores for all empty grid points within the sphere. The sphere with the highest score is determined as the putative ligand binding site [27].

2.2.3. Energetic Methods

SITEHOUND [45] is a widely used energetic method for the prediction of protein–ligand binding sites, which utilizes a chemical probe to explore the surface of the protein structure, determining regions that may have optimum energy for binding. SITEHOUND uses two different chemical probes: a carbon probe to identify drug-like binding sites, and a phosphate probe to locate binding sites for ligands having a phosphate group. Affinity maps or molecular interaction fields are then used to describe the interaction of each probe with the protein surface. These affinity maps are subsequently filtered to remove unfavourable interaction energies. The next step is to utilize agglomerative hierarchical clustering to cluster the remaining interaction points based on their spatial proximity. These clustered points are ranked by total interaction energy and result in a list of potential ligand-binding pocket locations [45].

2.2.4. Miscellaneous Methods

A recent review by Petrey *et al*. [62] highlights the essential need for template-based 3D modelling methods in the prediction of protein function [62]. The majority of these methods predict putative protein–ligand binding sites and ligand binding site residues, while some methods additionally predict Enzyme Commission Numbers (EC) and Gene Ontology (GO) terms. We have developed a number of versions of a template-based method, called FunFOLD [3,4,57], which starts with a 3D model of the target protein predicted from sequence, for example using the IntFOLD server [63,64]. Each version of the algorithm has worked on the assumption that proteins with the same fold that bind to similar biologically relevant ligands are likely to have similar binding sites. The latest FunFOLD3 pipeline is composed of updated versions of two main algorithms, FunFOLD [4] and FunFOLDQA [1], and it produces output comprising predicted EC and GO terms, ligand-binding site residues, putative ligands, binding site quality scores, and per-atom *p*-values to comply with the CAMEO-LB format [65].

FunFOLD firstly superposes, using TM-align [66], a list of structural templates containing biologically relevant ligands (determined using the BioLip database [67]) onto the target 3D model. Template-model superpositions with a TM-score $\geqslant 0.4$ are retained. The next step is to superimpose all retained templates onto the target model and assign ligands from the template files into clusters using agglomerative hierarchical clustering. The identified ligand clusters are located at the potential ligand-binding sites. Ligands are determined to be components of a cluster if the contact distance is less than or equal to 0.5 Å plus the Van der Waal radii of the contacting atoms. The putative ligand-binding site containing the largest ligand cluster is determined to be the most probable ligand-binding site of the protein. The identification of the putative ligand-binding site residues is carried out via a residue voting method [3,4].

The next component of the FunFOLD3 pipeline is the FunFOLDQA algorithm [1], which evaluates the quality of FunFOLD predictions, subsequently producing a set of confidence scores. The algorithm outputs scores for five sequence- and structure-based features that are combined using a neural network, outputting predicted Binding-site Distance Test (BDT) [68] and Matthews Correlation Coefficient (MCC) [69] scores. The FunFOLD3 [57] pipeline additionally outputs a set of per-residue binding probability scores to comply with the CAMEO-LB format [65]. Furthermore, the FunFOLD3 method outputs a putative ligand binding site, putative ligand binding site residues, putative ligands that may bind to the target protein, along with predicted EC and GO [70,71] terms (see Section 4) for each target protein [3,4].

The COACH [49] method is similar to FunFOLD and is one of the most accurate ligand-binding site prediction methods that utilizes both sequence and structural homology in the prediction pipeline. The structure component (TM-SITE) of the pipeline firstly locates putative ligand-binding pockets using ConCavity [72]. TM-SITE then uses fifteen residues within the binding pocket structure to search against the BioLip database to find structures containing similar binding pockets, in addition to searching for similar structures (using TM-align [66]) to the target protein containing biologically relevant ligands within BioLip [67]. All templates and sub-structural templates are superposed onto the target and scored based on empirically determined cutoffs. Ligand binding site residues are then determined using a similar strategy to FunFOLD [4], but using average linkage clustering and assigning a confidence score to each predicted ligand binding site residue. The sequence component of the algorithm, S-SITE, uses residue conservation of sequence profiles to predict ligand binding site residues, subsequently scoring the confidence of each predicted binding site residue. COACH then uses a consensus of predictions, combining the results from TM-SITE and S-SITE along with COFACTOR [50], FINDSITE [25], and ConCavity [72]. Similar to FunFOLD3, COACH predicts a putative ligand binding site, putative ligand binding site residues, putative ligands that may bind to the target protein, along with predicted EC and GO terms for each target protein.

A somewhat alternative approach to that of FunFOLD and COACH is used by GASS [52]. GASS (Genetic Active Site Search) is developed by Izidoro *et al.* [52], who have employed a genetic algorithm to predict ligand binding site residues for putative enzymes. Their method takes a list of templates from the CSA [73] with predefined binding site residues. They then simulate evolutionary effects (crossover and mutations) over this population of templates, according to predefined mutational probabilities, for a specific number of user-defined generations. The resultant binding site residue predictions are then assessed using a fitness function, which ranks individual sets of predictions. The fitness function is similar to an RMSD (root-mean-square deviation) for the ligand binding site residues, with the main difference being that the square distance of the results is not averaged [74].

Several structure-based methods that exploit surface accessibility have also been developed, such as LIGSITE^{csc} [55]. LIGSITE^{csc} uses the Connolly surface in its ligand binding site prediction protocol. The first step of the protocol is to encapsulate the protein structure into a 3D grid of 1 Å steps. In the second step of the protocol, each point in the grid is labelled as either protein, surface, or solvent. In the third step, the Connolly algorithm is utilized to calculate the solvent-excluded surface. In the fourth step, surface-solvent-surface events are then determined. In the fifth step, if the surface-solvent-surface events in a grid exceed a minimum threshold, set to six grid locations, this is determined to be a pocket. Each pocket cluster is then ranked in relation to the number of grid points within the cluster. The top three pockets are then retained. In the final step, the top three pockets are re-ranked in relation to the conservation of pocket surface residues [55].

Further structure-based methods have used physiochemical properties to determine ligand binding cavities. For example, the method by Andersson *et al.* (2010, [56]) works initially by identifying solvent accessible patches. In the second step, data is collected from each patch based on 408 surface descriptors, divided into eight categories. These descriptors include neighbouring amino acids, secondary structure, polarity of adjacent amino acids, close hydrogen bond donors and acceptors, electrostatic potential, shape, polarity, and flexibility. In the third step, the descriptor results are divided into bins and scaled to be usable for Principal Component Analysis (PCA). In the fourth step, PCA is carried out and the relationships between pockets are analysed. This method produces results for all putative pockets, leaving the user to determine which pocket is the most suitable ligand binding pocket for their particular task [56].

3. Methods for the Evaluation of Protein–Ligand Binding Site Residue Predictions

Assessment of protein–ligand binding site residue predictions have been carried out in CASP [8,75,76] and CAMEO [65] using a number of different scores, which include the Matthews Correlation Coefficient (MCC) [69] and the Binding-site Distance Test (BDT) score [68]. The MCC score is a statistical measure that compares observed ligand binding site residues to predictions by assessing the number of residues assigned as true positives, false positives, true negatives, and false negatives. This results in a score between −1 and 1, with scores close to zero representing random predictions and scores close to one representing near perfect predictions. The MCC score may only be a good choice for scoring sequence-based predictions, when no structural information is available, as the MCC score does not consider the 3D nature of the protein within the scoring metric.

To overcome the limitations of the MCC score, we developed the Binding-site Distance Test (BDT score) [68]. The BDT score utilizes the distance in 3D space between a predicted ligand binding site residue and an observed ligand binding site residue in the scoring process. The BDT score has a range from zero to one, where scores close to zero represent random predictions and scores close to one represent near perfect predictions. Predicted ligand binding sites closer in 3D space to the observed ligand binding site are scored higher than ligand binding sites predicted farther from the observed ligand binding site. In the CASP9 and CASP10 FN assessments [75,76], the BDT score was used by the official assessors in addition to the MCC score. Furthermore, the BDT score is used in the CAMEO [65] project as one of the standard assessment metrics.

4. Prediction of Enzyme Commission Numbers (EC) and Gene Ontology Terms (GO)

In addition to the determination of protein–ligand binding sites and their associated binding site residues, it is also useful to determine the likely function of a protein. Functionality can be generally assigned using Gene Ontology (GO) terms [70,71], or more specifically for enzymes, using Enzyme Commission numbers (EC).

The Gene Ontology Commission was formed in 2000 [70] to develop a controlled vocabulary for describing genes, as a result of the large increase of sequence data from genomics projects. Gene Ontology (GO) terms, often referred to as a shared vocabulary for genes, comprise over 40,000 terms. GO terms are broadly divided into three categories: cellular components, molecular function (a weak analogy to EC codes), and biological processes, which are further subdivided in a hierarchical graph-like structure. Each protein has the potential to be assigned to multiple GO classes and sub-classes. Moreover, each GO term has a unique serial number, in addition to a textual description [70,71].

The Enzyme Commission (EC) was set up in 1956 as part of the International Union of Pure and Applied Chemistry (IUPAC), publishing the first version of EC numbers in 1961. Today, the EC classification is maintained by the Nomenclature Committee of the International Union for Biochemistry and Molecular Biology (NC-IUBMB) and the enzyme list is curated and maintained by the Tipton group at Trinity College Dublin [77]. The list officially classifies enzymes by the overall reactions they catalyse, in order to reduce the ambiguous names enzymes previously acquired. Enzymes are hierarchically classified by four-digit EC numbers. The first number designates the broad classification into: 1. Oxidoreductases; 2. Transferases; 3. Hydrolases; 4. Lyases; 5. Isomerases; and 6. Ligases. The second class usually designates the type of molecule involved in the reaction. The third class designates the type of reaction involved, while the fourth class is essentially a serial number, which has been utilized to differentiate enzymes within the subclasses [77].

Recently, a number of methods have been developed specifically to predict GO and EC terms. A large number of these methods have been developed as rapid methods that utilize sequence information only. The majority of methods predict function based on Gene Ontology (GO) terms (which include: INGA [78], EFI-EST [79], SIFTER [80], GEO2Enrichr [81], PANNZER [82], and PILL [83]) with fewer utilizing EC numbers (EFI-EST [79] and DomSign [84]) for functional annotation. Furthermore, a number of structure-based methods for the prediction of protein–ligand binding sites have incorporated methods for predicting GO and EC terms, including COACH [49] and FunFOLD3 [3,4,57] (See Section 2.2.4). However, as these methods build 3D models as part of their prediction pipeline, they are somewhat more computationally intensive than the sequence-only methods.

The prediction of EC and GO terms, in addition to the prediction of protein–ligand binding sites and their associated ligand binding site residues, further enriches the information that can be gleaned for a particular protein. This highlights the biological need for *in silico* methods in function prediction and rational drug design, contributing to future *in silico*, *in vitro*, and *in vivo* experiments for both biomedical and bioenvironmental research applications.

5. CASP, CAFA, and CAMEO—Their Role in Development and Assessment of Protein–Ligand Binding Site Prediction Algorithms

The development of methods for the prediction of protein–ligand binding sites and function prediction has been driven in recent years as a direct result of community wide prediction experiments, such as the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [8,75,76], the Continuous Automated Model EvaluatiOn (CAMEO) project [65], and the Critical Assessment of Function Annotation (CAFA) [85].

Ligand binding site residue prediction was first introduced in CASP8 (as the FN category) [8], with the concept then involving the prediction of putative ligand binding site residues, which may functionally interact with a biologically relevant bound ligand. Since it is not presently possible to clearly distinguish between catalytic, active, and binding site residues, using computational methods, the algorithms simply predict protein–ligand binding site residues. In CASP8, the top performing methods LEE [86] and 3DLigandSite [87] used a similar prediction strategy, combining information from homology models along with the templates used to construct the models that contained biologically-relevant bound ligands. In CASP9 [75] and CASP10 [76], successful methods for the prediction of protein–ligand binding sites built upon and further refined this template-based approach.

Following on from CASP10 [76], it was decided to move the FN prediction category to a continuous assessment strategy, due to the lack of available targets containing bound biologically-relevant ligands during the short three month CASP prediction period. Hence, the CASP FN category moved to the CAMEO continuous assessment project [65]. The move to fully automated assessment resulted in a change of prediction format, with the additional prediction of which ligand category (I—Ion, O—Organic, N—Nucleotide, and P—Peptide) a protein may bind. Participating servers must also provide a p-value representing the likelihood that each residue (or atom) binds a ligand in each category. The CAMEO assessment runs weekly on structures containing biologically-relevant ligands using target sequences of structures that are on hold for release by the Protein Data Bank (PDB) [65]. The CAMEO project provides a better picture of how each method performs on a large and diverse dataset, containing a wide variety of proteins bound to a wide variety of ligands.

Complementary to CAMEO and CASP is the CAFA [85] experiment, which has also been a major driver for the development of function prediction methods. The goal of CAFA is to functionally annotate proteins on a large scale using GO terms [70,71]. The CAFA1 dataset contained >48,000 proteins as of October 2010, for which predictions were made. Following the prediction season, methods were evaluated on 866 of the proteins, which had acquired annotations over the eleven months following the close of the prediction season. Methods that compete in CAFA [85] include a large number of the methods described in the proceeding sections, comprising sequence-based methods, structure-based methods and combinations of both.

6. The Application of *in Silico* Protein–Ligand Binding Site Prediction Methods: Impact on *in Vitro* Studies

In addition to the theoretical and computational uses of protein–ligand binding site prediction algorithms previously highlighted, methods for the prediction of protein–ligand binding sites have been used in numerous *in silico*/*in vitro* studies. These studies have focused on a wide range of subjects as diverse as calcium-binding proteins [88], olfactory proteins [89], the CollagenQ protein–COLQ [90], human PE5 proteins [91], barley powdery mildew proteins [92,93], and spider mite glutathione S-transferases [94], which have led to biological findings of relevance to the study of health and disease and to food security [88–95].

We firstly describe a number of case studies from research projects investigating proteins implicated in health and disease. The first study [88] analysed a large number of calcium-binding proteins present in biological systems on a genome-wide scale, termed: calciomics. As calcium impacts every aspect of cellular life, Ca^{2+} binding proteins can be implicated in a wide range of diseases, thus this *in silico* study investigates their potential roles [88]. Another *in silico* proteome-wide study, this time on PE5 proteins (plasma membrane transporters and receptors) from the human proteome, was undertaken by Dong *et al.* to correct misannotations of these highly misannotated proteins [91]. Furthermore, Don and Riniker undertook *in silico* analysis of olfactory receptor proteins, members of the G-protein coupled receptor (GPCR) family, to enable the future design of therapeutics targeting olfactory-related and GPCR-related diseases [89]. In addition, Arredondo *et al.* combined modelling and the prediction of protein–ligand binding sites with *in vitro* studies to investigate a number of COLQ mutants and determine their mode of action [90]. These COLQ mutants cause human deficiency of endplate acetylcholinesterase, which results in the impairment of the interaction of COLQ with the basal lamina. This leads to a reduction in the duration of synaptic activation, which can lead to synaptic-related diseases.

Focusing on projects that have implications on food security, we highlight a study on the barley powdery mildew proteome [92]. This research involved the combination of proteogenomic along with structural and functional (protein–ligand binding sites and binding site residues) predictions, in order to investigate the pathogenic properties of barley powdery mildew. Basically, IntFOLD [63,64] was used to construct models for the entire proteome, which were validated utilizing ModFOLD3 [63]. Subsequently, FunFOLD [4] was used to predict protein–ligand binding sites for these models. This resulted in interesting conclusions about the *Blumera graminis* f.sp. *hordei* proteome. Firstly, the proteins are structurally diverse and remotely homologous to known proteins, potentially containing novel folds, as it was only possible to model six proteins with a model quality score above 0.4. Secondly, FunFOLD was able to help in the assignment of functionality for these six proteins, all were carbohydrate-binding and probably glycosyl hydrolases. Moreover, this putative functionality was experimentally elucidated, highlighting the utility of protein–ligand interaction methods to aid functional elucidation [92]. An additional study with relevance to food security from Pavlidi *et al.* [94] involves the functional characterization of a particular glutathione S-transferase, which may enable the two-spotted spider mite (*Tetranychus urticae*) to have acaricide/insecticide resistance. *Tetranychus urticae* has been shown to be one of the most damaging agricultural pests globally. The spider mite has three glutathione S-transferase enzymes; TuGSTd10, TuGSTd14, and TuGSTm09. Subsequently, assays determined that TuGSTd14 was the glutathione S-transferase involved in the acaricide/insecticide resistance. The structure of TuGSTd14 was predicted using IntFOLD [63,64] and protein–ligand binding site residues predicted using FunFOLD [3]. These *in silico* results were utilized to determine the key structural characteristics, including residues that were involved in the substrate binding specificity [94].

The studies described above, on proteins related to health and disease [88–91] in addition to food security [92–94] highlight the utility of protein–ligand binding site prediction methods, which can contribute to the interpretation of the function and the biochemical interactions of key proteins and enzymes, impacting our ability to tackle urgent global problems.

7. Conclusions

A large number of predictive methods are available to predict and analyse protein–ligand binding sites. These methods incorporate different approaches, providing numerous different data types ranging from lists of ligand binding site residues, 3D atomic coordinates of ligand binding sites, lists of putative binding ligands, EC, and GO terms. The results produced by these *in silico* methods can be useful to generate new hypotheses and drive further experiments that can impact on major challenges in biology.

Acknowledgments: Daniel Barry Roche is a recipient of a Young Investigator Fellowship from the Institut de Biologie Computationnelle, Université de Montpellier (ANR Investissements D'Avenir Bio-informatique: projet IBC).

Author Contributions: Daniel Barry Roche co-conceived the idea, drafted the manuscript, contributed figures and carried out final editing of the manuscript; Danielle Allison Brackenridge contributed text and figures; Liam James McGuffin co-conceived the idea and carried out final editing of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roche, D.B.; Buenavista, M.T.; McGuffin, L.J. FunFOLDQA: A quality assessment tool for protein–ligand binding site residue predictions. *PLoS ONE* **2012**, *7*, e38219. [CrossRef] [PubMed]
2. Roche, D.B.; Buenavista, M.T.; McGuffin, L.J. Predicting protein structures and structural annotation of proteomes. In *Encyclopedia of Biophysics*; Roberts, G.C.K., Ed.; Springer: Berlin, Germany, 2012; Volume 1.
3. Roche, D.B.; Buenavista, M.T.; McGuffin, L.J. The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Res.* **2013**, *41*, 303–307. [CrossRef] [PubMed]
4. Roche, D.B.; Tetchner, S.J.; McGuffin, L.J. FunFOLD: An improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinforma.* **2011**, *12*, 160. [CrossRef] [PubMed]
5. Rang, H.P.; Ritter, J.M.; Flower, R.J.; Henderson, G. *Rang and Dale's Pharmacology*, 8th ed.; Elsevier Churchill Livingstone: London, UK, 2015.
6. Walsh, A.A.; Szklarz, G.D.; Scott, E.E. Human cytochrome P450 1A1 structure and utility in understanding drug and xenobiotic metabolism. *J. Biol. Chem.* **2013**, *288*, 12932–12943. [CrossRef] [PubMed]
7. Yang, K.H.; Lee, J.H.; Lee, M.G. Effects of CYP inducers and inhibitors on the pharmacokinetics of intravenous theophylline in rats: Involvement of CYP1A1/2 in the formation of 1,3-DMU. *J. Pharm. Pharmacol.* **2008**, *60*, 45–53. [CrossRef] [PubMed]
8. Lopez, G.; Ezkurdia, I.; Tress, M.L. Assessment of ligand binding residue predictions in CASP8. *Proteins* **2009**, *77*, 138–146. [CrossRef] [PubMed]
9. Kauffman, C.; Karypis, G. Ligand-binding residue prediction. In *Introduction to Protein Structure Prediction: Methods and Algorithms*; Rangwala, H., Karypis, G., Eds.; Wiley: Hoboken, NJ, USA, 2010.
10. Yuriev, E.; Holien, J.; Ramsland, P.A. Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. *J. Mol. Recognit.* **2015**, *28*, 581–604. [CrossRef] [PubMed]
11. Ye, K.; Feenstra, K.A.; Heringa, J.; Ijzerman, A.P.; Marchiori, E. Multi-relief: A method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. *Bioinformatics* **2008**, *24*, 18–25. [CrossRef] [PubMed]
12. Yu, D.-J.; Hu, J.; Yang, J.; Shen, H.-B.; Tang, J.; Yang, J.-Y. Designing template-free predictor for targeting protein–ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 994–1008. [CrossRef] [PubMed]
13. Chen, P.; Huang, J.H.Z.; Gao, X. Ligandrfs: Random forest ensemble to identify ligand-binding residues from sequence information alone. *BMC Bioinforma.* **2014**. [CrossRef] [PubMed]
14. Yu, D.J.; Hu, J.; Li, Q.M.; Tang, Z.M.; Yang, J.Y.; Shen, H.B. Constructing query-driven dynamic machine learning model with application to protein–ligand binding sites prediction. *IEEE Trans. Nanobiosci.* **2015**, *14*, 45–58.
15. Sankararaman, S.; Kolaczkowski, B.; Sjolander, K. Intrepid: A web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Res.* **2009**, *37*, 390–395. [CrossRef] [PubMed]
16. Sankararaman, S.; Sha, F.; Kirsch, J.F.; Jordan, M.I.; Sjolander, K. Active site prediction using evolutionary and structural information. *Bioinformatics* **2010**, *26*, 617–624. [CrossRef] [PubMed]
17. Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N. Consurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **2010**, *38*, 529–533. [CrossRef] [PubMed]

18. Wass, M.N.; Sternberg, M.J. Confunc—functional annotation in the twilight zone. *Bioinformatics* **2008**, *24*, 798–806. [CrossRef] [PubMed]
19. Wierschin, T.; Wang, K.; Welter, M.; Waack, S.; Stanke, M. Combining features in a graphical model to predict protein binding sites. *Proteins* **2015**, *83*, 844–852. [CrossRef] [PubMed]
20. Kononenko, I. Estimating attributes: Analysis and extensions of relief. In Proceedings of the European Conference on Machine Learning; Springer-Verlag New York, Inc.: Catania, Italy, 1994; pp. 171–182.
21. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, 202–205. [CrossRef] [PubMed]
22. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
23. UniProt, C. UniProt: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, 204–212.
24. McGuffin, L.J.; Bryson, K.; Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404–405. [CrossRef] [PubMed]
25. Brylinski, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 129–134. [CrossRef] [PubMed]
26. Fuller, J.C.; Martinez, M.; Henrich, S.; Stank, A.; Richter, S.; Wade, R.C. Ligdig: A web server for querying ligand-protein interactions. *Bioinformatics* **2015**, *13*, 1147–1149. [CrossRef] [PubMed]
27. Xie, Z.R.; Liu, C.K.; Hsiao, F.C.; Yao, A.; Hwang, M.J. LISE: A server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. *Nucleic Acids Res.* **2013**, *41*, 292–296. [CrossRef] [PubMed]
28. Zhu, X.; Xiong, Y.; Kihara, D. Large-scale binding ligand prediction by improved patch-based method patch-surfer2.0. *Bioinformatics* **2015**, *31*, 707–713. [CrossRef] [PubMed]
29. Spitzer, R.; Cleves, A.E.; Jain, A.N. Surface-based protein binding pocket similarity. *Proteins* **2011**, *79*, 2746–2763. [CrossRef] [PubMed]
30. Erdin, S.; Ward, R.M.; Venner, E.; Lichtarge, O. Evolutionary trace annotation of protein function in the structural proteome. *J. Mol. Biol.* **2010**, *396*, 1451–1473. [CrossRef] [PubMed]
31. Krivak, R.; Hoksza, D. Improving protein–ligand binding site prediction accuracy by classification of inner pocket points using local features. *J. Cheminform.* **2015**, *7*, 12. [PubMed]
32. Kokubo, H.; Tanaka, T.; Okamoto, Y. Two-dimensional replica-exchange method for predicting protein–ligand binding structures. *J. Comput. Chem.* **2013**, *34*, 2601–2614. [CrossRef] [PubMed]
33. Chang, L.; Ishikawa, T.; Kuwata, K.; Takada, S. Protein-specific force field derived from the fragment molecular orbital method can improve protein–ligand binding interactions. *J. Comput. Chem.* **2013**, *34*, 1251–1257. [CrossRef] [PubMed]
34. Estrada, T.; Zhang, B.; Cicotti, P.; Armen, R.S.; Taufer, M. A scalable and accurate method for classifying protein–ligand binding geometries using a mapreduce approach. *Comput. Biol. Med.* **2012**, *42*, 758–771. [CrossRef] [PubMed]
35. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding protein–ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623–637. [CrossRef] [PubMed]
36. Santos, J.C.A.; Nassif, H.; Page, D.; Muggleton, S.H.; Sternberg, M.J.E. Automated identification of protein–ligand interaction features using inductive logic programming: A hexose binding case study. *BMC Bioinform.* **2012**, *13*, 162. [CrossRef] [PubMed]
37. Jacob, L.; Vert, J.P. Protein-ligand interaction prediction: An improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156. [CrossRef] [PubMed]
38. Seifert, M.H.; Schmitt, F.; Herz, T.; Kramer, B. Propose: A docking engine based on a fully configurable protein–ligand interaction model. *J. Mol. Model.* **2004**, *10*, 342–357. [CrossRef] [PubMed]
39. Das, S.; Sillitoe, I.; Lee, D.; Lees, J.G.; Dawson, N.L.; Ward, J.; Orengo, C.A. Cath funfhmmer web server: Protein functional annotations using functional family assignments. *Nucleic Acids Res.* **2015**, *43*, 148–153. [CrossRef] [PubMed]
40. He, W.; Liang, Z.; Teng, M.; Niu, L. Mfasd: A structure-based algorithm for discriminating different types of metal-binding sites. *Bioinformatics* **2015**, *31*, 1938–1944. [CrossRef] [PubMed]
41. Konc, J.; Janezic, D. Probis-2012: Web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.* **2012**, *40*, 214–221. [CrossRef] [PubMed]

42. Konc, J.; Janezic, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168. [CrossRef] [PubMed]
43. Krotzky, T.; Fober, T.; Hullermeier, E.; Klebe, G. Extended graph-based models for enhanced similarity search in cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 878–890. [CrossRef] [PubMed]
44. Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406. [CrossRef]
45. Hernandez, M.; Ghersi, D.; Sanchez, R. Sitehound-web: A server for ligand binding site identification in protein structures. *Nucleic Acids Res.* **2009**, *37*, 413–416. [CrossRef] [PubMed]
46. Amari, S.; Aizawa, M.; Zhang, J.; Fukuzawa, K.; Mochizuki, Y.; Iwasawa, Y.; Nakata, K.; Chuman, H.; Nakano, T. Viscana: Visualized cluster analysis of protein–ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. *J. Chem. Inf. Model.* **2006**, *46*, 221–230. [CrossRef] [PubMed]
47. Lin, Y.; Yoo, S.; Sanchez, R. Sitecomp: A server for ligand binding site analysis in protein structures. *Bioinformatics* **2012**, *28*, 1172–1173. [CrossRef] [PubMed]
48. Kozakov, D.; Grove, L.E.; Hall, D.R.; Bohnuud, T.; Mottarella, S.E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **2015**, *10*, 733–755. [CrossRef] [PubMed]
49. Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588–2595. [CrossRef] [PubMed]
50. Roy, A.; Yang, J.; Zhang, Y. Cofactor: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **2012**, *40*, 471–477. [CrossRef] [PubMed]
51. Heo, L.; Shin, W.H.; Lee, M.S.; Seok, C. GalaxySite: Ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res.* **2014**, *42*, 210–214. [CrossRef] [PubMed]
52. Izidoro, S.C.; de Melo-Minardi, R.C.; Pappa, G.L. GASS: Identifying enzyme active sites with genetic algorithms. *Bioinformatics* **2014**. [CrossRef] [PubMed]
53. Guo, Z.; Li, B.; Cheng, L.T.; Zhou, S.; McCammon, J.A.; Che, J. Identification of protein–ligand binding sites by the level-set variational implicit-solvent approach. *J. Chem. Theory Comput.* **2015**, *11*, 753–765. [CrossRef] [PubMed]
54. Salentin, S.; Schreiber, S.; Haupt, V.J.; Adasme, M.F.; Schroeder, M. Plip: Fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **2015**, *43*, 443–447. [CrossRef] [PubMed]
55. Huang, B.; Schroeder, M. Ligsitecsc: Predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19. [CrossRef] [PubMed]
56. Andersson, C.D.; Chen, B.Y.; Linusson, A. Mapping of ligand-binding cavities in proteins. *Proteins* **2010**, *78*, 1408–1422. [CrossRef] [PubMed]
57. Roche, D.B.; McGuffin, L.J. *In silico* identification and characterization of protein–ligand binding sites, methods in molecular biology. In *Structure based and Computational Design of Ligand Binding Proteins*; Stoddard, B., Baker, D., Eds.; Humana Press: New York, NY, USA, 2015.
58. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; *et al.* The ChEBi reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res.* **2013**, *41*, 456–463. [CrossRef] [PubMed]
59. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
60. Okuda, S.; Yamada, T.; Hamajima, M.; Itoh, M.; Katayama, T.; Bork, P.; Goto, S.; Kanehisa, M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, 423–426. [CrossRef] [PubMed]
61. Xie, Z.R.; Hwang, M.J. An interaction-motif-based scoring function for protein–ligand docking. *BMC Bioinform.* **2010**, *11*, 298. [CrossRef] [PubMed]
62. Petrey, D.; Chen, T.S.; Deng, L.; Garzon, J.I.; Hwang, H.; Lasso, G.; Lee, H.; Silkov, A.; Honig, B. Template-based prediction of protein function. *Curr. Opin. Struct. Biol.* **2015**, *32*, 33–38. [CrossRef] [PubMed]

63. Roche, D.B.; Buenavista, M.T.; Tetchner, S.J.; McGuffin, L.J. The intfold server: An integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res.* **2011**, *39*, 171–176. [CrossRef] [PubMed]
64. McGuffin, L.J.; Atkins, J.D.; Salehe, B.R.; Shuid, A.N.; Roche, D.B. IntFOLD: An integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.* **2015**. [CrossRef] [PubMed]
65. Haas, J.; Roth, S.; Arnold, K.; Kiefer, F.; Schmidt, T.; Bordoli, L.; Schwede, T. The protein model portal—A comprehensive resource for protein structure and model information. *Database* **2013**. [CrossRef] [PubMed]
66. Zhang, Y.; Skolnick, J. Tm-align: A protein structure alignment algorithm based on the TM-score. , **2005**, *33*, 2302–2309. [CrossRef] [PubMed]
67. Yang, J.; Roy, A.; Zhang, Y. Biolip: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41*, 1096–1103. [CrossRef] [PubMed]
68. Roche, D.B.; Tetchner, S.J.; McGuffin, L.J. The binding-site distance test score: A robust method for the assessment of predicted protein binding sites. *Bioinformatics* **2010**, *26*, 2920–2921. [CrossRef] [PubMed]
69. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **1975**, *405*, 442–451. [CrossRef]
70. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
71. Gene Ontology Consortium. Gene ontology consortium: Going forward. *Nucleic Acids Res.* **2015**, *43*, 1049–1056.
72. Capra, J.A.; Laskowski, R.A.; Thornton, J.M.; Singh, M.; Funkhouser, T.A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585. [CrossRef] [PubMed]
73. Furnham, N.; Holliday, G.L.; de Beer, T.A.; Jacobsen, J.O.; Pearson, W.R.; Thornton, J.M. The catalytic site atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **2014**, *42*, 485–489. [CrossRef] [PubMed]
74. Talavera, D.; Laskowski, R.A.; Thornton, J.M. Wssas: A web service for the annotation of functional residues through structural homologues. *Bioinformatics* **2009**, *25*, 1192–1194. [CrossRef] [PubMed]
75. Schmidt, T.; Haas, J.; Gallo Cassarino, T.; Schwede, T. Assessment of ligand-binding residue predictions in casp9. *Proteins* **2011**, *79*, 126–136. [CrossRef] [PubMed]
76. Gallo Cassarino, T.; Bordoli, L.; Schwede, T. Assessment of ligand binding site predictions in CASP10. *Proteins* **2014**, *82*, 154–163. [CrossRef] [PubMed]
77. McDonald, A.G.; Tipton, K.F. Fifty-five years of enzyme classification: Advances and difficulties. *FEBS J.* **2014**, *281*, 583–592. [CrossRef] [PubMed]
78. Piovesan, D.; Giollo, M.; Leonardi, E.; Ferrari, C.; Tosatto, S.C. Inga: Protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res.* **2015**, *43*, 134–140. [CrossRef] [PubMed]
79. Gerlt, J.A.; Bouvier, J.T.; Davidson, D.B.; Imker, H.J.; Sadkhin, B.; Slater, D.R.; Whalen, K.L. Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **2015**, *1854*, 1019–1037. [CrossRef] [PubMed]
80. Sahraeian, S.M.; Luo, K.R.; Brenner, S.E. Sifter search: A web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res.* **2015**, *43*, 141–147. [CrossRef] [PubMed]
81. Gundersen, G.W.; Jones, M.R.; Rouillard, A.D.; Kou, Y.; Monteiro, C.D.; Feldmann, A.S.; Hu, K.S.; Ma'ayan, A. Geo2enrichr: Browser extension and server app to extract gene sets from geo and analyze them for biological functions. *Bioinformatics* **2015**. [CrossRef] [PubMed]
82. Koskinen, P.; Toronen, P.; Nokso-Koivisto, J.; Holm, L. Pannzer: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* **2015**, *31*, 1544–1552. [CrossRef] [PubMed]
83. Yu, G.; Zhu, H.; Domeniconi, C. Predicting protein functions using incomplete hierarchical labels. *BMC Bioinform.* **2015**, *16*, 1. [CrossRef] [PubMed]

84. Wang, T.; Mori, H.; Zhang, C.; Kurokawa, K.; Xing, X.H.; Yamada, T. Domsign: A top-down annotation pipeline to enlarge enzyme space in the protein universe. *BMC Bioinforma.* **2015**, *16*, 96. [CrossRef] [PubMed]
85. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227. [CrossRef] [PubMed]
86. Oh, M.; Joo, K.; Lee, J. Protein-binding site prediction based on three-dimensional protein modeling. *Proteins* **2009**, *77*, 152–156. [CrossRef] [PubMed]
87. Wass, M.N.; Kelley, L.A.; Sternberg, M.J. 3DLigandsite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **2010**, *38*, 469–473. [CrossRef] [PubMed]
88. Zhou, Y.; Xue, S.; Yang, J.J. Calciomics: Integrative studies of Ca^{2+}-binding proteins and their interactomes in biological systems. *Metallomics* **2013**, *5*, 29–42. [CrossRef] [PubMed]
89. Don, C.G.; Riniker, S. Scents and sense: *In silico* perspectives on olfactory receptors. *J. Comput. Chem.* **2014**, *35*, 2279–2287. [CrossRef] [PubMed]
90. Arredondo, J.; Lara, M.; Ng, F.; Gochez, D.A.; Lee, D.C.; Logia, S.P.; Nguyen, J.; Maselli, R.A. Cooh-terminal collagen Q (COLQ) mutants causing human deficiency of endplate acetylcholinesterase impair the interaction of ColQ with proteins of the basal lamina. *Hum. Genet.* **2014**, *133*, 599–616. [CrossRef] [PubMed]
91. Dong, Q.; Menon, R.; Omenn, G.S.; Zhang, Y. Structural bioinformatics inspection of nextprot PE5 proteins in the human proteome. *J. Proteome Res.* **2015**, *14*, 3750–3761. [CrossRef] [PubMed]
92. Bindschedler, L.V.; McGuffin, L.J.; Burgis, T.A.; Spanu, P.D.; Cramer, R. Proteogenomics and *in silico* structural and functional annotation of the barley powdery mildew *blumeria graminis* f. sp. *hordei*. *Methods* **2011**, *54*, 432–441. [CrossRef] [PubMed]
93. Pedersen, C.; Ver Loren van Themaat, E.; McGuffin, L.J.; Abbott, J.C.; Burgis, T.A.; Barton, G.; Bindschedler, L.V.; Lu, X.; Maekawa, T.; Wessling, R.; *et al.* Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics* **2012**, *13*, 694. [CrossRef] [PubMed]
94. Pavlidi, N.; Tseliou, V.; Riga, M.; Nauen, R.; Van Leeuwen, T.; Labrou, N.E.; Vontas, J. Functional characterization of glutathione S-transferases associated with insecticide resistance in *Tetranychus urticae*. *Pestic. Biochem. Physiol.* **2015**, *121*, 53–60. [CrossRef] [PubMed]
95. Taylor, T.B.; Mulley, G.; Dills, A.H.; Alsohim, A.S.; McGuffin, L.J.; Studholme, D.J.; Silby, M.W.; Brockhurst, M.A.; Johnson, L.J.; Jackson, R.W. Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system. *Science* **2015**, *347*, 1014–1017. [CrossRef] [PubMed]

Consortium *Genome Biology* (2019) 20:244
https://doi.org/10.1186/s13059-019-1835-8

RESEARCH Open Access

The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou[1,2], Yuxiang Jiang[3], Timothy R. Bergquist[4], Alexandra J. Lee[5], Balint Z. Kacsoh[6,7], Alex W. Crocker[8],
Kimberley A. Lewis[8], George Georghiou[9], Huy N. Nguyen[1,10], Md Nafiz Hamid[1,2], Larry Davis[2], Tunca Dogan[11,37],
Volkan Atalay[12], Ahmet S. Rifaioglu[12,13], Alperen Dalkıran[12], Rengul Cetin Atalay[14], Chengxin Zhang[15], Rebecca L.
Hurto[16], Peter L. Freddolino[15,16], Yang Zhang[15,16], Prajwal Bhat[17], Fran Supek[18,19], José M. Fernández[20,21], Branislava
Gemovic[22], Vladimir R. Perovic[22], Radoslav S. Davidović[22], Neven Sumonja[22], Nevena Veljkovic[22], Ehsaneddin
Asgari[23,24], Mohammad R.K. Mofrad[25], Giuseppe Profiti[26,27], Castrense Savojardo[26], Pier Luigi Martelli[26], Rita
Casadio[26], Florian Boecker[28], Heiko Schoof[29], Indika Kahanda[30], Natalie Thurlby[31], Alice C. McHardy[32,33],
Alexandre Renaux[34,35,36], Rabie Saidi[37], Julian Gough[38], Alex A. Freitas[39], Magdalena Antczak[40], Fabio Fabris[39],
Mark N. Wass[40], Jie Hou[41,42], Jianlin Cheng[42], Zheng Wang[43], Alfonso E. Romero[44], Alberto Paccanaro[44],
Haixuan Yang[45], Tatyana Goldberg[129], Chenguang Zhao[49], Liisa Holm[50], Petri Törönen[50], Alan J. Medlar[50],
Elaine Zosa[51], Itamar Borukhov[51], Ilya Novikov[53], Angela Wilkins[54], Olivier Lichtarge[54], Po-Han Chi[55], Wei-Cheng T
seng[56], Michal Linial[57], Peter W. Rose[58], Christophe Dessimoz[59,60,61], Vedrana Vidulin[62], Saso Dzeroski[63,64], Ian Sillit
oe[65], Sayoni Das[66], Jonathan Gill Lees[66,67], David T. Jones[69,70], Cen Wan[68,69], Domenico Cozzetto[68,69], Rui
Fa[68,69], Mateo Torres[44], Alex Warwick Vesztrocy[70,71], Jose Manuel Rodriguez[72], Michael L. Tress[73], Marco
Frasca[74], Marco Notaro[74], Giuliano Grossi[74], Alessandro Petrini[74], Matteo Re[74], Giorgio Valentini[74], Marco
Mesiti[74], Daniel B. Roche[76], Jonas Reeb[76], David W. Ritchie[77], Sabeur Aridhi[77], Seyed Ziaeddin Alborzi[77,79],
Marie-Dominique Devignes[77,78,79], Da Chen Emily Koo[80], Richard Bonneau[81,82], Vladimir Gligorijević[83], Meet
Barot[84], Hai Fang[85], Stefano Toppo[86], Enrico Lavezzo[86], Marco Falda[87], Michele Berselli[86], Silvio C.E.
Tosatto[88,89], Marco Carraro[89], Damiano Piovesan[89], Hafeez Ur Rehman[90], Qizhong Mao[91,92], Shanshan Zhang[91],
Slobodan Vucetic[91], Gage S. Black[93,94], Dane Jo[93,94], Erica Suh[93], Jonathan B. Dayton[93,94], Dallas J. Larsen[93,94],
Ashton R. Omdahl[93,94], Liam J. McGuffin[95], Danielle A. Brackenridge[95], Patricia C. Babbitt[96,98], Jeffrey M. Yunes[97,98],
Paolo Fontana[99], Feng Zhang[100,101], Shanfeng Zhu[102,103,104], Ronghui You[102,103,104], Zihan Zhang[102,104], Suyang
Dai[102,104], Shuwei Yao[102,103], Weidong Tian[105,106], Renzhi Cao[107], Caleb Chandler[107], Miguel Amezola[107], Devon Jo
hnson[107], Jia-Ming Chang[108], Wen-Hung Liao[108], Yi-Wei Liu[108], Stefano Pascarelli[109], Yotam Frank[110], Robert Hoeh
ndorf[111], Maxat Kulmanov[111], Imane Boudellioua[112,113], Gianfranco Politano[114], Stefano Di Carlo[114], Alfredo
Benso[114], Kai Hakala[115,116], Filip Ginter[115,117], Farrokh Mehryary[115,116], Suwisa Kaewphan[115,116,118], Jari
Björne[119,120], Hans Moen[117], Martti E.E. Tolvanen[121], Tapio Salakoski[119,120], Daisuke Kihara[122,123], Aashish Jain[124],
Tomislav Šmuc[125], Adrian Altenhoff[126,127], Asa Ben-Hur[128], Burkhard Rost[129,130], Steven E. Brenner[131], Christine A.
Orengo[66], Constance J. Jeffery[132], Giovanni Bosco[133], Deborah A. Hogan[6,8], Maria J. Martin[9], Claire
O'Donovan[9], Sean D. Mooney[4], Casey S. Greene[134,135], Predrag Radivojac[136*] and Iddo Friedberg[1*]

*Correspondence: predrag@northeastern.edu; idoerg@iastate.edu
[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA
[136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
Full list of author information is available at the end of the article

Abstract

Background: The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function.

Results: Here, we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in *Candida albicans* and *aeruginosa* genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility. We further performed targeted assays on selected genes in *Drosophila melanogaster*, which we suspected of being involved in long-term memory.

Conclusion: We conclude that while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than the expectations set by baseline methods in *C. albicans* and *D. melanogaster*, it leaves considerable room and need for improvement. Finally, we report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.

Keywords: Protein function prediction, Long-term memory, Biofilm, Critical assessment, Community challenge

Introduction

High-throughput nucleic acid sequencing [1] and mass-spectrometry proteomics [2] have provided us with a deluge of data for DNA, RNA, and proteins in diverse species. However, extracting detailed functional information from such data remains one of the recalcitrant challenges in the life sciences and biomedicine. Low-throughput biological experiments often provide highly informative empirical data related to various functional aspects of a gene product, but these experiments are limited by time and cost. At the same time, high-throughput experiments, while providing large amounts of data, often provide information that is not specific enough to be useful [3]. For these reasons, it is important to explore computational strategies for transferring functional information from the group of functionally characterized macromolecules to others that have not been studied for particular activities [4–9].

To address the growing gap between high-throughput data and deep biological insight, a variety of computational methods that predict protein function have been developed over the years [10–24]. This explosion in the number of methods is accompanied by the need to understand how well they perform, and what improvements are needed to satisfy the needs of the life sciences community. The Critical Assessment of Functional Annotation (CAFA) is a community challenge that seeks to bridge the gap between the ever-expanding pool of molecular data and the limited resources available to understand protein function [25–27].

The first two CAFA challenges were carried out in 2010–2011 [25] and 2013–2014 [26]. In CAFA1, we adopted a time-delayed evaluation method, where protein sequences that lacked experimentally verified annotations, or *targets*, were released for prediction. After the submission deadline for predictions, a subset of these targets accumulated experimental annotations over time, either as a consequence of new publications about these proteins or the biocuration work updating the annotation databases. The members of this set of proteins were used as *benchmarks* for evaluating the participating computational methods, as the function was revealed only after the prediction deadline.

CAFA2 expanded the challenge founded in CAFA1. The expansion included the number of ontologies used for predictions, the number of target and benchmark proteins, and the introduction of new assessment metrics that mitigate the problems with functional similarity calculation over concept hierarchies such as Gene Ontology [28]. Importantly, we provided evidence that the top-scoring methods in CAFA2 outperformed the top-scoring methods in CAFA1, highlighting that methods participating in CAFA improved over the 3-year period. Much of this improvement came as a consequence of novel methodologies with some effect of the expanded annotation databases [26]. Both CAFA1 and CAFA2 have shown that computational methods designed to perform function prediction outperform a conventional function transfer through sequence similarity [25, 26].

In CAFA3 (2016–2017), we continued with all types of evaluations from the first 2 challenges and additionally performed experimental screens to identify genes associated with specific functions. This allowed us to provide unbiased evaluation of the term-centric performance based on a unique set of benchmarks obtained by assaying *Candida albicans*, *Pseudomonas aeruginosa*, and *Drosophila melanogaster*. We also held a challenge following CAFA3, dubbed CAFA-π, to provide the participating teams another opportunity to develop or modify prediction models. The genome-wide screens on *C. albicans* identified 240 genes previously not known to be involved in biofilm formation, whereas the screens on *P. aeruginosa* identified 532 new genes involved in biofilm formation and 403 genes involved in motility. Finally, we used CAFA predictions to select genes from *D. melanogaster* and assay them for long-term memory involvement. This experiment allowed us to both evaluate prediction methods and identify 11 new fly genes involved in this biological process [29]. Here, we present the outcomes of the CAFA3 challenge, as well as the accompanying challenge CAFA-π, and discuss further directions for the community interested in the function of biological macromolecules.

Results

Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2

One of CAFA's major goals is to quantify the progress in function prediction over time. We therefore conducted a comparative evaluation of top CAFA1, CAFA2, and CAFA3 methods according to their ability to predict Gene Ontology [28] terms on a set of common benchmark proteins. This benchmark set was created as an intersection of CAFA3 benchmarks (proteins that gained experimental annotation after the CAFA3 prediction submission deadline) and CAFA1 and CAFA2 target proteins. Overall, this set contained 377 protein sequences with annotations in the Molecular Function Ontology (MFO), 717 sequences in the Biological Process Ontology (BPO), and 548 sequences in the Cellular Component Ontology (CCO), which allowed for a direct comparison of all methods that have participated in the challenges so far. The head-to-head comparisons in MFO, BPO, and CCO between the top 5 CAFA3 and CAFA2 methods are shown in Fig. 1. CAFA3 and CAFA1 comparisons are shown in Additional file 1: Figure S1.

We first observe that, in effect, the performance of baseline methods [25, 26] has not improved since CAFA2. The Naïve method, which uses the term frequency in the existing annotation database as a prediction score for every input protein, has the same $F_{\max}$ performance using both annotation databases in 2014 (when CAFA2 was held) and in 2017 (when CAFA3 was held), which suggests little change in term frequencies in the annotation database since 2014. In MFO, the BLAST method based on the existing annotations in 2017 is slightly but significantly better than the BLAST method based on 2014 training data. In BPO and CCO, however, the BLAST based on the later database has not outperformed its earlier counterpart, although the changes in effect size (absolute change in $F_{\max}$) in both ontologies are small.

When surveying all 3 CAFA challenges, the performance of both baseline methods has been relatively stable, with some fluctuations of BLAST. Such performance of direct sequence-based function transfer is surprising, given the steady growth of annotations in UniProt-GOA [30]; that is, there were 259,785 experimental annotations in 2011, 341,938 in 2014, and 434,973 in 2017, but there does not seem to be a definitive trend with the BLAST method, as they go up and down in $F_{\max}$ across ontologies. We conclude from these observations on the baseline methods that first, the ontologies are in different annotation states and should not be treated as a whole. In fact, the distribution of annotation depth and information content is very different across 3 ontologies, as shown in Additional file 1: Figures S15 and S16. Second, methods that perform direct function transfer based on sequence similarity do not necessarily benefit from a larger training dataset. Although the performance observed in our work is also dependent on the benchmark set, it appears that the annotation databases remain too sparsely populated to effectively exploit function transfer by sequence similarity, thus justifying the need for advanced methodology development for this problem.

Head-to-head comparisons of the top 5 CAFA3 methods against the top 5 CAFA2 methods show mixed results. In MFO, the top CAFA3 method, GOLabeler [23], outperformed all CAFA2 methods by a considerable margin, as shown in Fig. 2. The rest of the 4 CAFA3 top methods did not perform as well as the top 2 methods of CAFA2, although only to a limited extent, with little change in $F_{\max}$. Of the top 12 methods ranked in MFO, 7 are from CAFA3, 5 are from CAFA2, and none from CAFA1. Despite the increase in database size, the majority of function prediction methods do not seem to have improved in predicting protein function in MFO since 2014, except for 1 method that stood out. In BPO, the top 3 methods in CAFA3 outperformed their CAFA2 counterparts, but with very small margins. Out of the top 12 methods in BPO, 8 are from CAFA3, 4 are from CAFA2, and none from CAFA1. Finally, in CCO, although 8 out of the top 12 methods over all CAFA challenges come from CAFA3, the top method is from CAFA2. The differences between the top-performing methods are small, as in the case of BPO.

The performance of the top methods in CAFA2 was significantly better than of those in CAFA1, and it is interesting to note that this trend has not continued in CAFA3.

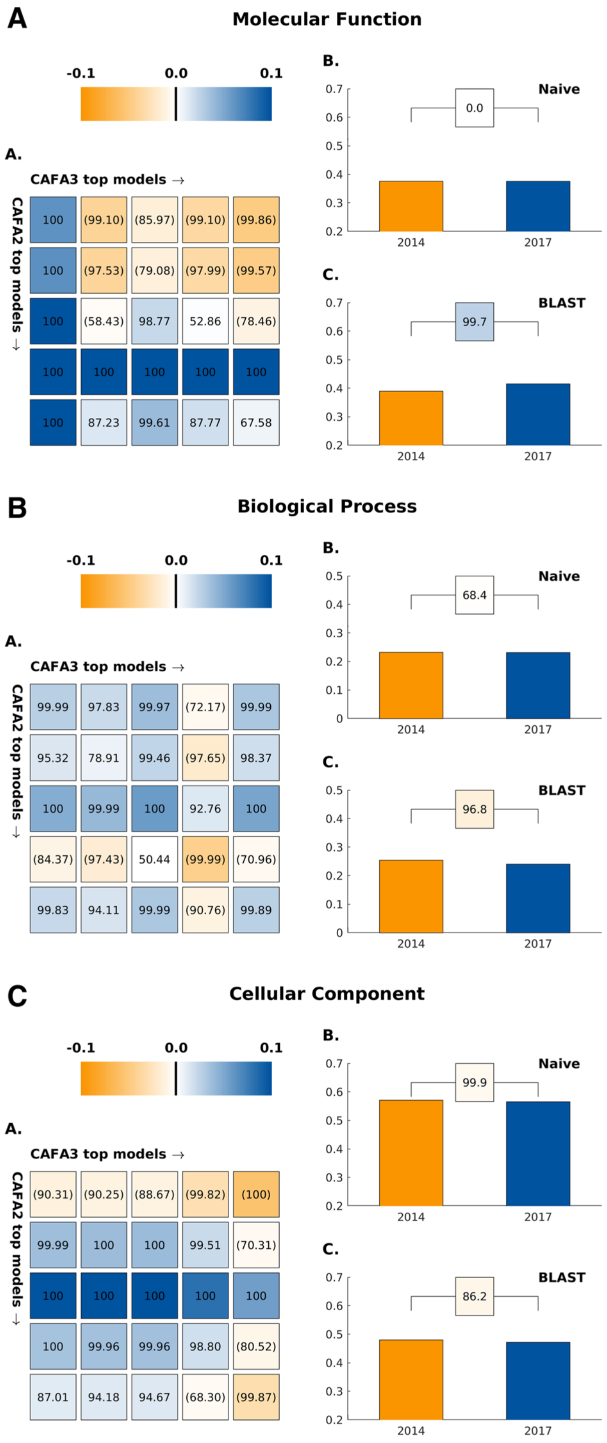


Fig. 1 A comparison in $F_{\max}$ between the top 5 CAFA2 models against the top 5 CAFA3 models. Colored boxes encode the results such that (1) the colors indicate margins of a CAFA3 method over a CAFA2 method in $F_{\max}$ and (2) the numbers in the box indicate the percentage of wins. **a** CAFA2 top 5 models (rows, from top to bottom) against CAFA3 top 5 models (columns, from left to right). **b** Comparison of the performance ($F_{\max}$) of Naïve baselines trained respectively on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. **c** Comparison of the performance ($F_{\max}$) of BLAST baselines trained on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. Statistical significance was assessed using 10,000 bootstrap samples of benchmark proteins

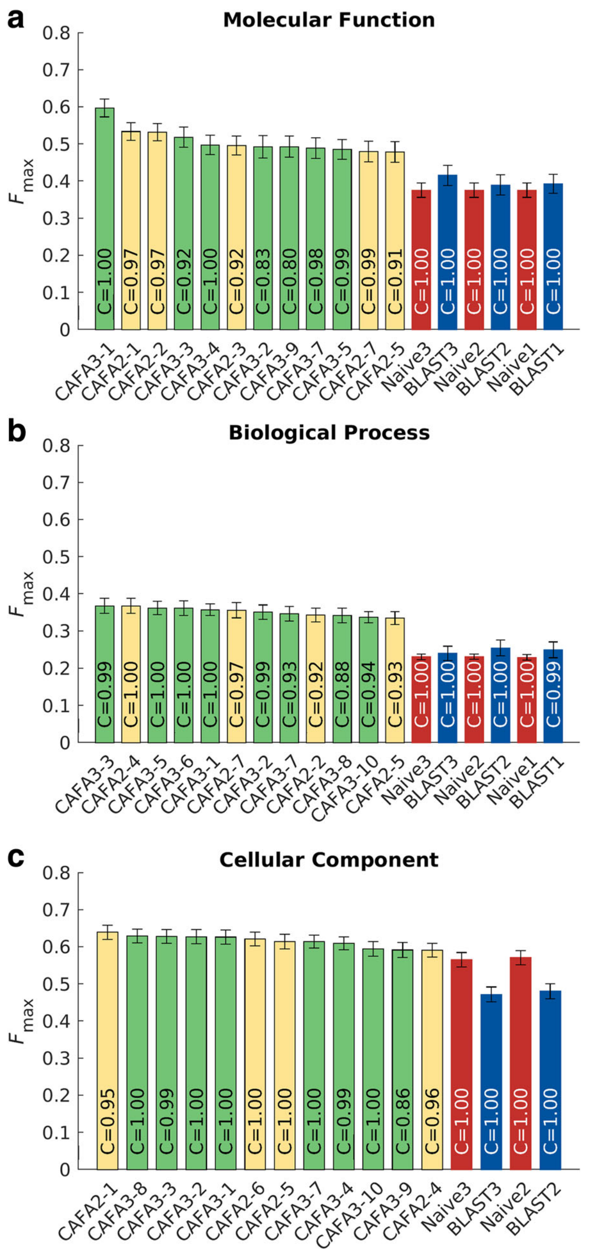


Fig. 2 Performance evaluation based on the $F_{\max}$ for the top CAFA1, CAFA2, and CAFA3 methods. The top 12 methods are shown in this barplot ranked in descending order from left to right. The baseline methods are appended to the right; they were trained on training data from 2017, 2014, and 2011, respectively. Coverage of the methods were shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark that are predicted by the methods. Color scheme: CAFA2, ivory; CAFA3, green; Naïve, red; BLAST, blue. Note that in MFO and BPO, CAFA1 methods were ranked, but since none made to the top 12 of all 3 CAFA challenges, they were not displayed. The CAFA1 challenge did not collect predictions for CCO. **a**: molecular function; **b**: Biological process; **c**: Cellular Component

This could be due to many reasons, such as the quality of the benchmark sets, the overall quality of the annotation database, the quality of ontologies, or a relatively short period of time between challenges.

Protein-centric evaluation

The *protein-centric* evaluation measures the accuracy of assigning GO terms to a protein. This performance is shown in Figs. 3 and 4.

We observe that all top methods outperform the baselines with the patterns of performance consistent with CAFA1 and CAFA2 findings. Predictions of MFO terms achieved the highest $F_{\max}$ compared with predictions in the other two ontologies. BLAST outperforms Naïve in predictions in MFO, but not in BPO or CCO. This is because sequence similarity-based methods such as BLAST tend to perform best when transferring basic biochemical annotations such as enzymatic activity. Functions in biological process, such as pathways, may not be as preserved by sequence similarity, hence the poor BLAST performance in BPO. The reasons behind the difference among the three ontologies include the structure and complexity of the ontology as well as the state of the annotation database, as discussed previously [26, 31]. It is less clear why the performance in CCO is weak, although it might be hypothesized that such performance is related to the structure of the ontology itself [31].

The top-performing method in MFO did not have as high an advantage over others when evaluated using the $S_{\min}$ metric. The $S_{\min}$ metric weights GO terms by conditional information content, since the prediction of more informative terms is more desirable than less informative, more general, terms. This could potentially explain the smaller gap between the top predictor and the rest of the pack in $S_{\min}$. The weighted $F_{\max}$ and normalized $S_{\min}$ evaluations can be found in Additional file 1: Figures S4 and S5.

Species-specific categories

The benchmarks in each species were evaluated individually as long as there were at least 15 proteins per species. Here, we present the results from eukaryotic and bacterial species (Fig. 5). We observed that different methods could perform differently on different species. As shown in Fig. 6, bacterial proteins make up a small portion of all benchmark sequences, so their effects on the performances of the methods are often masked. Species-specific analyses are therefore useful to researchers studying certain organisms. Evaluation results on individual species including human (Additional file 1: Figure S6), *Arabidopsis thaliana* (Additional file 1: Figure S7) and *Escherichia coli* (Additional file 1: Figure S10) can be found in Additional file 1: Figure S6-S14.

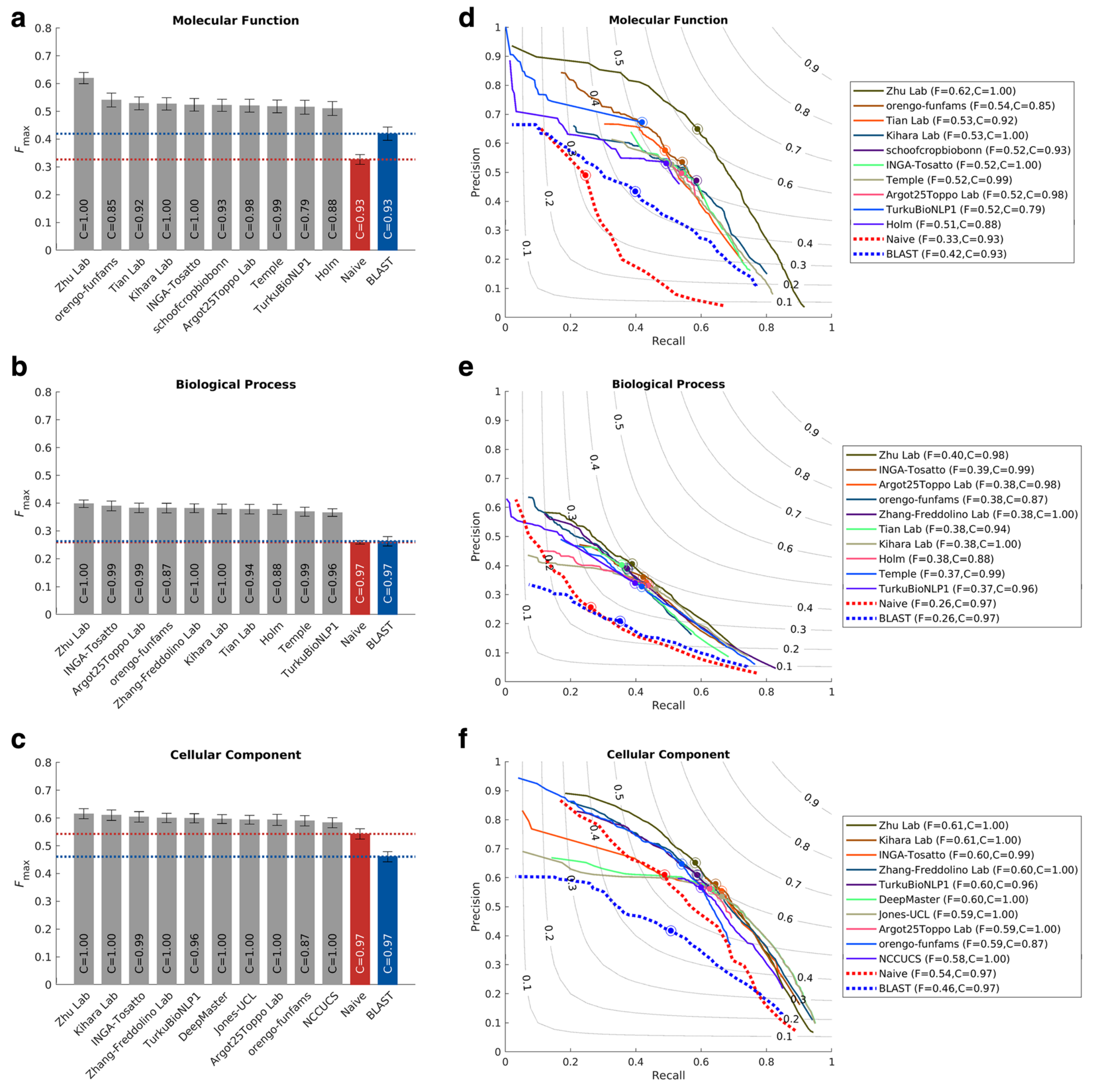


Fig. 3 Performance evaluation based on the F_{max} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing the F_{max} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: precision-recall curves for the top 10 methods. The perfect prediction should have $F_{max} = 1$, at the top right corner of the plot. The dot on the curve indicates where the maximum *F* score is achieved

Diversity of methods

It was suggested in the analysis of CAFA2 that ensemble methods that integrate data from different sources have the potential of improving prediction accuracy [32]. Multiple data sources, including sequence, structure, expression profile, genomic context, and molecular interaction data, are all potentially predictive of the function of the protein. Therefore, methods that take advantage of these rich sources as well as existing techniques from other research groups might see improved performance. Indeed, the one method that stood out from the rest in CAFA3 and performed significantly better than all methods across three challenges is a machine learning-based ensemble method [23]. Therefore, it is important to analyze what information sources and prediction algorithms are better at predicting function. Moreover, the similarity

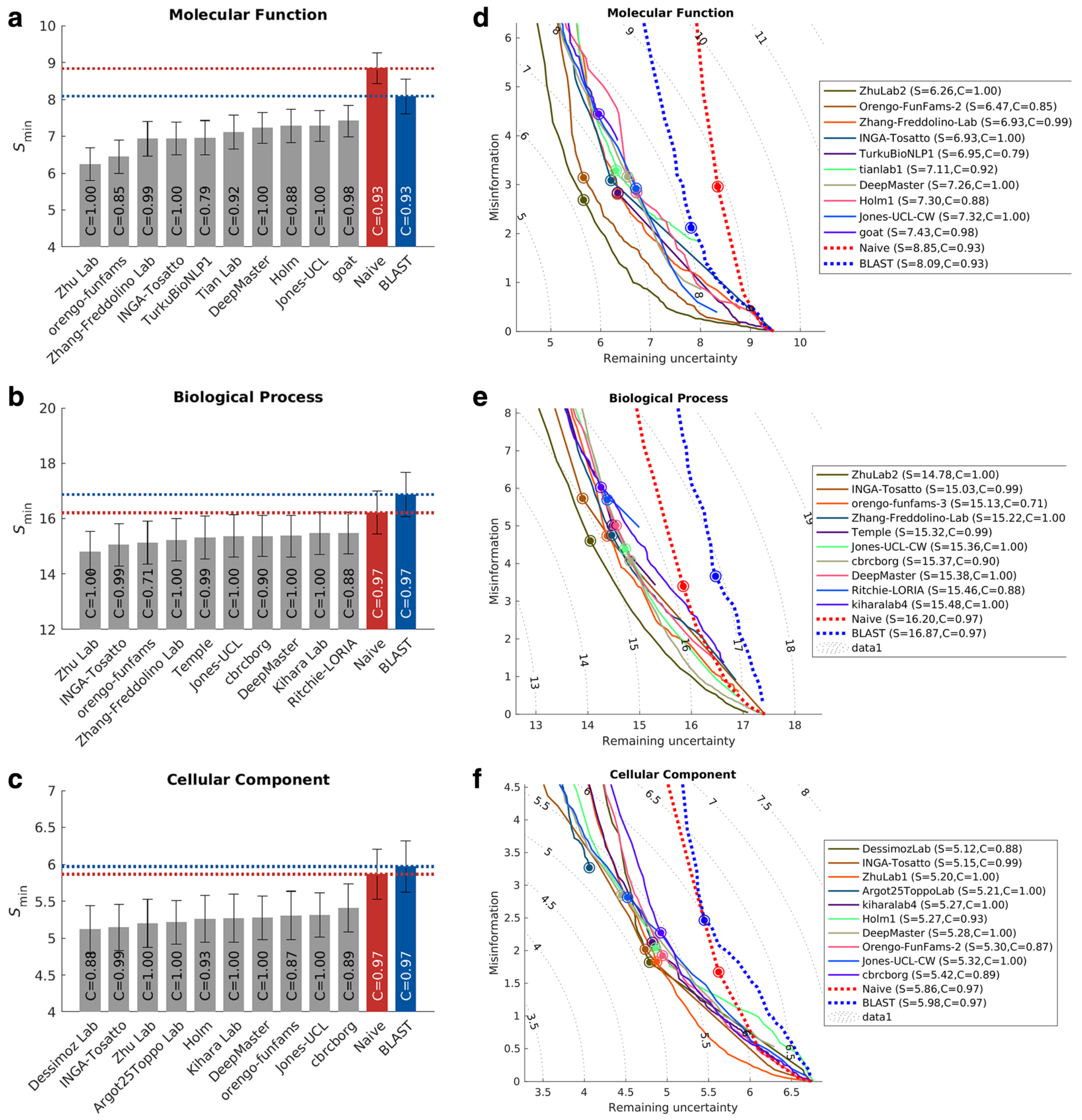


Fig. 4 Performance evaluation based on S_{min} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing S_{min} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: remaining uncertainty-missing information (RU-MI) curves for the top 10 methods. The perfect prediction should have $S_{min} = 0$, at the bottom left corner of the plot. The dot on the curve indicates where the minimum semantic distance is achieved

of the methods might explain the limited improvement in the rest of the methods in CAFA3.

The top CAFA2 and CAFA3 methods are very similar in performance, but that could be a result of aggregating predictions of different proteins to one metric. When computing the similarity of each pair of methods as the Euclidean distance of prediction scores (Fig. 7), we are not interested whether these predictions are correct according to the benchmarks, but simply whether they are similar to one another. The diagonal blocks in Fig. 7 show that

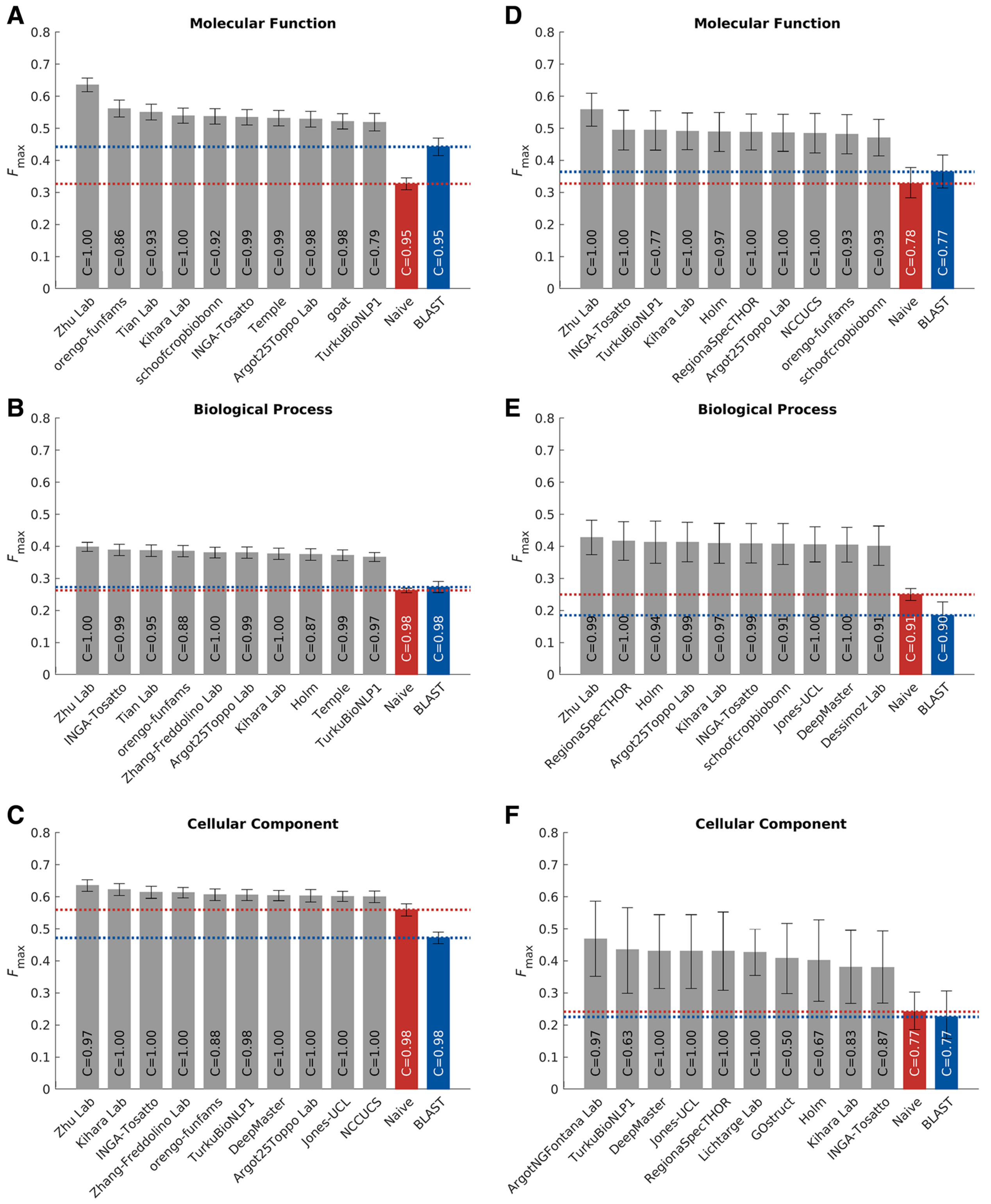


Fig. 5 Evaluation based on the F_{max} for the top-performing methods in eukaryotic and bacterial species

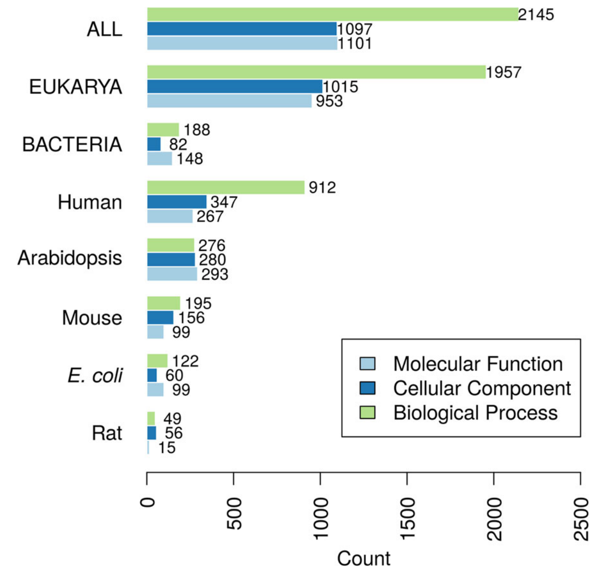


Fig. 6 Number of proteins in each benchmark species and ontology

CAFA1 top methods are more diverse than CAFA2 and CAFA3. The off-diagonal blocks shows that CAFA2 and CAFA3 methods are more similar with each other than with CAFA3 methods. It is clear that some methods are heavily based on the Naïve and BLAST baseline methods.

Participating teams also provided keywords that describe their approach to function prediction with their submissions. A list of keywords was given to the participants, listed in Additional file 1. Figure 8 shows the frequency of each keyword. In addition, we have weighted the frequency of the keywords with the prediction accuracy of the specific method. Machine learning and sequence alignment remain the most used approach by scientists predicting in all three ontologies. By raw count, machine learning is more popular than sequence in all three ontologies, but once adjusted by performance, their difference shrinks. In MFO, sequence alignment even overtakes machine learning as the most popular keyword after adjusting for performance. This indicates that methods that use sequence alignments are more helpful in predicting the correct function than the popularity of their use suggests.

Evaluation via molecular screening

Databases with proteins annotated by biocuration, such as UniProt knowledge base and UniProt Gene Ontology Annotation (GOA) database, have been the primary source of benchmarks in the CAFA challenges. New to CAFA3, we also evaluated the extent to which methods participating in CAFA could predict the results of genetic screens in model organisms done specifically for this project. Predicting GO terms for a protein (protein-centric) and predicting which proteins are associated with a given function (term-centric) are related but different computational problems: the former is a multi-label classification problem with a structured output, while the latter is a binary classification task. Predicting the results of a genome-wide screen for a single or a small number of functions fits the term-centric formulation. To see how well all participating CAFA methods perform term-centric predictions, we mapped the results from the protein-centric CAFA3 methods onto these terms. In addition, we held a separate CAFA challenge, CAFA-π, whose purpose was to attract additional submissions from algorithms that specialize in term-centric tasks.

We performed screens for three functions in three species, which we then used to assess protein function prediction. In the bacterium *Pseudomonas aeruginosa* and the fungus *Candida albicans*, we performed genome-wide screens capable of uncovering genes with two functions, biofilm formation (GO:0042710) and motility (for *P. aeruginosa* only) (GO:0001539), as described in the “Methods” section. In *Drosophila melanogaster*, we performed targeted assays, guided by previous CAFA submissions, of a selected set of genes and assessed whether or not they affected long-term memory (GO:0007616).

We discuss the prediction results for each function below in detail. The performance, as assessed by the genome-wide screens, was generally lower than in the protein-centric evaluations that were curation driven. We hypothesize that it may simply be more difficult to perform term-centric prediction for broad activities such as biofilm formation and motility. For *P. aeruginosa*, an existing compendium of gene expression data was already available [33]. We used the Pearson correlation over this collection of data to provide a complementary baseline to the standard BLAST approach used throughout CAFA. We found that an expression-based method outperformed the CAFA participants, suggesting that success on certain term-centric challenges will require the use of different types of data. On the other hand, the performance of the methods in predicting long-term memory in the Drosophila genome was relatively accurate.

Biofilm formation

In March 2018, there were 3019 annotations to biofilm formation (GO:0042710) and its descendent terms across all species, of which 325 used experimental evidence codes. These experimentally annotated proteins included 131 from the Candida Genome Database [34] for *C. albicans* and 29 for *P. aeruginosa*, the 2 organisms that we screened.

Of the 2746 genes we screened in the *Candida albicans* colony biofilm assay, 245 were required for the formation of wrinkled colony biofilm formation (Table 1). Of these, only 5 were already annotated in UniProt: *MOB*,

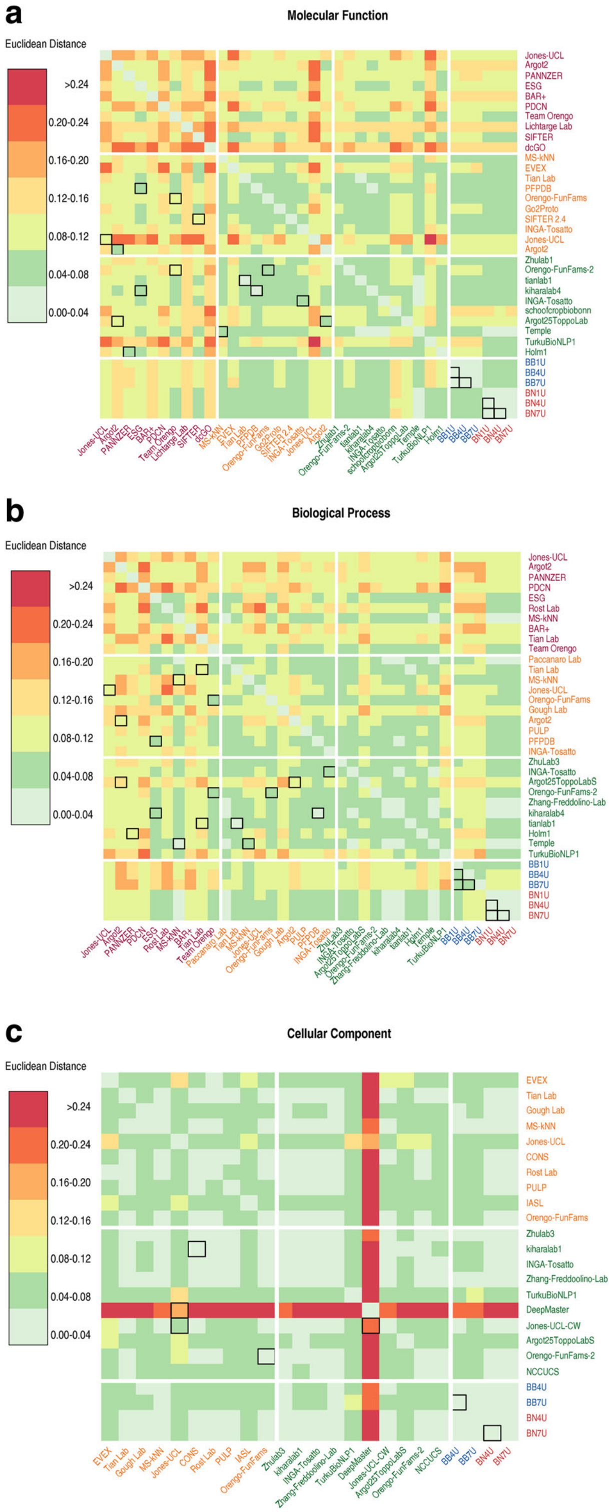
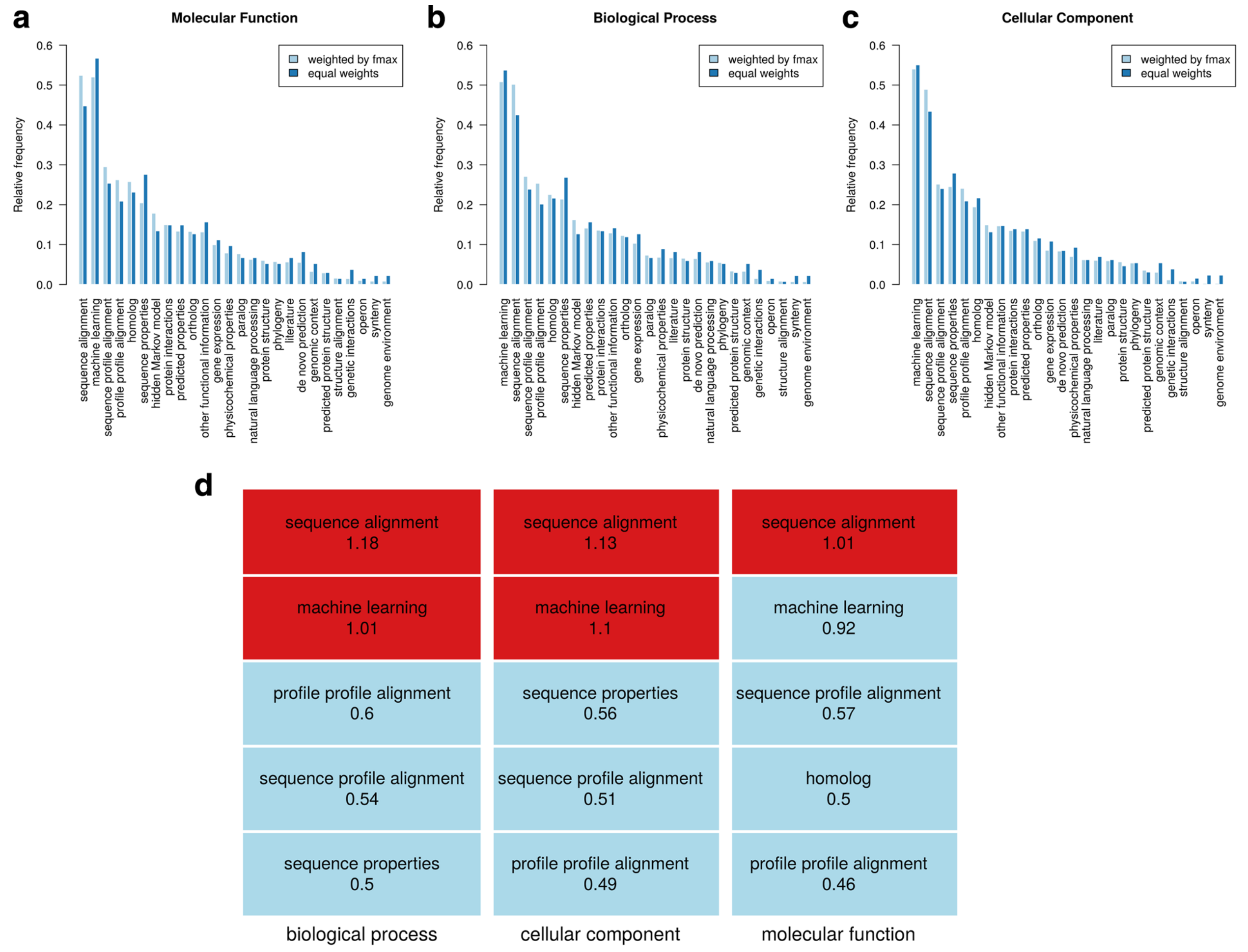


Fig. 7 Heatmap of similarity for the top 10 methods in CAFA1, CAFA2, and CAFA3. Similarity is represented by Euclidean distance of the prediction scores from each pair of methods, using the intersection set of benchmarks in the "Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2" section. The higher (darker red color) the euclidean distance, the less similar the methods are. Top 10 methods from each of the CAFA challenges are displayed and ranked by their performance in $F_{\max}$. Cells highlighted by black borders are between a pair of methods that come from the same PI. **a**: Molecular Function; **b**: Biological Process; **c**: Cellular Component



biological process	cellular component	molecular function
sequence alignment 1.18	sequence alignment 1.13	sequence alignment 1.01
machine learning 1.01	machine learning 1.1	machine learning 0.92
profile profile alignment 0.6	sequence properties 0.56	sequence profile alignment 0.57
sequence profile alignment 0.54	sequence profile alignment 0.51	homolog 0.5
sequence properties 0.5	profile profile alignment 0.49	profile profile alignment 0.46

Fig. 8 Keyword analysis of all CAFA3 participating methods. **a**–**c**: both relative frequency of the keywords and weighted frequencies are provided for three respective GO ontologies. The weighted frequencies accounts for the performance of the the particular model using the given keyword. If that model performs well (with high $F_{\max}$), then it gives more weight to the calculation of the total weighted average of that keyword. **d** shows the ratio of relative frequency between the $F_{\max}$-weighted and equal-weighted. Red indicates the ratio is greater than one while blue indicates the ratio is less than one. Only the top five keywords ranked by ratio are shown. The larger the ratio, the more difference there is between the $F_{\max}$-weighted and the equal-weighted

EED1 (*DEF1*), and *YAK1*, which encode proteins involved in hyphal growth, an important trait for biofilm formation [35–38]. Also, *NUP85*, a nuclear pore protein involved in early phase arrest of biofilm formation [39] and *VPS1*, contributes to protease secretion, filamentation, and biofilm formation [40]. Of the 2063 proteins that we did not find to be associated with biofilm formation, 29 were annotated with the term in the GOA database in *C. albicans*. Some of the proteins in this category highlight the need for additional information to GO term annotation. For example, Wor1 and the pheromone receptor are key for biofilm formation in strains under conditions in which the mating pheromone is produced [41], but not required in the monocultures of the commonly studied a/α mating type strain used here.

Table 1 Number of proteins in *Candida albicans* and *Pseudomonas aeruginosa* associated with the GO term "Biofilm formation" (GO:0042710) in the GOA databases versus experimental results

			GOA annotations	
	Total, 2308		Unannotated	Annotated
C. albicans	CAFA experiments	False	2034	29
		True	240	5
	Total, 4056		Unannotated	Annotated
P. aeruginosa	CAFA experiments	False	3491	25
		True	532	9

We used receiver operating characteristic (ROC) curves to measure the prediction accuracy. Area under ROC curves (AUROC) was used to compare the performance. AUROC is a common accuracy measure for classification problems where it evaluates how good a model is at distinguishing between the positive and negative classes. No method in CAFA-π or CAFA3 (not shown) exceeded an AUC of 0.60 on this term-centric challenge (Fig. 9) for either species. Performance for the best methods slightly exceeded a BLAST-based baselines. In the past, we have found that predicting BPO terms, such as biofilm formation, resulted in poorer method performance than predicting MFO terms. Many CAFA methods use sequence alignment as their primary source of information (the "Diversity of methods" section). For *Pseudomonas aeruginosa*, a pre-built expression compendium was available from prior work [33]. Where the compendium was available, simple gene expression-based baselines were the best-performing approaches. This suggests that successful term-centric prediction of biological processes may need to rely more heavily on information that is not sequence-based and, as previously reported, may require methods that use broad collections of gene expression data [42, 43].

Motility

In March 2018, there were 302,121 annotations for proteins with the GO term: cilium or flagellum-dependent cell motility (GO:0001539) and its descendent terms, which included cell motility in all eukaryotic (GO:0060285), bacterial (GO:0071973), and archaeal (GO:0097590) organisms. Of these, 187 had experimental

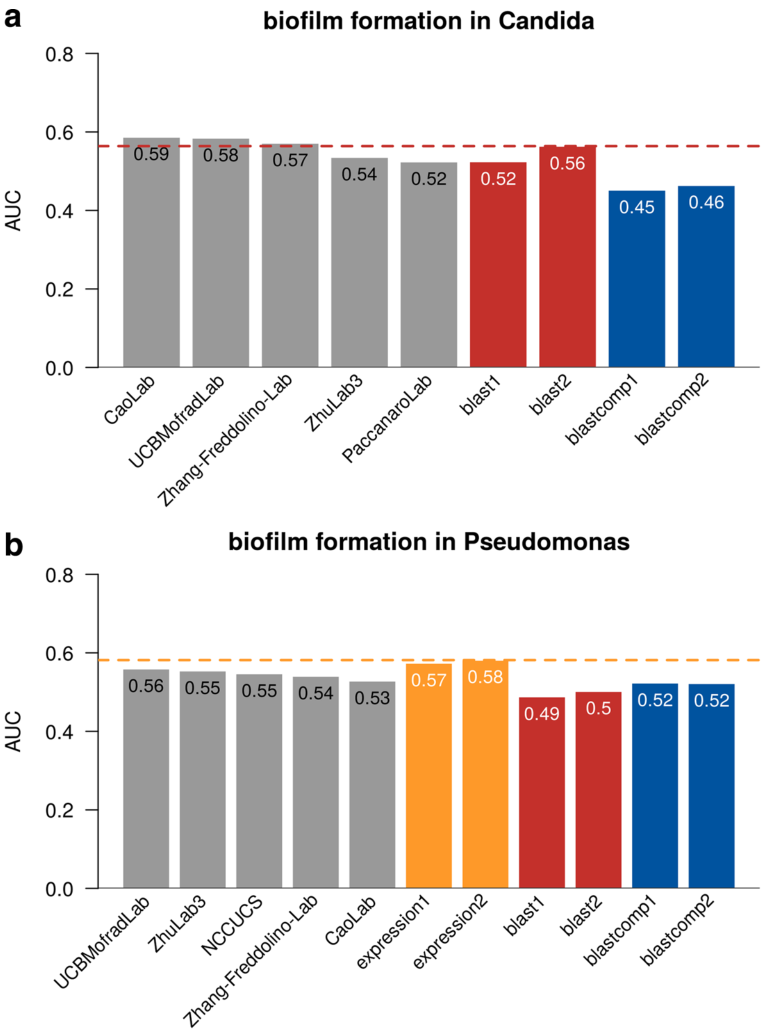


Fig. 9 AUROC of the top five teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. Four baseline models all based on BLAST were computed for *Candida*, while six baseline models were computed for *Pseudomonas*, including two based on expression profiles. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue, and expression are in yellow, see Table 3 for the description of the baselines

evidence codes, and the most common organism to have annotations was *P. aeruginosa*, on which our screen was performed (Additional file 1: Table S2).

As expected, mutants defective in the flagellum or its motor were defective in motility (*fliC* and other *fli* and *flg* genes). For some of the genes that were expected, but not detected, the annotation was based on the experiments performed in a medium different from what was used in these assays. For example, PhoB regulates motility but only when phosphate concentration is low [44]. Among the genes that were scored as defective in motility, some are known to have decreased motility due to over production of carbohydrate matrix material (*bifA*) [45], or the absence of directional swimming due to absence of chemotaxis functions (e.g., *cheW*, *cheA*) and others likely showed this phenotype because of a medium-specific requirement such as biotin (*bioA*, *bioC*, and *bioD*) [46]. Table 2 shows the contingency table for the number of proteins that are detected by our experiment versus GOA annotations.

The results from this evaluation were consistent with what we observed for biofilm formation. Many of the genes annotated as being involved in biofilm formation were identified in the screen. Others that were annotated as being involved in biofilm formation did not show up in the screen because the strain background used here, strain PA14, uses the exopolysaccharide matrix carbohydrate Pel [47] in contrast to the Psl carbohydrate used by another well-characterized strain, strain PAO1 [48, 49]. The *psl* genes were known to be dispensable for biofilm formation in the strain PA14 background, and this nuance highlights the need for more information to be taken into account when making predictions.

The CAFA-π methods outperformed our BLAST-based baselines but failed to outperform the expression-based baselines. Transferred methods from CAFA3 also did not outperform these baselines. It is important to note this consistency across terms, reinforcing the finding that term-centric prediction of biological processes is likely to require non-sequence information to be included.

Long-term memory in D. melanogaster

Prior to our experiments, there were 1901 annotations made in the long-term memory, including 283 experimental annotations. *Drosophila melanogaster* had the most annotated proteins of long-term memory with 217, while human has 7, as shown in Additional file 1: Table S3.

Table 2 Number of proteins in *Pseudomonas aeruginosa* associated with function motility (GO:0001539) in the GOA databases versus experimental results

		GOA annotations	
Total, 3630		Unannotated	Annotated
CAFA experiments	False	3195	12
	True	403	21

We performed RNAi experiments in *Drosophila melanogaster* to assess whether 29 target genes were associated with long-term memory (GO:0007616). Briefly, flies were exposed to wasps, which triggers a behavior that causes females to lay fewer eggs. The acute response is measured until 24 h post-exposure, and the long-term response is measured at 24 to 48 h post-exposure. RNAi was used to interfere with the expression of the 29 target genes in the mushroom body, a region of the fly brain associated with memory. Using this assay, we identified 3 genes involved in the perception of wasp exposure and 12 genes involved in the long-term memory. For details on the target selection and fly assay, see [29]. None of the 29 genes had an existing annotation in the GOA database. Because no genome-wide screen results were available, we did not release this as part of the CAFA-π and instead relied only on the transfer of methods that predicted the “long-term memory" at least once in *D. melanogaster* from CAFA3. Results from this assessment were more promising than our findings from the genome-wide screens in microbes (Fig. 10). Certain methods performed well, substantially exceeding the baselines.

Participation growth

The CAFA challenge has seen growth in participation, as shown in Fig. 11. To cope with the increasingly large data size, CAFA3 utilized the Synapse [50] online platform for submission. Synapse allowed for easier access for participants, as well as easier data collection for the organizers. The results were also released to the individual teams via this online platform. During the submission process, the online platform also allows for customized format checkers to ensure the quality of the submission.

Methods

Benchmark collection

In CAFA3, we adopted the same benchmark generation methods as CAFA1 and CAFA2, with a similar timeline (Fig. 12). The crux of a time-delayed challenge is the annotation growth period between time t_0 and t_1. All target proteins that have gained experimental annotation during this period are taken as benchmarks in all three ontologies. “No knowledge” (NK, no prior experimental annotations) and “Limited knowledge” (LK, partial prior experimental annotations) benchmarks were also distinguished based on whether the newly gained experimental annotation is in an ontology that already have experimental annotations or not. Evaluation results in Figs. 3 and 4 are made using the No knowledge benchmarks. Evaluation results on the Limited knowledge benchmarks are

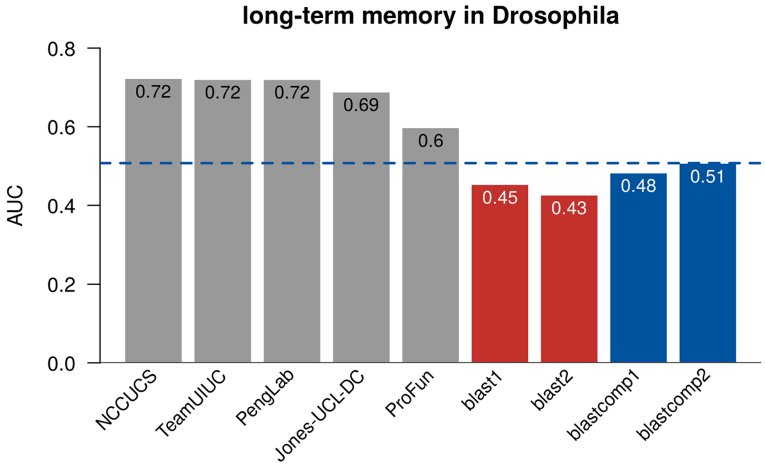


Fig. 10 AUROC of top five teams in CAFA3. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red and BLAST computational methods are in blue, see Table 3 for the description of the baselines

shown in Additional file 1: Figure S3. For more information regarding NK and LK designations, please refer to the Additional file 1 and the CAFA2 paper [26].

After collecting these benchmarks, we performed two major deletions from the benchmark data. Upon inspecting the taxonomic distribution of the benchmarks, we noticed a large number of new experimental annotations from *Candida albicans*. After consulting with UniProt-GOA, we determined these annotations have already existed in the Candida Genome Database long before 2018 but were only recently migrated to GOA. Since these annotations were already in the public domain before the CAFA3 submission deadline, we have deleted any annotation from *Candida albicans* with an assigned date prior to our CAFA3 submission deadline. Another major change is the deletion of any proteins with only a protein-binding (GO:0005515) annotation. Protein binding is a highly generalized function description, does not provide more specific information about the actual function of a protein, and in many cases may indicate a non-functional, non-specific binding. If it is the only annotation that a protein has gained, then it is hardly an advance in our understanding of that protein; therefore, we deleted these annotations from our benchmark set. Annotations with a depth of 3 make up almost half of all annotations in MFO before the removal (Additional file 1: Figure S15B). After the removal, the most frequent annotations became of depth 5 (Additional file 1: Figure S15A). In BPO, the most frequent annotations are of depth 5 or more, indicating a healthy increase of specific GO terms being added to our annotation database. In CCO, however, most new annotations in our benchmark set are of depths 3, 4, and 5 (Additional file 1: Figure S15). This difference could partially explain why the same computational methods perform very differently in different ontologies and benchmark sets. We have also calculated the total

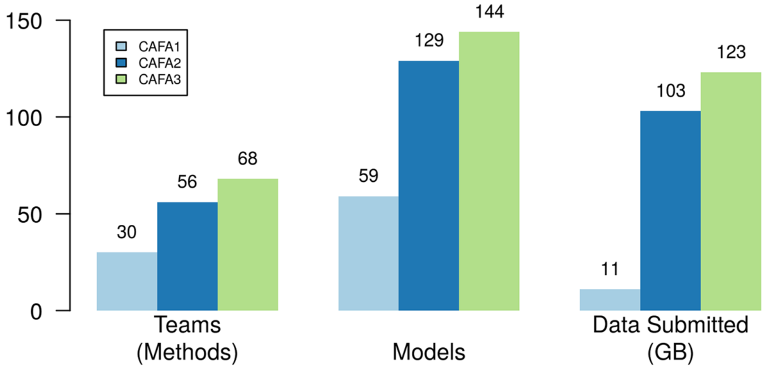


Fig. 11 CAFA participation has been growing. Each principal investigator is allowed to head multiple teams, but each member can only belong to one team. Each team can submit up to three models

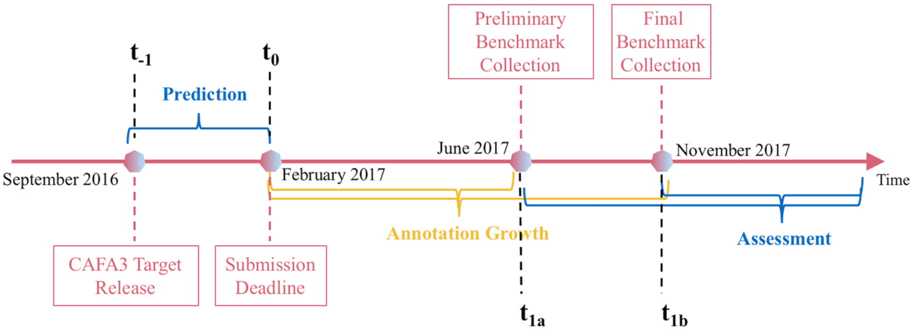


Fig. 12 CAFA3 timeline

information content per protein for the benchmark sets shown in Additional file 1: Figure S16. Taxonomic distributions of the proteins in our final benchmark set are shown in Fig. 6.

Additional analyses were performed to assess the characteristics of the benchmark set, including the overall information content of the terms being annotated.

Protein-centric evaluation

Two main evaluation metrics were used in CAFA3, the $F_{\max}$ and the $S_{\min}$. The $F_{\max}$ based on the precision-recall curve (Fig. 3), while the $S_{\min}$ is based on the remaining uncertainty/missing information (RU-MI) curve as described in [51] (Fig. 4), where S stands for semantic distance. The shortest semantic distance across all thresholds is used as the $S_{\min}$ metric. The RU-MI curve takes into account the information content of each GO term in addition to counting the number of true positives, false positives, etc., see Additional file 1 for the precise definition of $F_{\max}$ and $S_{\min}$. The information theory-based evaluation metrics counter the high-throughput low-information annotations such as protein binding, but down-weighing these terms according to their information content, as the ability to predict such non-specific functions are not as desirable and useful and the ability to predict more specific functions.

The two assessment modes from CAFA2 were also used in CAFA3. In the partial mode, predictions were evaluated only on those benchmarks for which a model made at least one prediction. The full evaluation mode evaluates all benchmark proteins, and methods were penalized for not making predictions. Evaluation results in Figs. 3 and 4 are made using the full evaluation mode. Evaluation results using the partial mode are shown in Additional file 1: Figure S2.

Two baseline models were also computed for these evaluations. The Naïve method assigns the term frequency as the prediction score for any protein, regardless of any protein-specific properties. BLAST was based on the results using the Basic Local Alignment Search Tool (BLAST) software against the training database [52]. A term will be predicted as the highest local alignment sequence identity among all BLAST hits annotated from the training database. Both of these methods were trained on the experimentally annotated proteins and their sequences in Swiss-Prot [53] at time t_0.

Microbe screens

To assess the matrix production, we used mutants from the PA14 NR collection [54]. Mutants were transferred from the − 80 °C freezer stock using a sterile 48-pin multiprong device into 200 μl LB in a 96-well plate. The cultures were incubated overnight at 37 °C, and their OD600 was measured to assess growth. Mutants were then transferred to tryptone agar with 15 g of tryptone and 15 g of agar in 1L amended with Congo red (Aldrich, 860956) and Coomassie brilliant blue (J.T. Baker Chemical Co., F789-3). Plates were incubated at 37 °C overnight followed by 4-day incubation at room temperature to allow the wrinkly phenotype to develop. Colonies were imaged and scored on day 5. To assess motility, mutants were revived from freezer stocks as described above. After overnight growth, a sterile 48-pin multiprong transfer device with a pin diameter of 1.58 mm was used to stamp the mutants from the overnight plates into the center of swim agar made with M63 medium with 0.2% glucose and casamino acids and 0.3% agar). Care was taken to avoid touching the bottom of the plate. Swim plates were incubated at room temperature (19–22 °C) for approximately 17 h before imaging and scoring. Experimental procedures in *P. aeruginosa* to determine proteins that are associated with the two functions in CAFA-π are shown in Fig. 13.

Biofilm formation in *Candida albicans* was assessed in single gene mutants from the Noble [55] and GRACE [56] collections. In the Noble Collection, mutants of *C. albicans* have had both copies of the candidate gene deleted. Most of the mutants were created in biological duplicate. From this collection, 1274 strains corresponding to 653 unique genes were screened. The GRACE Collection

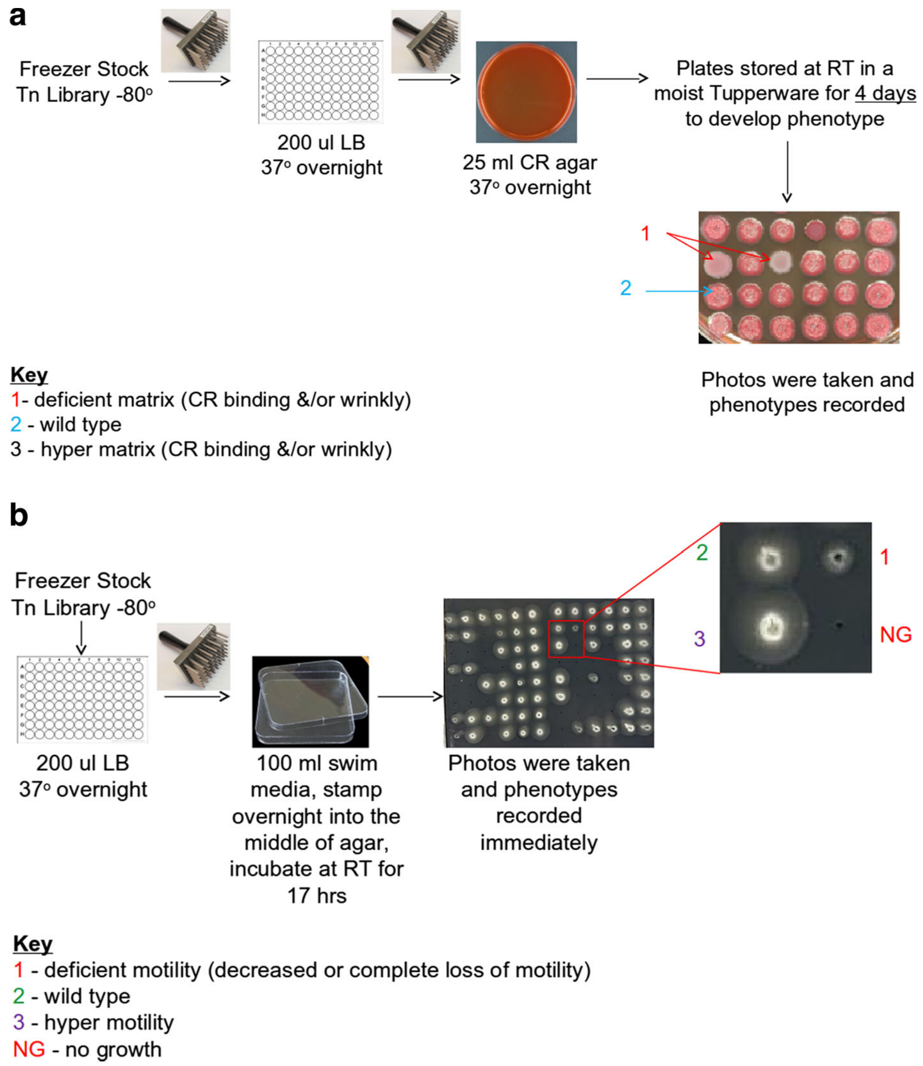


Fig. 13 Experimental procedure of determining genes associated with the functions biofilm formation (**a**) and motility (**b**) in *P. aeruginosa*

provided mutants with one copy of each gene deleted and the other copy placed under the control of a doxycycline-repressible promoter. To assay these strains, we used a medium supplemented with 100 μg/ml doxycycline strains, when rendered them functional null mutants. We screened 2348 mutants from the GRACE Collection, 255 of which overlapped with mutants in the Noble Collection, for 2746 total unique mutants screened in total. To assess the defects in biofilm formation or biofilm-related traits, we performed 2 assays: (1) colony morphology on agar medium and (2) biofilm formation on a plastic surface (Fig. 14). For both of these assays, we used Spider medium, which was designed to induce hyphal growth in *C. albicans* [57] and which promotes biofilm formation [39]. Strains were first replicated from frozen 96-well plates to YPD agar plates. Strains were then replicated from YPD agar to YPD broth and grown overnight at 30 °C. From YPD broth, strains were introduced onto Spider agar plates and into 96-well plates of Spider broth. When strains from the GRACE Collection were assayed, 100 μg/ml doxycycline was included in the agar and broth, and aluminum foil was used to protect the media from light. Spider agar plates inoculated with *C. albicans* mutants were incubated at 37 °C for 2 days before colony morphologies were scored. Strains in Spider broth were shaken at 225 rpm at 37 °C for 3 days and then assayed for biofilm formation at the air-liquid interface as follows. First, broth was removed by slowly tilting the plates and pulling the liquid away by running a gloved hand over the surface. Biofilms were stained by adding 100 μl of 0.1 percent crystal violet dye in water to each well of the plate. After 15 min, plates were gently washed in three

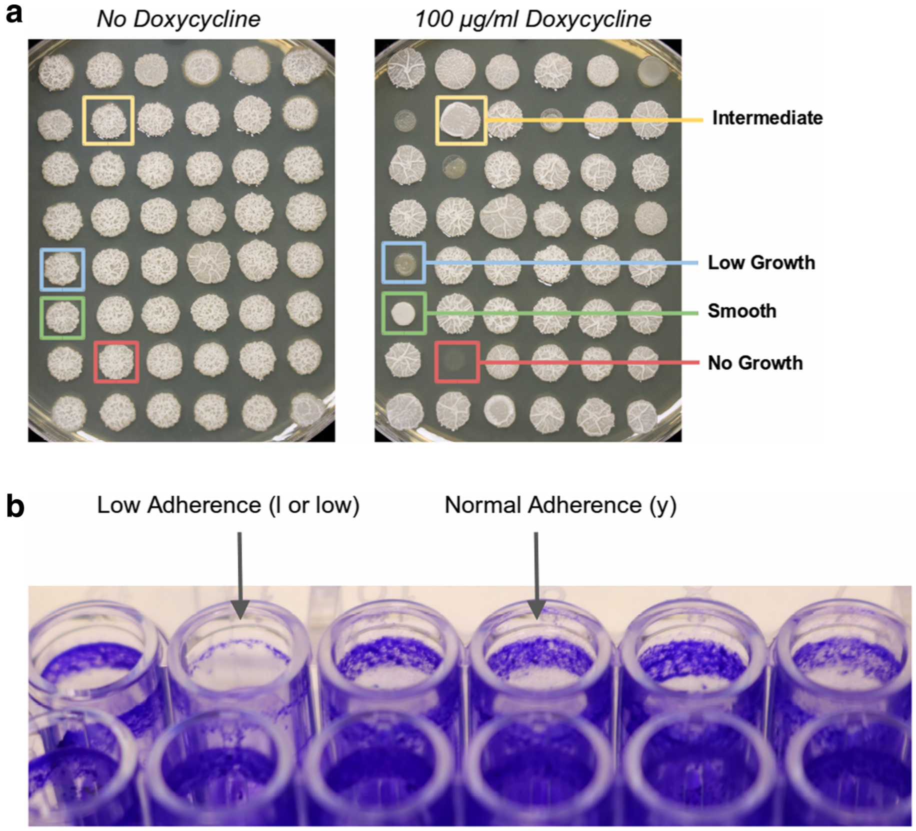


Fig. 14 a: different phenotypes in response to doxycycline treatment: low growth, smooth, no growth and intermediate. **b**: adherence phenotypes. See text for details

baths of water to remove dye without disturbing biofilms. To score biofilm formation for agar plates, colonies were scored by eye as either smooth, intermediate, or wrinkled. A wild-type colony would score wrinkled, and mutants with intermediate or smooth appearance were considered defective in colony biofilm formation. For biofilm formation on a plastic surface, the presence of a ring of cell material in the well indicated normal biofilm formation, while low or no ring formation mutants were considered defective. Genes whose mutations resulted defects in both or either assay were considered true for biofilm function. A complete list of the mutants identified in the screens is available in Additional file 1: Table S1.

A protein is considered true in the biofilm function, if its mutant phenotype is smooth or intermediate under doxycycline.

Term-centric evaluation

The evaluations of the CAFA-π methods were based on the experimental results in the "Microbe screens" section. We adopted $F_{\max}$ based on both precision-recall curves and area under ROC curves. There are a total of six baseline methods, as described in Table 3.

Discussion

Since 2010, the CAFA community has been the home to a growing group of scientists across the globe sharing the goal of improving computational function prediction. CAFA has been advancing this goal in three ways. First, through independent evaluation of computational methods against the set of benchmark proteins, thus providing a direct comparison of the methods' reliability and performance at a given time point. Second, the challenge assesses the quality of the current state of the annotations, whether they are made computationally or not, and is set up to reliably track it over time. Finally, as described in this work, CAFA has started to drive the creation of new experimental annotations by facilitating synergies between different groups of researchers interested in function of biological macromolecules. These annotations not only represent new biological discoveries, but simultaneously serve to provide benchmark data for rigorous method evaluation.

CAFA3 and CAFA-π feature the latest advances in the CAFA series to create advanced and accurate methods for protein function prediction. We use the repeated nature of the CAFA project to identify certain trends

Table 3 Baseline methods in term-centric evaluation of protein function prediction

	Model number	Training data	Score assignment
Expression	1	Gene expression compendium for *P. aeruginosa PAO1*	Highest correlation score out of all pairwise correlations
	2		Top 10 average correlation score
BLAST	1	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	Highest sequence identity out of all pairwise BLASTp hits
blastcomp	1	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	

via historical assessments. The analysis revealed that the performance of CAFA methods improved dramatically between CAFA1 and CAFA2. However, the protein-centric results for CAFA3 are mixed when compared to historical methods. Though the best-performing CAFA3 method outperformed the top CAFA2 methods (Fig. 1), this was not consistently true for other rankings. Among all 3 CAFA challenges, CAFA2 and CAFA3 methods inhabit the top 12 places in MFO and BPO. Between CAFA2 and CAFA3, the performance increase is more subtle. Based on the annotations of methods (Additional file 1), many of the top-ranking methods are improved versions of the methods that have been evaluated in CAFA2. Interestingly, the top-performing CAFA3 method, which consistently outperformed the methods from all past CAFAs in the major categories (GOLabeler [23]), utilized 5 component classifiers trained from different features; those included GO term frequency, sequence alignment, amino acid trigram, domains, motifs, and biophysical properties. It performs best in the Molecular Function Ontology, where sequence features perform best. Another method which did not participate in CAFA3 yet seems to perform well under CAFA parameters is NetGO [58], which utilizes the information from STRING, a network association database [59] in addition to sequence information. Taken together, the strong predictive performance of mRNA co-expression data (Figs. 9 and 15) leads us to hypothesize that including more varied sources of data can lead to additional large improvements in protein function prediction. We are looking forward to testing this hypothesis in future CAFA challenges. It should be noted that CAFA uses both $F_{\max}$ and $S_{\min}$.

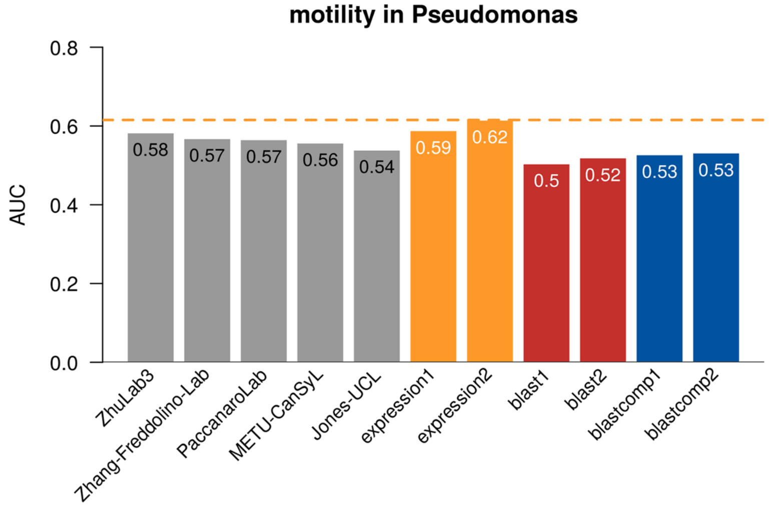


Fig. 15 AUROC of top 5 teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue and expression are in yellow. See Table 3 for description of the baselines

$F_{\max}$'s strength lies in its interpretability, as it is simply the maximum F_1 given for each model. At the same time, precision/recall-based assessment does not capture the hierarchical nature of ontologies or the differences in information content between different GO terms. For that reason, we also use the $S_{\min}$ score which incorporates information content, but is somewhat less interpretable than $F_{\max}$ and less robust to the problems of incomplete annotation [60, 61]. Additionally, since the information content of a GO term is derived from its frequency in the corpus [62], it is somewhat malleable depending on the corpus from which it is derived. We therefore use both measures for scoring, to achieve a more comprehensive picture of the models' performance.

For this iteration of CAFA, we performed genome-wide screens of phenotypes in *P. aeruginosa* and *C. albicans* as well as a targeted screen in *D. melanogaster*. This not only allowed us to assess the accuracy with which methods predict genes associated with select biological processes, but also to use CAFA as an additional driver for new biological discovery. Note that high-throughput screening for a single phenotype should be interpreted with caution as the phenotypic effect may be the result of pleiotropy, and the phenotype in question may be expressed as part of a set of other phenotypes. The results of genome-wide screenings typically lack context for the observed phenotypic effects, and each genotype-phenotype association should be examined individually to ascertain how immediate is the phenotypic effect from the seeming genotypic cause.

In sum, our experimental work identified more than a thousand new functional annotations in three highly divergent species. Though all screens have certain limitations, the genome-wide screens also bypass questions of biases in curation. This evaluation provides key insights: CAFA3 methods did not generalize well to selected terms. Because of that, we ran a second effort, CAFA-π, in which participants focused solely on predicting the results of these targeted assays. This targeted effort led to improved performance, suggesting that when the goal is to identify genes associated with a specific phenotype, tuning methods may be required.

For CAFA evaluations, we have included both Naïve and sequence-based (BLAST) baseline methods. For the evaluation of *P. aeruginosa* screen results, we were also able to include a gene expression baseline from a previously published compendium [33]. Intriguingly, the expression-based predictions outperformed the existing methods for this task. In future CAFA efforts, we will include this type of baseline expression-based method across evaluations to continue to assess the extent to which this data modality informs gene function prediction. The results from the CAFA3 effort suggest that gene expression may be particularly important for successfully predicting term-centric biological process annotations.

The primary takeaways from CAFA3 are as follows: (1) genome-wide screens complement annotation-based efforts to provide a richer picture of protein function prediction; (2) the best-performing method was a new method, instead of a light retooling of an existing approach; (3) gene expression, and more broadly, systems data may provide key information to unlocking biological process predictions, and (4) performance of the best methods has continued to improve. The results of the screens released as part of CAFA3 can lead to a re-examination of approaches which we hope will lead to improved performance in CAFA4.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1835-8.

Additional file 1: Additional figures and tables referenced in the article.

Additional file 2: Review History.

Acknowledgements

NZ and IF acknowledge the invaluable input from Michael C Gerten and Shatabdi Sen and all members of the Friedberg Lab for the ongoing support for stimulating discussions.

Review history

The review history is available as Additional file 2.

Authors' contributions

The experiment was designed by IF, PR, CSG, SDM, COD, MJM, and NZ. NZ, YJ, MNH, HNN, AJL, and LD performed the computational analyses. NZ, SDM, and TM were responsible for managing the participants' submissions to the CAFA challenge. KAL, AWC, and DAH performed the experimental work in *C. albicans* and *P. aeruginosa*. BZK and GB performed the experimental work in *D. melanogaster*. CJJ, MJM, COD, and GG provided the novel biocurated data for the benchmarks and incorporated data into UniprotKB. All other co-authors developed the computational function prediction methods participating in the challenge, performed the computational protein function predictions, and submitted the results for analysis in CAFA3. NZ, IF, CSG, DAH, and PR wrote the manuscript. All authors have read and approved the final manuscript.

Funding

The work of IF was funded, in part, by the National Science Foundation award DBI-1458359. The work of CSG and AJL was funded, in part, by the National Science Foundation award DBI-1458390 and GBMF 4552 from the Gordon and Betty Moore Foundation. The work of DAH and KAL was funded, in part, by the National Science Foundation award DBI-1458390, National Institutes of Health NIGMS P20 GM113132, and the Cystic Fibrosis Foundation CFRDP STANTO19R0. The work of AP, HY, AR, and MT was funded by BBSRC grants BB/K004131/1, BB/F00964X/1 and BB/M025047/1, Consejo Nacional de Ciencia y Tecnología Paraguay (CONACyT) grants 14-INV-088 and PINV15-315, and NSF Advances in BioInformatics grant 1660648. The work of JC was partially supported by an NIH grant (R01GM093123) and two NSF grants (DBI 1759934 and IIS1763246). ACM acknowledges the support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2155 "RESIST" - Project ID 39087428. DK acknowledges the support from the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941). PB acknowledges the support from the National Institutes of Health (R01GM60595). GB and BZK acknowledge the support from the National Science Foundation (NSF 1458390) and NIH DP1MH110234. FS was funded by the ERC StG 757700 "HYPER-INSIGHT" and by the Spanish Ministry of Science, Innovation and Universities grant BFU2017-89833-P. FS further acknowledges the funding from the Severo Ochoa award to the IRB Barcelona. TS was funded by the Centre of Excellence project "BioProspecting of Adriatic

Sea", co-financed by the Croatian Government and the European Regional Development Fund (KK.01.1.1.01.0002). The work of SK was funded by ATT Tieto käyttöön grant and Academy of Finland. JB and HM acknowledge the support of the University of Turku, the Academy of Finland and CSC – IT Center for Science Ltd. TB and SM were funded by the NIH awards UL1 TR002319 and U24 TR002306. The work of CZ and ZW was funded by the National Institutes of Health R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of PWR was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. PR acknowledges NSF grant DBI-1458477. PT acknowledges the support from Helsinki Institute for Life Sciences. The work of AJM was funded by the Academy of Finland (No. 292589). The work of FZ and WT was funded by the National Natural Science Foundation of China (31671367, 31471245, 91631301) and the National Key Research and Development Program of China (2016YFC1000505, 2017YFC0908402]. CS acknowledges the support by the Italian Ministry of Education, University and Research (MIUR) PRIN 2017 project 2017483NH8. SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). PLF and RLH were supported by the National Institutes of Health NIH R35-GM128637 and R00-GM097033. JG, DTJ, CW, DC, and RF were supported by the UK Biotechnology and Biological Sciences Research Council (BB/N019431/1, BB/L020505/1, and BB/L002817/1) and Elsevier. The work of YZ and CZ was funded in part by the National Institutes of Health award GM083107, GM116960, and AI134678; the National Science Foundation award DBI1564756; and the Extreme Science and Engineering Discovery Environment (XSEDE) award MCB160101 and MCB160124. The work of BG, VP, RD, NS, and NV was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 173001. The work of YWL, WHL, and JMC was funded by the Taiwan Ministry of Science and Technology (106-2221-E-004-011-MY2). YWL, WHL, and JMC further acknowledge the support from "the Human Project from Mind, Brain and Learning" of the NCCU Higher Education Sprout Project by the Taiwan Ministry of Education and the National Center for High-performance Computing for computer time and facilities. The work of IK and AB was funded by Montana State University and NSF Advances in Biological Informatics program through grant number 0965768. BR, TG, and JR are supported by the Bavarian Ministry for Education through funding to the TUM. The work of RB, VG, MB, and DCEK was supported by the Simons Foundation, NIH NINDS grant number 1R21NS103831-01 and NSF award number DMR-1420073. CJJ acknowledges the funding from a University of Illinois at Chicago (UIC) Cancer Center award, a UIC College of Liberal Arts and Sciences Faculty Award, and a UIC International Development Award. The work of ML was funded by Yad Hanadiv (grant number 9660 /2019). The work of OL and IN was funded by the National Institute of General Medical Science of the National Institute of Health through GM066099 and GM079656. Research Supporting Plan (PSR) of University of Milan number PSR2018-DIP-010-MFRAS. AWV acknowledges the funding from the BBSRC (CASE studentship BB/M015009/1). CD acknowledges the support from the Swiss National Science Foundation (150654). CO and MJM are supported by the EMBL-European Bioinformatics Institute core funds and the CAFA BBSRC BB/N004876/1. GG is supported by CAFA BBSRC BB/N004876/1. SCET acknowledges funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778247 (IDPfun) and from COST Action BM1405 (NGP-net). SEB was supported by NIH/NIGMS grant R01 GM071749. The work of MLT, JMR, and JMF was supported by the National Human Genome Research Institute of the National of Health, grant numbers U41 HG007234. The work of JMF and JMR was also supported by INB Grant (PT17/0009/0001 - ISCIII-SGEFI / ERDF). VA acknowledges the funding from TUBITAK EEEAG-116E930. RCA acknowledges the funding from KanSil 2016K121540. GV acknowledges the funding from Università degli Studi di Milano - Project "Discovering Patterns in Multi-Dimensional Data" and Project "Machine Learning and Big Data Analysis for Bioinformatics". SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). RY and SY are supported by the 111 Project (NO. B18015), the key project of Shanghai Science & Technology (No. 16JC1420402), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), and ZJLab. ST was supported by project Ribes Network POR-FESR 3S4H (No. TOPP-ALFREVE18-01) and PRID/SID of University of Padova (No. TOPP-SID19-01). CZ and ZW were supported by the NIGMS grant R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of MK and RH was supported by the funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01-01 and URF/1/3790-01-01. The work of SDM is funded, in part, by NSF award DBI-1458443.

Availability of data and materials

Data repository: A data repository providing all additional data, analyses, and all anonymous prediction results for all methods are available at https://figshare.com/articles/Supplementary_data/8135393/3 [63]. Code: The assessment software used in this paper is available under GNU-GPLv3 license on GitHub in both Matlab [64] (https://doi.org/10.5281/zenodo.3403452) and Python [65](http://doi.org/10.5281/zenodo.3401694).

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA. [2]Program in Bioinformatics and Computational Biology, Ames, IA, USA. [3]Indiana University Bloomington, Bloomington, Indiana, USA. [4]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA, USA. [5]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA. [6]Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [7]Department of Molecular and Systems Biology, Hanover, NH, USA. [8]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [9]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom. [10]Program in Computer Science, Ames, IA, USA. [11]Department of Computer Engineering, Hacettepe University, Ankara, Turkey. [12]Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey. [13]Department of Computer Engineering, Iskenderun Technical University, Hatay, Turkey. [14]CanSyL, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey. [15]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [16]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA. [17]Achira Labs, Bangalore, India. [18]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [20]INB Coordination Unit, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Catalonia, Spain. [21](former) INB GN2, Structural and Computational Biology Programme, Spanish National Cancer Research Centre, Barcelona, Catalonia, Spain. [22]Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia. [23]Molecular Cell Biomechanics Laboratory, Departments of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [24]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Berkeley, CA, USA. [25]Departments of Bioengineering and Mechanical Engineering, Berkeley, CA, USA. [26]Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy. [27]National Research Council, IBIOM, Bologna, Italy. [28]University of Bonn: INRES Crop Bioinformatics, Bonn, North Rhine-Westphalia, Germany. [29]INRES Crop Bioinformatics, University of Bonn, Bonn, Germany. [30]Gianforte School of Computing, Montana State University, Bozeman, Montana, USA. [31]University of Bristol, Computer Science, Bristol, Bristol, United Kingdom. [32]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Brunswick, Germany. [33]RESIST, DFG Cluster of Excellence 2155, Brunswick, Germany. [34]Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel, Brussels, Belgium. [35]Machine Learning Group, Université libre de Bruxelles, Brussels, Belgium. [36]Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium. [37]European Molecular Biolo gy Labora tory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [38]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom. [39]University of Kent, School of Computing, Canterbury, United Kingdom. [40]School of Biosciences, University of Kent, Canterbury, Kent, United Kingdom. [41]University of Missouri, Computer Science, Columbia, Missouri, USA. [42]Department of Electrical Engineering and Computer Science, University

of Missouri, Columbia, MO, USA. [43]University of Miami, Coral Gables, Florida, USA. [44]Centre for Systems and Synthetic Biology, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, United Kingdom. [45]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Galway, Ireland. [46]Technical University of Munich, Garching, Germany. [47]Faculty for Informatics, Garching, Germany. [48]Department for Bioinformatics and Computational Biology, Garching, Germany. [49]School of Computing Sciences and Computer Engineering, Hattiesburg, Mississippi, USA. [50]Institute of Biotechnology, Helsinki Institute of Life Sciences, University of Helsinki, Finland, Helsinki, Finland. [51]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [52]Compugen Ltd., Holon, Israel. [53]Baylor College of Medicine, Department of Biochemistry and Molecular Biology, Houston, TX, USA. [54]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, USA. [55]National TsingHua University, Hsinchu, Taiwan. [56]Department of Electrical Engineering in National Tsing Hua University, Hsinchu City, Taiwan. [57]The Hebrew University of Jerusalem, Jerusalem, Israel. [58]University of California San Diego, San Diego Supercomputer Center, La Jolla, California, USA. [59]Department of Computational Biology and Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [60]Department of Genetics, Evolution & Environment, and Department of Computer Science, University College London, London, UK. [61]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [62]Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia. [63]Jozef Stefan Institute, Ljubljana, Slovenia. [64]Jozef Stefan International Postgraduate School, Ljubljana, Slovenia. [65]Research Department of Structural and Molecular Biology, University College London, London, England. [66]Research Department of Structural and Molecular Biology, University College London, London, United Kingdom. [67]Oxford Brookes University, Department of Health and Life Sciences, London, UK. [68]University College London, Department of Computer Science, London, United Kingdom. [69]The Francis Crick Institute, Biomedical Data Science Laboratory, London, United Kingdom. [70]Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom. [71]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [72]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain. [73]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [74]Università degli Studi di Milano - Computer Science Department - AnacletoLab, Milan, Milan, Italy. [75]Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Universite de Montpellier, Montpellier, France. [76]Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [77]University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France. [78]University of Lorraine, Nancy, Lorraine, France. [79]Inria, Nancy, France. [80]Department of Biology, New York University, New York, NY, USA. [81]NYU Center for Data Science, New York, NY 10010, USA. [82]Flatiron Institute, CCB, 10010 New York, NY, USA. [83]Center for Computational Biology (CCB), Flatiron Institute, Simons Foundation, New York, New York, USA. [84]Center for Data Science, New York University, New York 10011, NY, USA. [85]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [86]Department of Molecular Medicine, University of Padova, Padova, Italy. [87]Department of Biology - University of Padova, Padova, Italy. [88]CNR Institute of Neuroscience, Padova, Italy. [89]Department of Biomedical Sciences, University of Padua, Padova, Italy. [90]Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar, Khyber Pakhtoonkhwa, Pakistan. [91]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [92]University of California, Riverside, Philadelphia, PA, USA. [93]Department of Biology, Brigham Young University, Provo, UT, USA. [94]Bioinformatics Research Group, Provo, UT, USA. [95]School of Biological Sciences, University of Reading, Reading, England, United Kingdom. [96]Department of Pharmaceutical Chemistry, San Francisco, CA, USA. [97]UC Berkeley - UCSF Graduate Program in Bioengineering, University of California, San Francisco 94158, CA, USA. [98]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco 94158, CA, USA. [99]Research and Innovation Center, Edmund Mach Foundation, 38010 San Michele all'Adige, Italy. [100]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, Shanghai, China. [101]Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [102]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China. [103]Institute of Science and Technology for Brain-Inspired Intelligence and Shanghai Institute of Artificial Intelligence Algorithms, Fudan University, Shanghai, China. [104]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China. [105]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [106]Department of Pediatrics, Brain Tumor Center, Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [107]Pacific Lutheran University, Department of Computer Science, Tacoma, WA, USA. [108]Department of Computer Science, National Chengchi University, Taipei, Taiwan. [109]Okinawa Institute of Science and Technology, Tancha, Okinawa, Japan. [110]Tel Aviv University, Tel Aviv, Israel. [111]Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Jeddah, Saudi Arabia. [112]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [113]Computer, Electrical and Mathematical Sciences Engineering Division (CEMSE), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [114]Politecnico di Torino, Control and Computer Engineering Department, Torino, TO, Italy. [115]University of Turku, Department of Future Technologies, Turku NLP Group, Turku, Finland. [116]University of Turku Graduate School (UTUGS), Turku, Finland. [117]University of Turku, Turku, Finland. [118]Turku Centre for Computer Science (TUCS), Turku, Finland. [119]Department of Future Technologies, Faculty of Science and Engineering, University of Turku, FI-20014 Turku, Finland. [120]Turku Centre for Computer Science (TUCS), Agora, Vesilinnantie 3, FI-20500, Turku, Finland. [121]University of Turku, Department of Future Technologies, Turku, Finland. [122]Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA. [123]Department of Pediatrics, University of Cincinnati, Cincinnati 45229, OH, USA. [124]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [125]Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia. [126]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [127]SIB Swiss Institute of Bioinformatics, Zurich, Switzerland. [128]Department of Computer Science, Colorado State University, Fort Collins, CO, USA. [129]Department of Informatics, Bioinformatics & Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [130]Institute for Food and Plant Sciences WZW, Technische Universität München, Freising, Germany. [131]University of California, Berkeley, CA, USA. [132]Biological Sciences, University of Illinois at Chicago, Chicago, Illinois, USA. [133]Geisel School of Medicine at Dartmouth, Department of Molecular and Systems Biology, Hanover, NH, USA. [134]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [135]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, Pennsylvania, USA. [136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA.

Received: 16 May 2019 Accepted: 24 September 2019
Published online: 19 November 2019

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6): 333–51.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
3. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol. 2013;9(5): 1003063.
4. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. 2003;60(12):2637–50.
5. Friedberg I. Automated protein function prediction–the genomic challenge. Brief Bioinform. 2006;7(3):225–42.
6. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3:88.
7. Rentzsch R, Orengo CA. Protein function prediction–the power of multiplicity. Trends Biotechnol. 2009;27(4):210–9.
8. Shehu A, Barbara D, Molloy K. A survey of computational methods for protein function predictions. Cham: Springer; 2016, pp. 225–98.
9. Cozzetto D, Jones DT. Computational methods for annotation transfers from sequence. Methods Mol Biol. 2017;1446:55–67.

10. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA. 1999;96(8):4285–8.
11. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol. 2002;319(5):1257–65.
12. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. J Comput Biol. 2003;10(6):947–60.
13. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA. 2004;101(41):14754–9.
14. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005;21(21 Suppl 1):302–10.
15. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005;1(5):45.
16. Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005;6:247.
17. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 2006;15(6):1550–6.
18. Wass MN, Sternberg MJ. Confunc–functional annotation in the twilight zone. Bioinformatics. 2008;24(6):798–806.
19. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008;9(Suppl 1):4.
20. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol. 2010;8(2):357–76.
21. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins. 2011;79(7):2086–96.
22. Piovesan D, Tosatto SCE. INGA 2.0: improving protein function prediction for the dark proteome. Nucleic Acids Res. 2019;47(W1):373–8. https://doi.org/10.1093/nar/gkz375.
23. You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics. 2018;34(14):2465–73.
24. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multi-task deep neural networks. PLoS One. 2018;13(6):0198216.
25. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–7.
26. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184.
27. Friedberg I, Radivojac P. Community-wide evaluation of computational function prediction. Methods Mol Biol. 2017;1446:133–46.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
29. Kacsoh BZ, Barton S, Jiang Y, Zhou N, Mooney SD, Friedberg I, Radivojac P, Greene CS, Bosco G. New Drosophila long-term memory genes revealed by assessing computational function prediction methods. G3. 2019;9(1):251–67.
30. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. 2015;43(Database issue):1057–63.
31. Peng Y, Jiang Y, Radivojac P. Enumerating consistent sub-graphs of directed acyclic graphs: an insight into biomedical ontologies. Bioinformatics. 2018;34(13):313–22.
32. Wang L, Law J, Kale SD, Murali TM, Pandey G. Large-scale protein function prediction using heterogeneous ensembles. F1000Res. 2018;7.
33. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene CS. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst. 2017;5(1):63–71.
34. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45(Database issue):592–6.
35. Goyard S, Knechtle P, Chauvel M, Mallet A, Prevost MC, Proux C, Coppee JY, Schwarz P, Dromer F, Park H, Filler SG, Janbon G, d'Enfert C. The Yak1 kinase is involved in the initiation and maintenance of hyphal growth in Candida albicans. Mol Biol Cell. 2008;19(5):2251–66.
36. Gutierrez-Escribano P, Gonzalez-Novo A, Suarez MB, Li CR, Wang Y, de Aldana CR, Correa-Bordes J. CDK-dependent phosphorylation of Mob2 is essential for hyphal development in Candida albicans. Mol Biol Cell. 2011;22(14):2458–69.
37. Lassak T, Schneider E, Bussmann M, Kurtz D, Manak JR, Srikantha T, Soll DR, Ernst JF. Target specificity of the Candida albicans Efg1 regulator. Mol Microbiol. 2011;82(3):602–18.
38. Martin R, Moran GP, Jacobsen ID, Heyken A, Domey J, Sullivan DJ, Kurzai O, Hube B. The Candida albicans-specific gene EED1 encodes a key regulator of hyphal extension. PLoS One. 2011;6(4):18394.
39. Richard ML, Nobile CJ, Bruno VM, Mitchell AP. Candida albicans biofilm-defective mutants. Eukaryot Cell. 2005;4(8):1493–502.
40. Bernardo SM, Khalique Z, Kot J, Jones JK, Lee SA. Candida albicans VPS1 contributes to protease secretion, filamentation, and biofilm formation. Fungal Genet Biol. 2008;45(6):861–77.
41. Yi S, Sahni N, Daniels KJ, Lu KL, Huang G, Srikantha T, Soll DR. Self-induction of a/a or α/α biofilms in Candida albicans is a pheromone-based paracrine system requiring switching. Eukaryot Cell. 2011;10(6):753–60.
42. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, Paw J, Clore JJ, Mendoza RM, Luis BS, Nislow C, Giaever G, Costanzo M, Troyanskaya OG, Caudy AA. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLOS Genetics. 2009;5(3):1–16. https://doi.org/10.1371/journal.pgen.1000407.
43. Hibbs MA, Myers CL, Huttenhower C, Hess DC, Li K, Caudy AA, Troyanskaya OG. Directing experimental biology: a case study in mitochondrial biogenesis. PLOS Comput Biol. 2009;5(3):1–12. https://doi.org/10.1371/journal.pcbi.1000322.
44. Blus-Kadosh I, Zilka A, Yerushalmi G, Banin E. The effect of pstS and phoB on quorum sensing and swarming motility in Pseudomonas aeruginosa. PLoS One. 2013;8(9):74444.
45. Kuchma SL, Brothers KM, Merritt JH, Liberati NT, Ausubel FM, O'Toole GA. BifA, a cyclic-Di-GMP phosphodiesterase, inversely regulates biofilm formation and swarming motility by Pseudomonas aeruginosa PA14. J Bacteriol. 2007;189(22):8165–78.
46. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas Genome Database. Nucleic Acids Res. 2016;44(D1):646–53.
47. Friedman L, Kolter R. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. Mol Microbiol. 2004;51(3):675–90.
48. Friedman L, Kolter R. Two genetic loci produce distinct carbohydrate-rich structural components of the Pseudomonas aeruginosa biofilm matrix. J Bacteriol. 2004;186(14):4457–65.
49. Jackson KD, Starkey M, Kremer S, Parsek MR, Wozniak DJ. Identification of psl, a locus encoding a potential exopolysaccharide that is essential for Pseudomonas aeruginosa PAO1 biofilm formation. J Bacteriol. 2004;186(14):4466–75.
50. Synapse. https://www.synapse.org/. Accessed 1 Jan 2016.
51. Clark WT, Radivojac P. Information-theoretic evaluation of predicted ontological annotations. Bioinformatics. 2013;29(13):53–61.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

53. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):158–69.
54. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci USA. 2006;103(8):2833–8.
55. Noble SM, French S, Kohn LA, Chen V, Johnson AD. Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. Nat Genet. 2010;42(7):590–8.
56. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Storms R, Boone C, Bussey H. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50(1):167–81.
57. Liu H, Kohler J, Fink GR. Suppression of hyphal formation in Candida albicans by mutation of a STE12 homolog. Science. 1994;266(5191):1723–6.
58. You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S. NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res. 2019;47(W1):379–87. https://doi.org/10.1093/nar/gkz388.
59. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):447–52. https://doi.org/10.1093/nar/gku1003.
60. Dessimoz C, Skunca N, Thomas PD. CAFA and the open world of protein function predictions. Trends Genet. 2013;29(11):609–10.
61. Jiang Y, Clark WT, Friedberg I, Radivojac P. The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. Bioinformatics. 2014;30(17):609–16.
62. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–83. https://doi.org/10.1093/bioinformatics/btg153. Accessed 1 Aug 2019.
63. Zhou N. Supplementary data. figshare. 2019. https://doi.org/10.6084/m9.figshare.8135393.v3. https://figshare.com/articles/Supplementary_data/8135393/3.
64. Jiang Y. CAFA2. Zenodo. 2019. https://doi.org/10.5281/zenodo.3403452.
65. Zhou N, Gerten M, Friedberg I. CAFA_assessment_tool. Zenodo. 2019. https://doi.org/10.5281/zenodo.3401694.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

W408–W413 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 2 May 2019*
doi: 10.1093/nar/gkz322

IntFOLD: an integrated web resource for high performance protein structure and function prediction

Liam J. McGuffin[1,*], Recep Adiyaman[1], Ali H.A. Maghrabi[1], Ahmad N. Shuid[1,2], Danielle A. Brackenridge[1], John O. Nealon[1] and Limcy S. Philomina[1]

[1]School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK and [2]Infectomics cluster, Advanced Medical and Dental Institute, University of Science, Malaysia, Bertam, 13200, Kepala Batas, Pulau Pinang, Malaysia

Received February 14, 2019; Revised April 05, 2019; Editorial Decision April 18, 2019; Accepted April 23, 2019

ABSTRACT

The IntFOLD server provides a unified resource for the automated prediction of: protein tertiary structures with built-in estimates of model accuracy (EMA), protein structural domain boundaries, natively unstructured or disordered regions in proteins, and protein–ligand interactions. The component methods have been independently evaluated via the successive blind CASP experiments and the continual CAMEO benchmarking project. The IntFOLD server has established its ranking as one of the best performing publicly available servers, based on independent official evaluation metrics. Here, we describe significant updates to the server back end, where we have focused on performance improvements in tertiary structure predictions, in terms of global 3D model quality and accuracy self-estimates (ASE), which we achieve using our newly improved ModFOLD7_rank algorithm. We also report on various upgrades to the front end including: a streamlined submission process, enhanced visualization of models, new confidence scores for ranking, and links for accessing all annotated model data. Furthermore, we now include an option for users to submit selected models for further refinement via convenient push buttons. The IntFOLD server is freely available at: http://www.reading.ac.uk/bioinf/IntFOLD/.

INTRODUCTION

Despite recent advances in the experimental methods for determining protein tertiary structures and their interactions, the sequence-to-structure gap has been relentlessly increasing. The gap in our knowledge of protein sequences versus known structures is being exacerbated by onset of ever cheaper and more efficient genome sequencing methods. At the time of writing, we now have close to two hundred million unique protein sequences in UniProt (1), but the number of protein structures in the Protein Data Bank (PDB) (2) remains <150 000. In order to realize the promise of next generation sequencing, it is clear that we must rely on computational tools for predicting structures and building 3D models of proteins directly from sequence so that we may close the knowledge gap. While the routine use of predicted 3D models by life scientists continues to grow, the protein structure prediction community has faced a number of challenges, which may have restricted the more wide spread acceptance of 3D protein models by non-experts (3). For example, until relatively recently we have not had methods that can confidently estimate the likely quality of 3D protein models, although these tools are now becoming increasingly accurate and more widely available (4).

The structure prediction community has made great advances over the past 20+ years with several major improvements in template based modelling (TBM), free modelling (FM) and estimates of 3D model accuracy (EMA) coming in the last few CASP (Critical Assessment of Structure Prediction) experiments (5–7). Successive versions of the IntFOLD server components have been independently benchmarked in the CASP experiments, from CASP9 to CASP13, and continually by the CAMEO project (8). Many of our own advances in performance over the years have come through improvements in our ModFOLD methods for EMA, and in particular our Accuracy Self Estimate (ASE) scoring for our 3D models (5,9).

Previous versions of the IntFOLD server were described in the Web Server issues of this journal in 2011 (10) and 2015 (11). Since its inception, the server has had ∼15,000 unique users and it has completed ∼200 000 predictions. The server's component methods have been applied in order to model protein structures and their interactions for a diverse range of specialisations accross the life sciences. For example, our tools have been used: to model novel proteins in the *Drosophila melanogaster* genome (12), to reveal new interactions and mechanisms for the regulation of mammalian GCKIII kinases (13,14), to explain the evolutionary

[*]To whom correspondence should be addressed. Tel: +44 118 378 6332; Fax: +44 118 378 8106; Email: l.j.mcguffin@reading.ac.uk

resurrection of flagellar motility in *Pseudomonas fluorescens* (15), to structurally and functionally annotate the proteome of barley powdery mildew (*Blumeria graminis* f. sp. *hordei*) (16), and to understand the effect of the missense mutation associated with dermatosparaxis (17).

In this paper, we describe the significant modifications to IntFOLD server and its component methods, which have led to successive performance gains since our last paper on the server from 2015. As well as reporting the major enhancements 'under the hood' to the server backend, we also report on the provision of new data outputs and user interface improvements.

MATERIALS AND METHODS

The IntFOLD server provides a single point of access to an integrated suite of six component methods: IntFOLD-TS, for tertiary structure prediction (9–11,18,19); ModFOLD, for 3D model Accuracy Self-Estimate (ASE) scoring (9,20); ReFOLD, for 3D model refinement (9,21); DISOclust, for disorder prediction (22,23); DomFOLD for structural domain prediction (10,11) and FunFOLD for ligand binding site prediction (24,25). These component methods have been independently evaluated in the various CASP (5,7,26–28) experiments over the years and are continually benchmarked by the CAMEO project (8) (also see results section). The major enhancement to the server methodology, since the last web server paper, has been to the underlying Tertiary Structure (TS) prediction algorithm. Since its inception, the high performance tertiary structure prediction algorithms with integrated model quality assessment have been at the core of IntFOLD server (10,11,18), and these factors have been key contributors to the historical success of the component methods (5,7,9,18,26–30). For version 5 of the IntFOLD server, the algorithms for both 3D model selection and ASE scoring have been upgraded via the integration of our new ModFOLD7_rank method.

The IntFOLD-TS method is the major component of the server and its output of high quality 3D models forms the basis for subsequent prediction algorithms. The IntFOLD5-TS method was newly developed for CASP13 and worked via iterative multi-template based modelling (19) using the target-template alignments from 14 alternative methods (SP3 (31), SPARKS2 (31), HHsearch (32), COMA (33), SPARKSX (34), CNFsearch (35) and the eight alternative threading methods that are integrated into the current LOMETS package (36)). The multiple target-template alignments for 3D modelling were then selected using ASE scoring via the ModFOLD7_rank method, with the aim of minimising local errors in final generated models. Additionally, the HHpred (37) method and the template free method I-TASSER light (38) (for sequence <500 residues; run in 'light mode' with wall-time restricted to 5h) contributed models for ranking. All of the final models were pooled and then scored and ranked using the ModFOLD7_rank method and presented to the user in descending order of global model quality. The ASE scores from ModFOLD7_rank were included in the temperature factor column of each of the PDB formatted model files. The integration of ASE scores in this way allows users to conveniently view the local model quality as temperature gradient that can be mapped onto their 3D models using their favourite molecular viewing software, for example PyMOL (http://www.pymol.org/).

The ModFOLD7_rank method is our latest update to Quality Assessment (QA) that combines the strengths of multiple pure-single and quasi-single model methods for improving prediction accuracy, building on the successful strategy that was used in ModFOLD6 (4,9,20). For the IntFOLD5 server our major emphasis was on increasing the performance of per-residue accuracy prediction for our own models, as well as improving our model ranking and score consistency for our models. Each IntFOLD5 model was considered individually using 6 pure-single model methods (CDA (20), SSA (20), ProQ2 (4), ProQ2D (39), ProQ3D (39) and VoroMQA (40)), and four alternative quasi-single model methods (DBA (20), MF5s (20), MFcQs (20) and ResQ (41)). For producing final local score outputs, Artificial Neural networks (NNs) were used to combine the component per-residue/local quality scores from each of the 10 alternative scoring methods, resulting in a final consensus of per-residue quality scores for each model. For producing the global score outputs, we made several variants that combined the mean global scores from the different methods and each were optimized for different aspects of the quality estimation problem. For the IntFOLD5 server, the accurate ranking of our models was the main objective, so for this reason we integrated the ModFOLD7_rank variant, which was optimized for ranking.

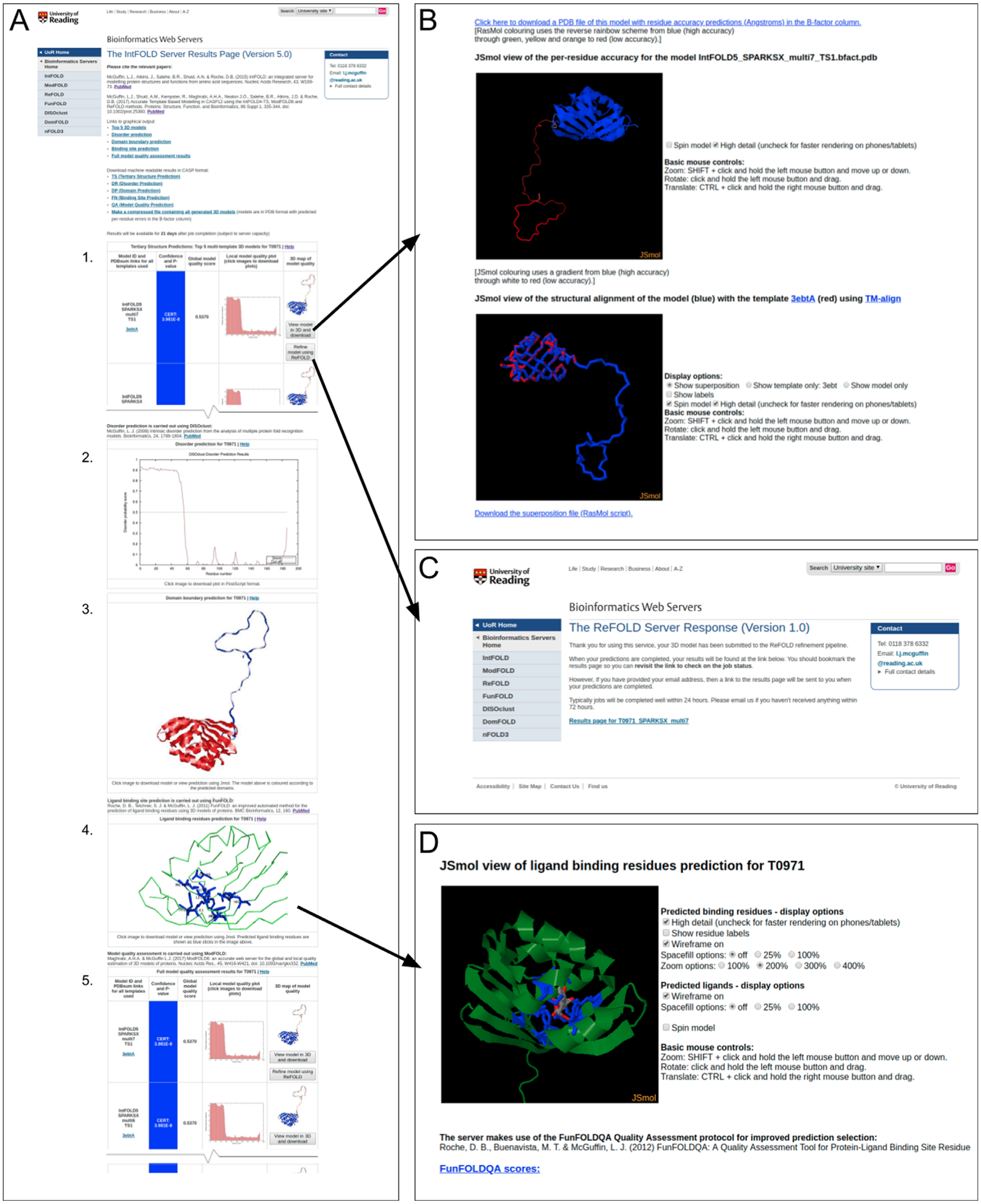
As well as improvements in performance to underlying algorithms, several new user interface upgrades were implemented. These included a streamlined submission form, recalibrated *P*-values for confidence scoring of model quality estimates, the ability to download compressed archives of all annotated models, and the ability to interact with models and then further refine them with a few clicks via simple push buttons. The server inputs and outputs are described in more detail below.

RESULTS AND DISCUSSION

Server inputs and outputs

Inputs. A single amino acid sequence for the protein chain is the only required input for the server. However, users also have the option to provide a short memorable name for their prediction job and an email address, which will only be used to provide a notification of the link to the results when the predictions are completed. If users do not wish to be notified via email, then they can bookmark the link to the results page for later viewing.

Graphical outputs. Examples of the graphical outputs from the IntFOLD5 server are shown in Figure 1. The graphical output is presented as a single table that graphically summarises all prediction data using thumbnail images of ASE plots and models, links to the template information and colour coded scoring (Figure 1A). It is always recommended to choose the model with the highest score or lowest *P*-value. The confidence rating relates to the *P*-value. For example, a 'CERT' rating relates to models where $P < 0.001$, i.e., less than a 1/1000 chance that the model is incorrect (see help pages for other ratings). So all 'CERT' mod-



Bioinformatics Web Servers

The IntFOLD Server Results Page (Version 5.0)

Please cite the relevant papers:

McGuffin, L.J., Adkins, J., Salehe, B.R., Shuid, A.N. & Roche, D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic Acids Research, 43, W169-73. PubMed

McGuffin, L.J., Shuid, A.M., Kempster, R., Maghrabi, A.H.A., Nealon J.O., Salehe, B.R., Atkins, J.D. & Roche, D.B. (2017) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. Proteins: Structure, Function, and Bioinformatics, 86 Suppl 1, 335-344, doi: 10.1002/prot.25360. PubMed

Tel: 0118 378 6332
Email: l.j.mcguffin@reading.ac.uk
Full contact details

Links to graphical output:
- Top 5 3D models
- Disorder prediction
- Domain boundary prediction
- Binding site prediction
- Full model quality assessment results

Download machine readable results in CASP format:
- TS (Tertiary Structure Prediction)
- DR (Disorder Prediction)
- DP (Domain Prediction)
- FN (Binding Site Prediction)
- QA (Model Quality Prediction)
- Make a compressed file containing all generated 3D models (models are in PDB format with predicted per-residue errors in the B-factor column)

Results will be available for **21 days** after job completion (subject to server capacity)

Tertiary Structure Predictions: Top 5 multi-template 3D models for T0971 | Help

Model ID and PDBsum links for all templates used	Confidence and P-value	Global model quality score	Local model quality plot (click images to download plots)	3D map of model quality
IntFOLD5 SPARKSX multi7 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download Refine model using ReFOLD
IntFOLD5 SPARKSX				

Disorder prediction is carried out using DISOclust:
McGuffin, L. J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics, 24, 1789-1804. PubMed

Disorder prediction for T0971 | Help

DISOclust Disorder Prediction Results
Disorder probability score
Residue number

Click image to download plot in PostScript format.

Domain boundary prediction for T0971 | Help

Click image to download model or view prediction using Jmol. The model above is coloured according to the predicted domains.

Ligand binding site prediction is carried out using FunFOLD:
Roche, D. B., Tetchner, S. J. & McGuffin, L. J. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics, 12, 160. PubMed

Ligand binding residues prediction for T0971 | Help

Click image to download model or view prediction using Jmol. Predicted ligand binding residues are shown as blue sticks in the image above.

Model quality assessment is carried out using ModFOLD:
Maghrabi, A.H.A. & McGuffin L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. Nucleic Acids Res., 45, W416-W421, doi: 10.1093/nar/gkx332. PubMed

Full model quality assessment results for T0971 | Help

Model ID and PDBsum links for all templates used	Confidence and P-value	Global model quality score	Local model quality plot (click images to download plots)	3D map of model quality
IntFOLD5 SPARKSX multi7 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download Refine model using ReFOLD
IntFOLD5 SPARKSX multi6 TS1 3ebtA	CERT: 3.981E-8	0.5370		View model in 3D and download

Click here to download a PDB file of this model with residue accuracy predictions (Angstroms) in the B-factor column.
[RasMol colouring uses the reverse rainbow scheme from blue (high accuracy) through green, yellow and orange to red (low accuracy).]

JSmol view of the per-residue accuracy for the model IntFOLD5_SPARKSX_multi7_TS1.bfact.pdb

□ Spin model ☑ High detail (uncheck for faster rendering on phones/tablets)

Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

[JSmol colouring uses a gradient from blue (high accuracy) through white to red (low accuracy).]

JSmol view of the structural alignment of the model (blue) with the template 3ebtA (red) using TM-align

Display options:
● Show superposition ○ Show template only: 3ebt ○ Show model only
□ Show labels
☑ Spin model ☑ High detail (uncheck for faster rendering on phones/tablets)
Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

Download the superposition file (RasMol script).

Bioinformatics Web Servers

The ReFOLD Server Response (Version 1.0)

Thank you for using this service, your 3D model has been submitted to the ReFOLD refinement pipeline.

When your predictions are completed, your results will be found at the link below. You should bookmark the results page so you can **revisit the link to check on the job status**.

However, if you have provided your email address, then a link to the results page will be sent to you when your predictions are completed.

Typically jobs will be completed well within 24 hours. Please email us if you haven't received anything within 72 hours.

Results page for T0971_SPARKSX_multi7

Tel: 0118 378 6332
Email: l.j.mcguffin@reading.ac.uk
Full contact details

Accessibility | Site Map | Contact Us | Find us © University of Reading

JSmol view of ligand binding residues prediction for T0971

Predicted binding residues - display options
☑ High detail (uncheck for faster rendering on phones/tablets)
□ Show residue labels
☑ Wireframe on
Spacefill options: ● off ○ 25% ○ 100%
Zoom options: ○ 100% ● 200% ○ 300% ○ 400%

Predicted ligands - display options
☑ Wireframe on
Spacefill options: ● off ○ 25% ○ 100%

□ Spin model

Basic mouse controls:
Zoom: SHIFT + click and hold the left mouse button and move up or down.
Rotate: click and hold the left mouse button and drag.
Translate: CTRL + click and hold the right mouse button and drag.

The server makes use of the FunFOLDQA Quality Assessment protocol for improved prediction selection:
Roche, D. B., Buenavista, M. T. & McGuffin, L. J. (2012) FunFOLDQA: A Quality Assessment Tool for Protein-Ligand Binding Site Residue

FunFOLDQA scores:

Figure 1. The IntFOLD5 server results pages for CASP13 target T0971. (**A**) Graphical output from the main results page showing (from top to bottom): 1. The table with the top 5 selected 3D models and scores (table truncated here to fit); 2. The prediction of natively unstructured/disordered regions; 3. The predicted structural domain boundaries; 4. The ligand binding site prediction; 5. The full model quality rankings for all generated models (table truncated here to fit). The arrows point to additional pages that are linked to when users click on images/buttons on the main page. (**B**) Clicking the button titled 'View model in 3D and download' leads to dynamically generated pages showing interactive views of the model, and structural superpositions of the model with relevant template/s, which can be manipulated in 3D using the JSmol/HTML5 framework (http://www.jmol.org/) and/or downloaded for local viewing. (**C**) Clicking the button titled 'Refine model using ReFOLD' submits the 3D model to the ReFOLD service (21) for refinement guided by accurate quality estimates. (**D**) Clicking on the image of the ligand binding site prediction links to a dynamically generated page that provides numerous options for interactively viewing the likely protein–ligand interactions in 3D with JSmol.

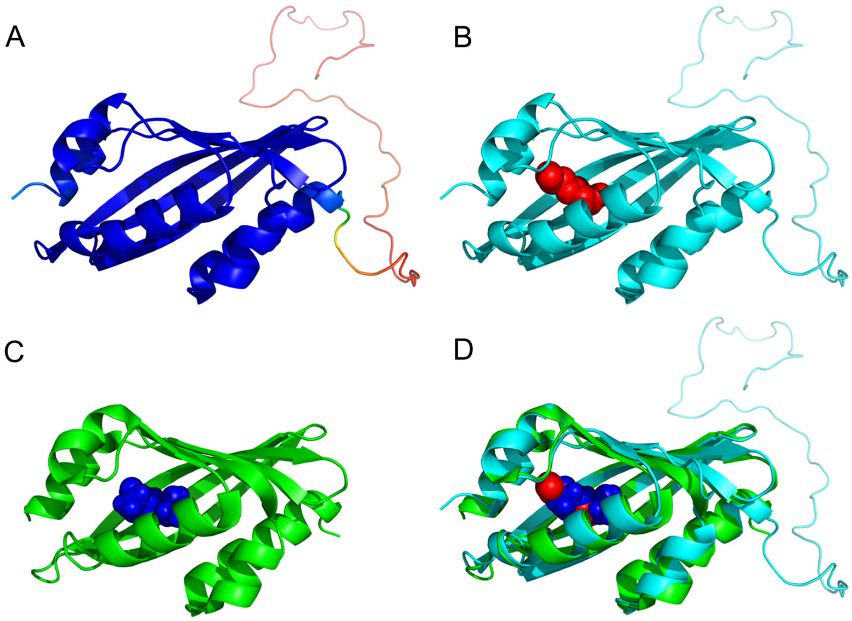


Figure 2. The IntFOLD5 server predictions for CASP13 target T0971 – comparison of models with the native crystal structure (PDB ID: 6d34). All images were rendered using PyMOL (http://www.pymol.org/). (**A**) The IntFOLD5 3D model coloured by accuracy self-estimate of local quality using the temperature coloured scheme from blue (indicating residues in the model predicted to be close to the native structure) to red (indicating residues in the model that are far from the native or unstructured). (**B**) The IntFOLD5 3D model with the main cluster of predicted ligands (red spheres) indicating the predicted location of binding site. (**C**) The crystal structure of T0971/6d34 with ligand (blue spheres). Note: the disordered domain predicted in the model is absent in the X-ray data. (**D**) Superposition of the IntFOLD5 model and the native structure.

els are highly likely to have the correct fold. However, the models with the lowest *P*-values are more likely to have the highest backbone accuracy and overall quality. Several new user interface options are available. Users have the option to download coordinates and view the detailed IntFOLD5-TS tertiary structure prediction results interactively in 3D (Figure 1B) and submit individual 3D models for further refinement using ReFOLD (Figure 1C) via simple push buttons. Downloadable coordinates and interactive 3D views of the protein ligand interactions can also be accessed via the FunFOLD results summary image (Figure 1D). In addition, clicking on the DISOclust disorder prediction profile images and the thumbnail images of the ASE score profiles from ModFOLD7_rank will allow users to view and/or download higher quality versions of the plots.

Figure 2 shows a comparison of the example models for CASP13 target T0971 (obtained via the pages shown in Figure 1) and the native structure (PDB ID 6d34). The 3D model of the protein (Figure 2A and B) is close to the native structure shown in Figure 2C. The predicted location of the ligand binding site is shown to be accurate (Figure 2B) and there is a close superposition of the model and native structure (Figure 2D), with a GDT_TS score of 95%. The ASE for the model, indicated by the colouring in Figure 2A, and the identification of the unstructured domain are also shown to be accurately predicted.

Machine readable outputs. All of the raw data files for the predictions are available to download via links on the results pages. The file formats comply with the CASP and/or CAMEO data standards. An additional new feature is the provision of a link that allows users to download all of the ASE annotated models in PDB format (with the error estimates, in Angstroms, in place of temperature factor data) as a zipped archive.

Independent benchmarking

Each major version of the server has been independently tested in each of the relevant categories of the CASP experiments (from CASP9 to CASP13, http://predictioncenter.org) and the performance has been competitive (9,18). Recently, the component methods have ranked among top independent servers in the Tertiary Structure (TS) prediction (5) and Estimates of Model Accuracy (EMA) categories (7), as well as ranking well in the historical categories of intrinsic disorder prediction and function prediction (26,27). The DISOclust method was designed to add a significant performance boost to DISOPRED (22), and the latest version of DISOPRED is integrated with the IntFOLD server. Additionally, the IntFOLD5 server components (IntFOLD, ModFOLD and FunFOLD) have been continuously benchmarked using the CAMEO resource (8) and they have been shown to be competitive in each respective category (see results from the 3D, QE and LB categories at https://www.cameo3d.org/). Furthermore, the GO term outputs from the FunFOLD component of the server have been benchmarked during the most recent CAFA experiment (https://www.biofunctionprediction.org/cafa/, paper in preparation).

CAMEO results summary. The TS predictions from the IntFOLD5 server are continuously evaluated by the CAMEO project (8). The IntFOLD versions have consistently ranked among the top few public servers accord-

Table 1. Independent benchmarking of tertiary structure predictions with CAMEO 3D data. Performance results for 3 months of data (26 October 2018 to 19 January 2019) are shown for *all* (250) targets and *all* (17) public methods. Data are sorted by average lDDT score for all targets. The scores for the IntFOLD-TS methods are indicated in bold. Data are taken from the CAMEO 3D front page http://www.cameo3d.org/ on 19 January 2019.

	Average lDDT		Average lDDT-BS	
Server name	All targets	Modelled targets	All targets	Modelled targets
IntFOLD5-TS	**68.04**	**68.04**	**70.94**	**70.94**
RaptorX	67.38	67.38	68.45	68.45
Robetta	65.51	69.1	63.24	66.11
HHpredB	64.06	64.06	68.59	68.59
SWISS-MODEL	62.22	62.97	64.85	65.56
IntFOLD4-TS	**55.02**	**68.1**	**58.12**	**73.25**
SPARKS-X	54.63	60.7	58.07	66.78
M4T-SMOTIF-TF	54.45	60.77	62.92	65.78
IntFOLD3-TS	**53.75**	**66.85**	**55.76**	**69.33**
PRIMO	51.74	57.48	58.32	64.65
PRIMO_BST_CL	51.71	57.45	58.32	64.65
NaiveBLAST	50.34	55.69	60.08	62.11
PRIMO_BST_3D	49.83	55.86	57.99	63.51
PRIMO_HHS_3D	48.27	55.87	56.49	62.62
PRIMO_HHS_CL	46.73	56.43	55.55	61.58
Princeton_TEMPLATE	24.46	54.61	25.63	58.95
Phyre2	24.06	52.77	29.27	67.31

Table 2. Independent benchmarking of IntFOLD versions with CAMEO 3D data showing the sequential improvement in server performance since the last webserver paper describing IntFOLD3. Performance results for 1 year of data (26 January 2018 to 19 January 2019) are shown for a common subset of 581 targets. The reference method is IntFOLD5-TS and the table is sorted by average lDDT. Data are downloaded from http://www.cameo3d.org/

	Avg. lDDT		Avg. CAD-score		Avg. lDDT-BS	
Server Name	Dif.	Ref.	Dif.	Ref.	Dif.	Ref.
IntFOLD5-TS	0	67.72	0	0.67	0	71.86
IntFOLD4-TS	0.53	67.18	0	0.66	0.23	71.62
IntFOLD3-TS	2.11	65.61	0.02	0.65	1.9	69.96

ing to lDDT_BS scores and lDDT scores. At the time of writing, IntFOLD5-TS ranks as the top publicly available method based on the last 3-month data for all targets (Table 1). Based on pairwise comparisons using a common subset of targets over the last year, IntFOLD5-TS ranks as the second best 3D server according to the CAMEO lDDT scores (Supplementary Tables S1 and S2). Moreover, the IntFOLD5-TS version of the method has been independently verified to be an improvement over our two previous methods (IntFOLD3-TS and IntFOLD4-TS) (Table 2).

CASP12 and 13 results summary. In the last few CASP experiments since the last webserver publication, the IntFOLD server has performed well at Template Based Modelling (TBM), ranking as high as third place and outperforming other servers in terms of Accuracy Self Estimates (ASE) (5). The IntFOLD4 and IntFOLD5 server performance rankings, for CASP12 and CASP13 targets respectively, are shown in Supplementary Tables S3–S6. The IntFOLD server methods have also been key to our group's success at CASP12 and 13 allowing us to rank as high as second place on the 'all group' TBM + TBM/FM domains. The McGuffin group performance is summarized in Supplementary Tables S7 and S8.

CONCLUSIONS

The IntFOLD server provides free access to an integrated set of high performance, fully automated tools for structure and function prediction of proteins from their amino acid sequences. The component methods of the server are continually benchmarked via the CAMEO project and they have been rigorously blind tested at recent CASP experiments. The IntFOLD methods have been independently verified to rank among the top performing servers in many prediction categories. Results from the IntFOLD server are presented to non-expert users in an intuitive manner with graphical output providing a visual summary of a complex set of data. More detailed results for individual predictions can be interactively viewed and the raw, machine readable data can be accessed in standard data formats.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Saudi Arabian Government (to A.H.A.M.); Republic of Turkey Ministry of National Education (to R.A.); Malaysian Government (to A.N.S). Funding for open access charge: Republic of Turkey Ministry of National Education; Saudi Arabian Embassy.

Conflict of interest statement. None declared.

REFERENCES

1. UniProt, C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Schwede,T. (2013) Protein modeling: what happened to the "protein structure gap"? *Structure*, **21**, 1531–1540.
4. Elofsson,A., Joo,K., Keasar,C., Lee,J., Maghrabi,A.H.A., Manavalan,B., McGuffin,L.J., Menendez Hurtado,D., Mirabello,C., Pilstal,R. *et al.* (2018) Methods for estimation of model accuracy in CASP12. *Proteins*, **86**(Suppl. 1), 361–373.
5. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Moult,J., Schwede,T. and Tramontano,A. (2018) Evaluation of the template-based modeling in CASP12. *Proteins*, **86**(Suppl. 1), 321–334.
6. Abriata,L.A., Tamo,G.E., Monastyrskyy,B., Kryshtafovych,A. and Dal Peraro,M. (2018) Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, **86**(Suppl. 1), 97–112.
7. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Schwede,T. and Tramontano,A. (2018) Assessment of model accuracy estimations in CASP12. *Proteins*, **86**(Suppl. 1), 345–360.
8. Haas,J., Barbato,A., Behringer,D., Studer,G., Roth,S., Bertoni,M., Mostaguir,K., Gumienny,R. and Schwede,T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, **86**(Suppl. 1), 387–398.
9. McGuffin,L.J., Shuid,A.N., Kempster,R., Maghrabi,A.H.A., Nealon,J.O., Salehe,B.R., Atkins,J.D. and Roche,D.B. (2018) Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. *Proteins*, **86**(Suppl. 1), 335–344.
10. Roche,D.B., Buenavista,M.T., Tetchner,S.J. and McGuffin,L.J. (2011) The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res.*, **39**, W171–W176.
11. McGuffin,L.J., Atkins,J.D., Salehe,B.R., Shuid,A.N. and Roche,D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.*, **43**, W169–W173.
12. Dunwell,T.L., McGuffin,L.J., Dunwell,J.M. and Pfeifer,G.P. (2013) The mysterious presence of a 5-methylcytosine oxidase in the Drosophila genome: possible explanations. *Cell Cycle*, **12**, 3357–3365.
13. Fuller,S.J., McGuffin,L.J., Marshall,A.K., Giraldo,A., Pikkarainen,S., Clerk,A. and Sugden,P.H. (2012) A novel non-canonical mechanism of regulation of MST3 (mammalian Sterile20-related kinase 3). *Biochem. J.*, **442**, 595–610.
14. Sugden,P.H., McGuffin,L.J. and Clerk,A. (2013) SOcK, MiSTs, MASK and STicKs: the GCKIII (germinal centre kinase III) kinases and their heterologous protein-protein interactions. *Biochem. J.*, **454**, 13–30.
15. Taylor,T.B., Mulley,G., Dills,A.H., Alsohim,A.S., McGuffin,L.J., Studholme,D.J., Silby,M.W., Brockhurst,M.A., Johnson,L.J. and Jackson,R.W. (2015) Evolution. Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system. *Science*, **347**, 1014–1017.
16. Bindschedler,L.V., McGuffin,L.J., Burgis,T.A., Spanu,P.D. and Cramer,R. (2011) Proteogenomics and in silico structural and functional annotation of the barley powdery mildew Blumeria graminis f. sp. hordei. *Methods*, **54**, 432–441.
17. Monteagudo,L.V., Ferrer,L.M., Catalan-Insa,E., Savva,D., McGuffin,L.J. and Tejedor,M.T. (2015) In silico identification and three-dimensional modelling of the missense mutation in ADAMTS2 in a sheep flock with dermatosparaxis. *Vet. Dermatol.*, **26**, 49–52.
18. McGuffin,L.J. and Roche,D.B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins*, **79**(Suppl. 10), 137–146.
19. Buenavista,M.T., Roche,D.B. and McGuffin,L.J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*, **28**, 1851–1857.
20. Maghrabi,A.H.A. and McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Res.*, **45**, W416–W421.
21. Shuid,A.N., Kempster,R. and McGuffin,L.J. (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. *Nucleic. Acids. Res.*, **45**, W422–W428.
22. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **24**, 1798–1804.
23. Atkins,J.D., Boateng,S.Y., Sorensen,T. and McGuffin,L.J. (2015) Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. *Int. J. Mol. Sci.*, **16**, 19040–19054.
24. Roche,D.B., Tetchner,S.J. and McGuffin,L.J. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.
25. Roche,D.B., Buenavista,M.T. and McGuffin,L.J. (2013) The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Res.*, **41**, W303–W307.
26. Noivirt-Brik,O., Prilusky,J. and Sussman,J.L. (2009) Assessment of disorder predictions in CASP8. *Proteins*, **77**(Suppl. 9), 210–216.
27. Schmidt,T., Haas,J., Gallo Cassarino,T. and Schwede,T. (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79**(Suppl. 10), 126–136.
28. Kryshtafovych,A., Barbato,A., Fidelis,K., Monastyrskyy,B., Schwede,T. and Tramontano,A. (2014) Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins*, **82**(Suppl. 2), 112–126.
29. McGuffin,L.J. (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. *Proteins*, **77**(Suppl. 9), 185–190.
30. Kryshtafovych,A., Barbato,A., Monastyrskyy,B., Fidelis,K., Schwede,T. and Tramontano,A. (2016) Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. *Proteins*, **84**(Suppl. 1), 349–369.
31. Zhou,H. and Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*, **61**(Suppl. 7), 152–156.
32. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
33. Margelevicius,M. and Venclovas,C. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics*, **11**, 89.
34. Yang,Y., Faraggi,E., Zhao,H. and Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.
35. Ma,J., Wang,S., Zhao,F. and Xu,J. (2013) Protein threading using context-specific alignment potential. *Bioinformatics*, **29**, i257–i265.
36. Wu,S. and Zhang,Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.*, **35**, 3375–3382.
37. Meier,A. and Soding,J. (2015) Automatic prediction of protein 3D structures by probabilistic Multi-template homology modeling. *PLoS Comput. Biol.*, **11**, e1004343.
38. Roy,A., Kucukural,A. and Zhang,Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.
39. Uziela,K., Menendez Hurtado,D., Shu,N., Wallner,B. and Elofsson,A. (2017) ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, **33**, 1578–1580.
40. Olechnovic,K. and Venclovas,C. (2017) VoroMQA: assessment of protein structure quality using interatomic contact areas. *Proteins*, **85**, 1131–1145.
41. Yang,J., Wang,Y. and Zhang,Y. (2016) ResQ: an approach to unified estimation of B-Factor and Residue-Specific error in protein structure prediction. *J. Mol. Biol.*, **428**, 693–701.

BOOK CHAPTER

Proteins and their interacting partners: An introduction to protein-ligand binding site prediction methods with a focus on FunFOLD3

Danielle Allison Brackenridge[1] and Liam James McGuffin[2]

[1]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK.
d.a.brackenridge@pgr.reading.ac.uk

[2]School of Biological Sciences, University of Reading, Reading RG6 6AS, UK.
l.j.mcguffin@reading.ac.uk

Abstract

Proteins are essential molecules with a diverse range of functions; elucidating their biological and biochemical roles from their interacting partners can be difficult and time consuming using *in vitro* and/or *in vivo* methods. Additionally, *in vivo* protein-ligand binding site elucidation is unable to keep pace with current growth in sequencing, leaving the majority of new sequences without known functions. Therefore, the development of new methods, which aim to predict the protein-ligand interactions and ligand-binding site residues directly from amino acid sequences, is becoming increasingly important. *In silico* prediction can utilise either sequence information, structural information or a combination of both. In this chapter, we will discuss the broad range of methods for ligand-binding site prediction from protein structure and we will describe our method FunFOLD3, for the prediction of protein-ligand interactions and ligand-binding sites based on template-based modelling. Additionally we will describe the step-by-step instructions on using the FunFOLD3 downloadable application, along with examples from the Critical Assessment of Techniques for Protein Structure Prediction (CASP) where FunFOLD3 has been used to aid ligand and ligand-binding site prediction. Finally, we will introduce our newer method, FunFOLD3-D, a version of FunFOLD3 which will aim to improve template based protein-ligand binding site prediction through the integration of docking, using AutoDock Vina.

Key words protein-ligand interactions, ligand-binding site prediction, Critical Assessment of Techniques for Protein Structure Prediction (CASP), protein structure prediction, template-based modelling, *in silico* prediction, FunFOLD3, docking

Introduction

Proteins are essential molecules involved in a wide variety of essential intra- and inter-cellular activities. The particular activities include, but are not limited to; maintaining cellular defences; enzymatic catalysis; metabolism and catabolism; maintenance of the structural integrity of cells and signalling within and between cells. Hence, studying protein-ligand interactions is an important step in the functional elucidation of proteins involved in these cellular processes.[1–4]

In silico methods are used to address the problem with the sequence-structure-function knowledge gaps. Bioinformatics approaches that utilise information from existing protein-ligand complexes are becoming increasingly important, because ligands that bind to a protein are pivotal to understanding protein function. In general, function prediction methods can be split into two broad categories: the sequence-based and structure-based alignment .[5]

Sequence comparison is used to infer homology and collect evidence about membership in a given family. The key requirement is to properly choose similarity measures and related cut-off values in order to avoid false positives and false negatives. When new sequences diverge with low homology (<30%) to those within known databases, then finding functionally annotated homologs becomes less likely. Sequence alignment relies on the evolutionarily related segments of two proteins, which could consist of binding sites and domains.[6] The main strength of sequence-based approaches for prediction of binding sites is that methods that utilise this approach have the ability to determine ligand-binding motif in proteins that may not have the same overall fold.[7] Homology-based methods require related proteins with significant identity to the query protein to be available in the PDB because the conservation of biochemical function drops rapidly for proteins sharing <35-40% sequence identity.[7] Therefore, a limitation of this approach is that methods do not work for remote homologs (<30% pairwise identity). For sequence-based methods, the homologous sequence of the target sequence is required, and a multiple sequence alignment (MSA) is constructed. Then, using the specific approach conserved residues are identified among all the sites in the MSA. The selection of homologous sequences to a query protein is a critical step in methods based on both sequence and structure approaches for prediction of protein functional site.[7]

Structure-based methods require the 3D dimensional structures of proteins, often also relying on available structural templates from distantly related proteins.[8] There are two principal methods to resolve structures experimentally; x-ray crystallography and nuclear magnetic resonance (NMR) with the former being the preferred structural tool.[8] Additionally, tertiary structures may be modelled directly from sequences using prediction servers.[9] Based on the observation of existing protein-ligand binding sites/complexes, it is evident that homologous proteins with similar global topology will often bind similar ligands and there will be conserved residues.[7] As a result, there are methods utilising both geometric match and evolutionary information to identify binding sites.[7] In general, these methods are broadly classified into geometry-based approaches and energetic based approaches. Geometry-based approaches identify binding residues by searching for pockets or cavities in a protein structure whereas, energetic-based approaches identify binding residues by using various derived interaction energies.[7] Whilst purely sequence-based and structure-based methods have different approaches, there are many methods which are based on a combination of both.[10]

Considerations when employing structure-based methods for prediction of protein–ligand binding sites have a number of limitations, including the following: 1. If a 3D model or experimental structure cannot be obtained, then it is not possible to make a prediction; in such cases the solution is to rely on purely sequence-based methods. 2. If templates with the same fold as the target protein that contain biologically relevant ligands cannot be detected, then it is not possible to make a prediction. 3. Most prediction servers, such as COACH[11] and FunFOLD[3,4,12], utilise in-house structure prediction pipelines to construct models for protein–ligand interaction predictions that may not always produce the best quality model for every target, which may result in over- and under-predicted protein–ligand binding sites. Nevertheless, despite these shortcomings, prediction methods are constantly under development and improvements can be gauged via the rigorous independent blind assessment scoring that is employed in competitive community-wide experiments and will be discussed later in the chapter.

Table 1. Availability of existing tools for ligand binding site prediction from protein structure introduced since 2009

Figure adapted from[10]

Method	Year	Type
SiteMap[13]	2009	Geometric
Fpocket[14]	2009	Geometric
SiteHound[15]	2009	Energetic
ConCavity[16]	2009	Conservation
3DLigandSite[17]	2010	Template
POCASA[18]	2010	Geometric
DoGSite[19]	2010	Consensus
FunFOLD[4]	2011	Template
MetaPocket 2.0[20]	2011	Consensus
MSPocket[21]	2011	Geometric
FTSite[22]	2012	Energetic
LISE[23]	2012	Knowledge/conservation
COFACTOR[24]	2012	Template
COACH[11]	2013	Template
G-LoSA[25]	2013	Template
eFindSite[26]	2013	Template
GalaxySite[27]	2014	Template/Docking

LIBRA[28]	2015	Template
P2RANK[10]	2015	Machine learning
bSiteFinder[29]	2016	Template
ISMBLabLIG[30]	2016	Machine learning
DeepSite[31]	2017	Machine learning

The natural step after determination of protein structure, is the prediction of ligand-binding sites. An important consideration when investigating protein-ligand interactions is whether any predicted ligands are biologically relevant. The most direct way to investigate the biological relevance of a ligand is by manual verifications.[32] These verifications can consist of reading literature, however given the growth of protein sequences and structures, such manual checking can be time consuming to carry out for numerous targets. Additionally, novel proteins may not have adequate amounts of literature available to deduce the biologically relevant ligands. As a result, there has been a need to develop automatic procedures to select biologically relevant ligands based on proteins available in PDB.[32] These consist of; FireDB[33], LigASite (LIGand Attachment SITE)[34], Binding MOAD (Mother of All Databases)[35], PDBbind[36], BindingDB[37] and BioLiP[32]. For the Critical Assessment of protein Structure Prediction competitions, biologically relevant ligands were defined using information from the literature, Swiss-Prot ligand annotations,[38] sequence conservation of functionally important residues and information from homologous structures.[39]

FunFOLD3

FunFOLD3 is a template-based method for protein-ligand binding site prediction[40] and it uses an automatic approach for cluster identification and residue selection.[4] The main requirement for FunFOLD3 is a 3D model and a list of templates as inputs.[4] FunFOLD3 will provide:

1. A list of residues in the target sequence that are most likely to bind a ligand
2. A list of putative binding ligand(s)
3. 3D models of the likely protein-ligand interactions
4. Lists of likely GO terms and EC identifiers

The FunFOLD3 method for predicting ligand-binding site residues is based on the concept that, target proteins with similar structures to known structures within the PDB, may also contain similar ligand binding sites.[3] The FunFOLD3 method predicts protein-ligand binding sites from a single sequence using predicted 3D structures (for example from the IntFOLD server), and lists of identified PDB structure templates.[40]

The FunFOLD3 algorithm utilises the TM-align[41] method to superpose templates containing biologically relevant ligands with the predicted 3D structures.[3] This method is a similar concept to methods from the Lee group[42] and Sternberg group.[17] However, the FunFOLD3 algorithm uses a novel automated method for ligand clustering and identification of binding residues.[4] The method also integrates protein-ligand binding site and quality assessment protocols for the prediction of protein function from sequence via structure.[3]

The input to the FunFOLD3 method is a 3D model of the protein under analysis and a list of template PDB IDs.[4] Once the 3D model formatting has been checked, the TM-align method is used to superimpose the template structures of the 3D protein model. Template-model superpositions with a TM-score ≥0.4 are retained.[40] This is because, TM-scores from 0.4 to 0.6 have previously been shown to mark the transition from unrelated to significantly related folds.[43] Then superpositions are combined and reoriented using a PyMOL script to determine putative ligands.[3] The next step is that ligands are assigned to clusters using a agglomerative hierarchical clustering algorithm. To determine the ligand binding site residues in the selected binding pocket, a novel residue-voting algorithm is used. Residues are determined to be in contact with a ligand cluster, if the residue is in contact with the ligand cluster.[3] Ligands are considered to be part of a cluster if the Van der Waals radius is ≤0.5 Å. Other authors also support using this cut-off.[5] The most probable ligand-binding site is the site with the largest ligand cluster.[40]

Blind evaluation of methods: CASP, CAMEO and CAFA

There are several community wide prediction experiments such as; Critical Assessment of techniques for protein Structure Prediction (CASP),[44] the Continuous Automated Model EvaluatiOn (CAMEO) project[45] and the Critical Assessment of Function Annotation (CAFA).[46] All of the listed methods can be used to analyse the performance of FunFOLD3. CASP is perhaps the premier community wide experiment for blind testing of modelling methods and is used to assess progress in capabilities.[47] Contributors are provided with amino acid sequences of unknown structures and are asked to deposit structure models.[47] The deposited models are then compared with newly determined experimental structures. [47] On of the scoring methods utilised in the CASP experiments for evaluating ligand binding site predicting is the Matthew Correlation Coefficient (MCC score). The MCC score provides a statistical score for the comparison of predicted ligand binding sites to observed ligand binding site residues. Residues will be assigned to one of the following; true positives, false positives, true negatives and false negatives. This provides a score of between -1 and 1, with 1 being a perfect positive prediction, whereas 0 is a random prediction. The calculation for MCC is given below:[44]

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

Equation 1. Matthews correlation coefficient

The equation above illustrates the calculation of the MCC, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives

The main disadvantage of the MCC score is that it is a purely a statistical measure and does not consider the overall tertiary structure of the protein.[48] In order to address this, a new scoring metric, the BDT score was developed and tested.[40] The BDT score takes into consideration the distance in 3D space that a predicted binding site residue is from the observed binding site residue. Similar to the MCC score, the BDT score of 1 indicates a perfect prediction while a score of 0 indicates a random prediction, but the score ranges from 0-1. The higher the BDT score, the closer the predicted site is to the observed site.

$$\mathrm{BDT}=\frac{\sum_{i=1}^{N_p}\max(S_{ij})}{\max(N_p, N_o)}$$

Equation 2. Binding site distance test score

The equation above illustrates the calculation of BDT, where *Sij* was the *S*-score between a predicted residue *i* and an observed residue *j*, *dij* was the Euclidean distance between the C-alpha coordinates of residues *i* and *j* and *d*0 was a distance threshold (values between 1 and 3 Å are recommended). The maximum *Sij* score, max(*Sij*), was then determined for each predicted residue. The final BDT score is the sum of the maximum *Sij* scores normalized by the greater value of the number of predicted residues (*Np*) and the number of observed residues (*No*)

The BDT score has been used in CASP experiments since CASP9[40] and the utilisation of BDT in CASP9 by the McGuffin group was cited in the publication about ligand-binding residues.[49] The BDT score was applied to predictions for the top ranked groups and no significant deviations to the MCC-based assessment were observed, supporting BDT as a robust alternative for evaluating the prediction of protein-ligand binding sites.

Methods

Instructions for installing and running theFunFOLD3 program have been described previously[12]. A downloadable version of the FunFOLD3 method is available as an executable JAR file, which can be run locally.

The system requirements are as follows:

1. A linux-based operating system such as Ubuntu

2. A recent version of Java (www.java.com/getjava/)

3. A recent version of PyMOL (www.pymol.org)

4. The TM-align program[41] (http://zhanglab.ccmb.med.umich.edu/TM-align/). The TMalign file is made executable: chmod +x TMalign.

5. wget and ImageMagick installed system wide.

6. The CIF chemical components database file[50] should be downloaded from here: ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif

7. The BioLip databases[51] containing ligand and receptor PDB files are also required (up to 30 GB of disc space may be required). The databases need to be downloaded in two sections: firstly all annotations prior to 6/3/2013 can be downloaded from here for the receptor database:

http://zhanglab.ccmb.med.umich.edu/BioLiP/download/receptor_2013-03-6.tar.bz2 (3.6 G) and from here for the ligand database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/ligand_2013-03-6.tar.bz2 (438 M). The text file of the BioLip annotations can be downloaded from here: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2. To update the databases to include annotations after 2013-03-6 it is recommended to download and use this perl script which will update the databases: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/download_all_sets.pl. The BioLip text file: http://zhanglab.ccmb.med.umich.edu/ BioLiP/download/BioLiP.tar.bz2 and all the weekly update text files should be concatenated to form a large text file containing all of the annotations. Furthermore, it is recommended to regularly update your BioLip and CIF databases. Additionally, a shell script is available as downloadBioLipdata.sh, which can be downloaded from here: http://www.reading.ac.uk/bioinf/downloads/, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

8. Please ensure your system environment is set to English, as utilising other languages may cause problems with the FunFOLD calculations: export LC_ALL=en_US.utf-8.

9. To run the program you can simply edit the shell script (FunFOLD3.sh)

10. For example, if the path of your model was “/home/dani/bin/FunFOLD3/MUProt_TS3”, your list of templates was “/home/dani/bin/FunFOLD3/T0470_PARENTNew.dat” (all templates should be listed on a single line separated by a space), your FASTA sequence file was “/home/dani/bin/FunFOLD3/T0470.fasta”, your output directory was “/home/dani/bin/FunFOLD3/” and your target was called T0470:

```
$JAVA_HOME/java -jar FunFOLD3.jar /home/dani/bin/FunFOLD3/MUProt_TS3 T0470
/home/dani/bin/FunFOLD3/ /home/dani/bin/FunFOLD3/T0470_PARENTNew.dat
/home/dani/bin/FunFOLD3/T0470.fasta $BIOLIP_TXT $BIOLIP_LIGAND $BIOLIP_RECEPTOR
$CIF
```

Or, using the shell script provided:

```
./FunFOLD3.sh /home/dani/bin/FunFOLD3/MUProt_TS3 T0470 /home/dani/bin/FunFOLD3/
/home/dani/bin/FunFOLD3/T0470_PARENTNew.dat /home/dani/bin/
FunFOLD3/T0470.fasta5
```

11.Basically, the user requires a model generated for their target protein, this can be achieved using a homology modeling method either in-house or via a web server such as IntFOLD (see Note 3). Additionally, the user needs a list of structurally similar templates. Again this list of templates can be generated from the list of templates used to generate the target protein model. The program utilizes the templates that have the same fold and contain biologically relevant ligands in the prediction process. Furthermore, it is important to download and install the BioLip databases[51] and CIF chemical components library file[50]. Additionally, it is important that the full paths for all input files are used, the output directory should also end with a "/" and must contain the input model, template list, and FASTA sequence file. A shell script is available called downloadBioLipdata.sh, which can be used to download and update the BioLip and CIF libraries. The shell script and the required perl script can be found on the downloads page, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

13. A number of output files are produced in the output directory (e.g. "/home/dani/bin/FunFOLD3/") and a log of the prediction process is output to screen as standard output. A description of the output files are as follows:

(a) The final ligand binding site prediction file "T0470_FN.txt" is supplied, conforming to CASP FN format. This file contains a list of predicted binding site residues, ligands, along with associated EC and GO terms.

(b) The final binding site prediction file "T0470_FN2_CAMEO-LB.txt" is additionally supplied in CAMEO-LB format. This file contains the predicted propensity that each ligand type is in contact with the predicted binding site residues.

(c) A PDB file “T0470_lig.pdb”, which contains superpositions of all templates, having the same fold and containing biologically relevant ligands, onto the model is produced.

(d) A reduced version of the PDB file “T0470_lig2.pdb”, which contains only the target model with all possible ligands is also produced.

(e) Another reduced version of the PDB file “T0470_lig3.pdb”, which contains only the target model with the predicted centroid ligand, is additionally output.

(f) A graphical representation of the protein–ligand interaction prediction “T0470_binding_site.png” is automatically generated using PyMOL.

(g) Finally, the PyMOL script “pymol.script” that was used to generate the image file is also output.

8. An example of output produced by FunFOLD3 for target T0470 can be found in the compressed directory: “T0470_Results.tar.gz” along with an example of the required input: “T0470_Input.tar.gz”. These example directories can be found on the downloads page: http://www.reading.ac.uk/bioinf/downloads/

FunFOLD3 for Function Predictions in CASP11, CASP12 and CASP13

FunFOLD3 has been used for prediction of protein ligand binding sites for all CASP targets since CASP11. Function prediction and modelling of protein ligand binding sites remains an important part of our prediction pipelines to aid with our manual evaluation of models during each CASP prediction season. The information from our FunFOLD3 method (regarding the function and locations of putative bound ligands) along with visual inspection was used for some targets in order to manually filter our modelled complexes prior to submission of our final models. The next CASP competition will start in 2020 with CASP14, and we will utilise FunFOLD3 with AutoDock Vina to improve ligand-binding site predictions for target proteins.

Following the release of the experimental data, all of the CASP11, CASP12 and CASP13 targets were analysed using the BioLip database to determine if their resolved structures contained biologically relevant ligands. Once targets with biologically relevant ligands were determined then the ligand-binding site residues were identified using Van der Waal radius of the contacting atom of a residue and the contacting ligand atom plus 0.5 Å. This resulted in a total of nine targets with PDB IDs containing biologically relevant ligands and binding site residues associated for CASP11, three for CASP12 and seven for CASP13. There can be a disparity between the number of predictions that were made and the availability of observed structures for analysis of these predictions, due to structures being cancelled by organisers or PDB IDs not being released for many CASP targets.

Figure 1A shows the predicted ligand-binding site for histidinol-phosphate aminotransferase (HisC) from *Sinorhizobium meliloti* (CASP11 ID T0819 and PDB ID 4wbt), with correctly predicted ligand-binding site residues in blue and incorrect predictions, defined as under and/or over-predictions in red. The pyridoxal-5'-phosphate (PLP) ligand is coloured yellow. This prediction resulted in a MCC score of 0.877 and a BDT score of 0.853. Figure 1B shows the observed binding site for T0819 with the binding site residues coloured in blue and the PLP ligand coloured yellow. This is an example of a good prediction and would therefore not benefit from refinement using docking, as when a prediction is already good there is less room for improvement.

The CASP12 target is T0911 *Escherichia Coli* Figure 1C shows the predicted ligand binding site, with correctly predicted binding site residues in blue and under- and over-predictions in red, the predicted dibromotyrosine ligand is coloured yellow. This prediction has a MCC score of -0.006 and BDT score of 0.006. Figure 1D shows the observed binding site for T0911 with the binding site residues coloured in blue and the gluconic ligand coloured yellow. A target such as this, would be a good candidate for refinement using docking because the correct ligand has been predicted. However, the ligand predicted is in an incorrect part of the protein and needs to be improved, initially by being in the correct section of the protein, at the very least and this could potentially lead to a better ligand-binding site prediction.

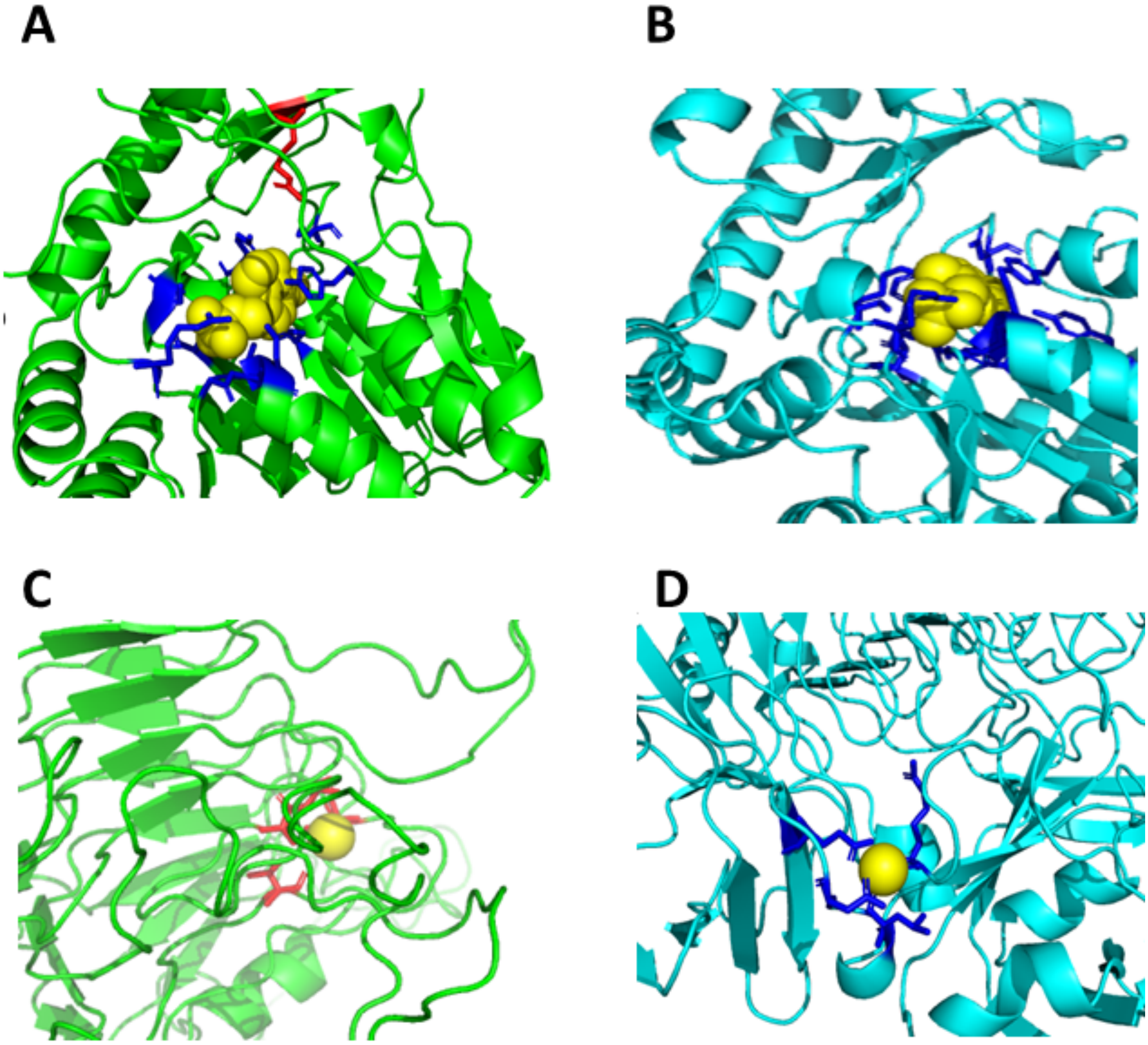

A
B
C
D

Figure 1. Comparison of FunFOLD3 ligand binding site predictions (A and C) for CASP targets, compared to the observed ligand binding sites (B and D). **(A)** Predicted ligand binding site residues for T0819 (PDB ID 4wbt) shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the atoms in the pyridoxal-5'-phosphate (PLP) ligand are shown as spheres and coloured yellow. BDT score of 0.853 and MCC score of 0.877. **(B)** The observed ligand binding site for T0819 (PDB ID 4wbt), with binding site residues shown as sticks and coloured in blue and the ligand PLP is coloured yellow. **(C)** Predicted ligand binding site residues for T0912 shown as sticks with incorrect predictions in red, the calcium ligand is shown as a sphere and coloured yellow. BDT score of 0.006 and MCC score of -0.006. **(D)** The observed ligand binding site residues shown as sticks for T0912 with binding site residues coloured in blue and the calcium coloured yellow.

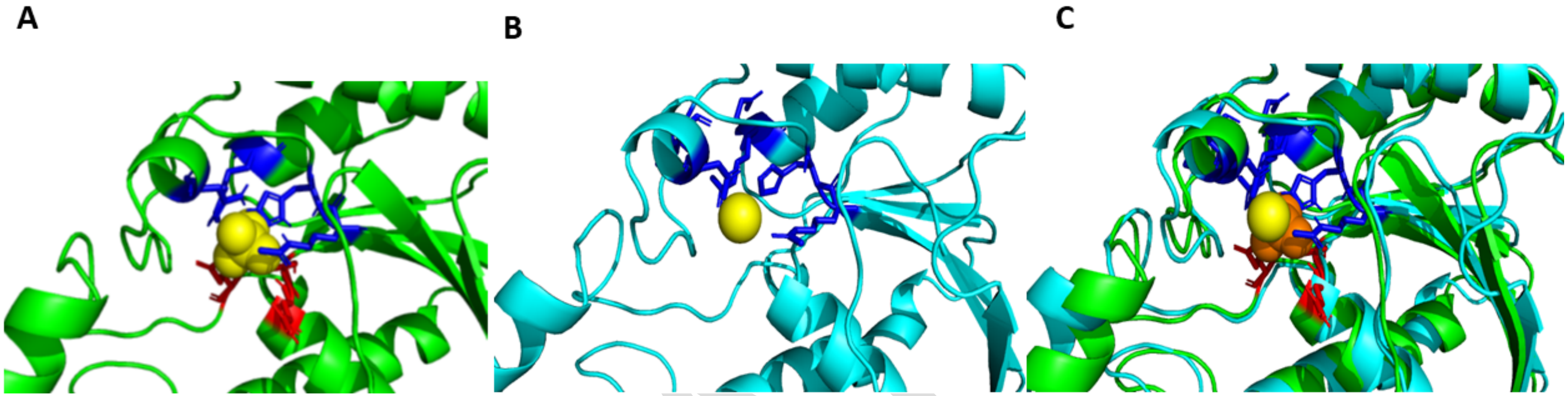

A
B
C

Figure 2. Comparison of FunFOLD3 ligand binding site predictions (a) for CASP13 T1016 target, compared to the observed ligand binding sites (b) and alignment of the predicted and observed proteins and ligand (c) **(A)** Predicted ligand binding site residues for CASP13 T1016 (PDB ID 6e4b) shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the phosphate (PO4) ligand atoms are shown as spheres and coloured yellow. BDT score of 0.646 and MCC score of 0.556 **(B)** The observed ligand binding site for T1016 (PDB ID 6e4b), with binding site residues shown as sticks and coloured in blue and the ligand cholrine coloured yellow. **(C)** Comparison of the predicted and observed protein structures and ligands with the predicted ligand PO4 coloured orange and the observed ligaind CL coloured yellow.

The results in Figure 2 show that despite an incorrect ligand prediction; PO_4 instead of CL FunFOLD3 was able to produce corect binding site residues, as shown by the blue sticks in Figure 2A. Figure 2C illustrates how close the two ligands were to one another in the binding pocket and the difference in size between the ligands, with the PO_4 ligand being larger and could explain why there were some incorrect bnding site residue predictions.

These results demonstrate not only the similarities in the structure but also the similarity in the location of the predicted and observed ligand, despite an incorrect prediction of the ligand type. The differing results from CASP11, CASP12 and CASP13 show the variability of results obained with FunFOLD3. Figure 1A and B show a prediction where the ligand is correct and there are few incorrect predicitions. In comparions, Figure 1C and D shows a correctly predicted ligand wbut in a different location within the structure aand Figure2A-C shows that FunFOLD3 can produce ligand-binding site predictions which are similar to the observed protein strcuture despite the predicted ligand not being the same ligand. Given the diversity of the predictions there is a clear need to refine and docking will be used to do this.

FunFOLD3-D

FunFOLD3-D is a new method, which utilises docking to improve the predicted ligand binding sites by rotating the predicted ligand in the predicted binding site space. The method outputs the predictions of both the ligand and ligand-binding site residues for nine alternative models, which should, in theory produce a more refined ligand-binding site. Note, nine models is the recommended number of binding models AutoDock Vina to generate. Currently, this method is under development and is being benchmarked using CASP11, CASP12 and CASP13 functional target predictions to determine if there has been an improvement in the ligand-binding site predictions. FunFOLD3-D will also be integrated into our CASP14 pipelines so we can objectively test our ligand-binding site predictions during a blind experiment. AutoDock Vina has been used to refine ligand-binding site predictions previously.[52] Wu et al., utilised molecular docking by AutoDock Vina in order to enhance the low quality of the predicted ligand-binding poses that usually had severe steric clashes to the protein structure. FunFOLD3-D will be different to COACH-D, due to a box calculation method. A grid size of 22.5Å was chosen as the grid space for docking because this space has been explored in literature[53] and ensures the space is large enough for the ligand to rotate. Note, this grid size may change once the final methodology has been finalised for FunFOLD3-D and is being used as a guide to start improving docking. Once the predicted ligand and receptor files have been docked, the output files will be analysed using

FunFOLD3 to produce new ligand-binding site residues and MCC and BDT scores calculated to objectively measure any changes in the predicted-ligand binding site residues.

Figure 3 shows the improved predictions, compared to the observed binding sites following docking. Figure 3A shows the predicted ligand-binding site for Glutathione S-transferase domain protein from *Haliangium ochraceum* (CASP11 ID T0849 and PDB ID 4w66), with correctly predicted ligand-binding site residues in blue and incorrect predictions in red. The GSH ligand is coloured yellow. This prediction by FunFOLD3 resulted in a MCC score of -0.05 and a BDT score of 0.0375. Figure 3B shows the observed binding site for T0849 with the binding site residues coloured in blue and the GSH ligand coloured yellow. Figure 3C shows the structure following docking with AutoDock Vina and the MCC score was 0.18 and BDT score was 0.28 for the best model (out of nine models). Residues predicted by FunFOLD3 were 9,10,14,15,54,55,56,67,68,108,113,226,230 and observed predictions were 168,171,179,182,183,190,194,197 with FunFOLD3-D predicting 13,107,108,111,168,169,171,172,215,217,218.

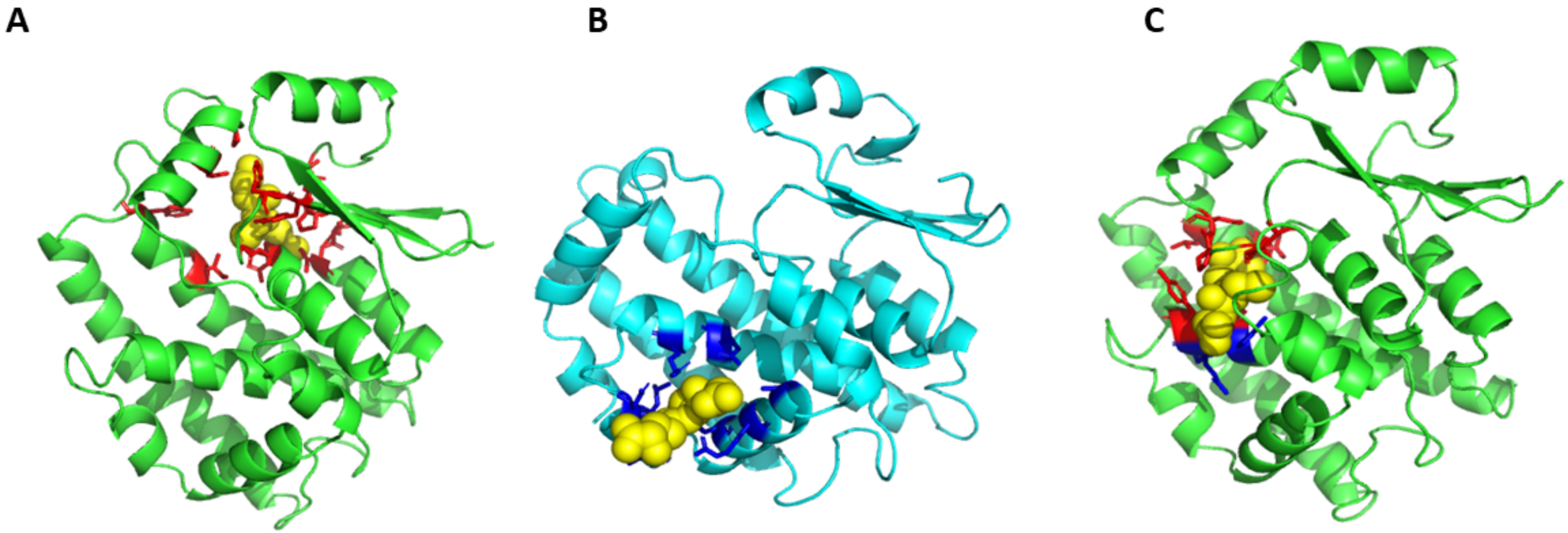

A
B
C

Figure 3. Comparison of FunFOLD3 and FunFOLD3-D ligand binding site predictions for CASP 11 target T0849 (PDB ID 4w66).

(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the glutathione (GSH) ligand is shown as a sphere and coloured yellow. BDT score of 0.0375 and MCC score of -0.05. **(B)** The observed ligand binding site residues shown as sticks for T0849 (PDB ID 4w66), with binding site residues coloured in blue and the ligand GSH coloured yellow **(C)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and over-predictions in red following docking with AutoDock Vina. The GSH ligand is coloured yellow. BDT score of 0.28 and MCC score of 0.18 was achieved for the top scoring docked model**.**

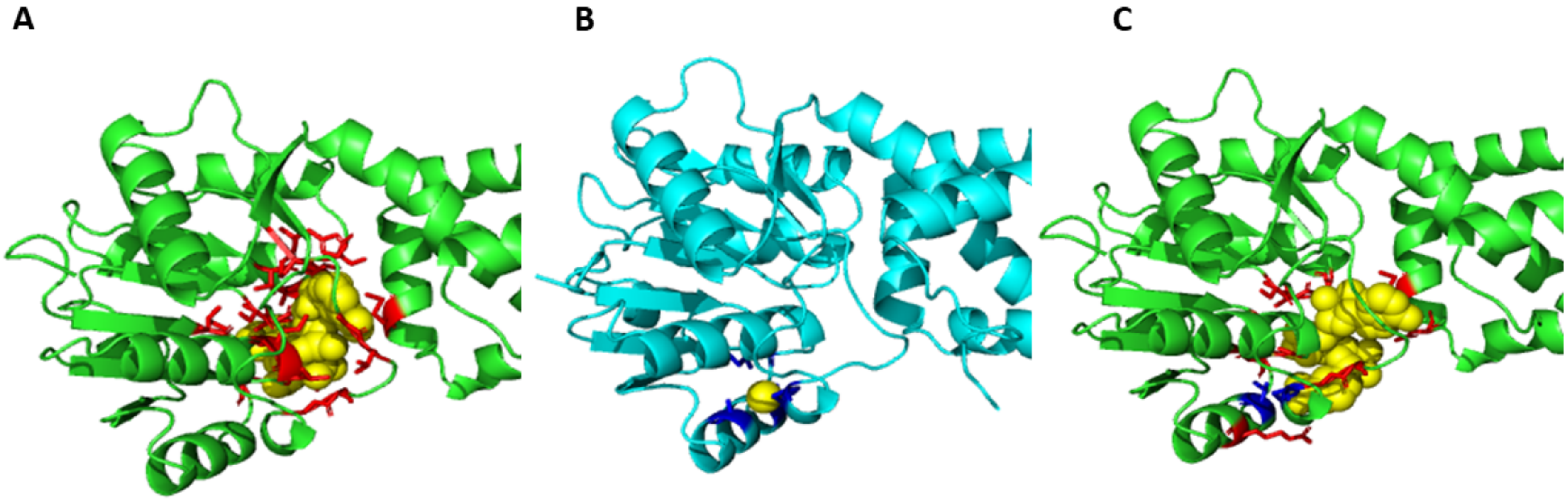

A
B
C

Figure 4. Comparison of FunFOLD3 and FunFOLD3-D ligand binding site predictions for CASP 11 target T0813 (PDB ID 4wji).

(A) Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in *blue* and incorrect predictions in *red*, the NAI ligand is shown as a sphere and coloured yellow. BDT score of 0.11and MCC score of 0.03. **(B)** The observed ligand binding site residues shown as sticks for T0813 (PDB ID 4wji), with binding site residues coloured in *blue* and the ligand magnesium (MG) coloured yellow **(C)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and over-predictions in red following docking with AutoDock Vina. The NAI ligand is coloured yellow. BDT score of 0.32 and MCC score of 0.25 was achieved for the top scoring docked model

Preliminary results from FunFOLD3-D shows that docking is able to enhance predictions of the same ligands within the same ligand-binding site pocket (Figures 3A-C) and also different ligands within the same pocket (Figure 4). Although, if the prediction is already good (e.g. MCC/BDT of >0.8), then there is less room for improvement and if the ligand is a small metal ion in within a small binding pocket, docking is unable to rotate the ligand with enough space to improve the predicted ligand-binding site.

Notes

1. As can be seen from the variation in MCC and BDT scores prediction of ligand-binding site residues is a difficult task and limitations can arise from what is deemed a biologically relevant ligand. Ligands, which are part of the PDB ID entry, are not necessarily biologically relevant ligand and information is required from the BioLip database and users must update this database regularly to ensure the information is as up-to-date as possible
2. Docking is quite time-consuming as each of the nine models which are produced following the refinement process needs to be analysed individually as currently there is no scoring method to pre-select the best model. This will be an area which will be improved on in the future
3. Not all proteins and ligands can be docked. If the ligand is a small metal ion bound in a tight space then there is not enough of a space for it to rotate and therefore will not benefit from docking

References

1. Roche, D. B., Buenavista, M. T. & McGuffin, L. J. FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. *PLoS One* **7**, e38219 (2012).

2. Roche, D.B. Buenavista, M.T, McGuffin, L. J. Predicting protein structures and structural annotation of proteomes. in *Encyclopedia of biophysics* (ed. Roberts, G. C. .)

469 (Springer: Berling and Heidelberg, 2012).

3. Roche, D. B., Buenavista, M. T. & McGuffin, L. J. The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Res.* **41**, W303-7 (2013).

4. Roche, D. B., Tetchner, S. J. & McGuffin, L. J. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics* **12**, 160 (2011).

5. Konc, J. & Janežič, D. ProBiS-ligands: a web server for prediction of ligands by examination of protein binding sites. *Nucleic Acids Res.* **42**, W215-20 (2014).

6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).

7. Dukka, B. K. Structure-based Methods for Computational Protein Functional Site Prediction. *Comput. Struct. Biotechnol. J.* **8**, e201308005 (2013).

8. Danishuddin, M. & Khan, A. U. Structure based virtual screening to discover putative drug candidates: necessary considerations and successful case studies. *Methods* **71**, 135–45 (2015).

9. McGuffin, L. J. *et al.* IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res.* **47**, W408–W413 (2019).

10. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **10**, 39 (2018).

11. Yang, J., Roy, A. & Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **29**, 2588–95 (2013).

12. Roche, D. B. & McGuffin, L. J. In silico Identification and Characterization of Protein-Ligand Binding Sites. *Methods Mol. Biol.* **1414**, 1–21 (2016).

13. Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **49**, 377–389 (2009).

14. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

15. Ghersi, D. & Sanchez, R. EasyMIFs and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **25**, 3185–3186 (2009).

16. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. & Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **5**, e1000585 (2009).

17. Wass, M. N., Kelley, L. A. & Sternberg, M. J. E. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **38**, W469-73 (2010).

18. Yu, J., Zhou, Y., Tanaka, I. & Yao, M. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **26**, 46–52 (2010).

19. Volkamer, A., Griewel, A., Grombacher, T. & Rarey, M. Analyzing the Topology of

Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.* **50**, 2041–2052 (2010).

20. Zhang, Z., Li, Y., Lin, B., Schroeder, M. & Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088 (2011).

21. Zhu, H. & Pisabarro, M. T. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **27**, 351–358 (2011).

22. Ngan, C.-H. *et al.* FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **28**, 286–287 (2012).

23. Xie, Z.-R. & Hwang, M. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **28**, 1579–1585 (2012).

24. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40**, W471–W477 (2012).

25. Lee, H. S. & Im, W. Ligand Binding Site Detection by Local Structure Alignment and Its Performance Complementarity. *J. Chem. Inf. Model.* **53**, 2462–2470 (2013).

26. Brylinski, M. & Feinstein, W. P. eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput. Aided. Mol. Des.* **27**, 551–567 (2013).

27. Heo, L., Shin, W.-H., Lee, M. S. & Seok, C. GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res.* **42**, W210–W214 (2014).

28. Viet Hung, L., Caprari, S., Bizai, M., Toti, D. & Polticelli, F. LIBRA: LIgand Binding site Recognition Application. *Bioinformatics* btv489 (2015). doi:10.1093/bioinformatics/btv489

29. Gao, J. *et al.* bSiteFinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming. *J. Cheminform.* **8**, 38 (2016).

30. Jian, J.-W. *et al.* Predicting Ligand Binding Sites on Protein Surfaces by 3-Dimensional Probability Density Distributions of Interacting Atoms. *PLoS One* **11**, e0160315 (2016).

31. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

32. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2013).

33. Lopez, G., Valencia, A. & Tress, M. FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* **35**, D219-23 (2007).

34. Dessailly, B. H., Lensink, M. F., Orengo, C. A. & Wodak, S. J. LigASite a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic*

Acids Res. **36**, D667–D673 (2007).

35. Benson, M. L. *et al.* Binding MOAD, a high-quality protein ligand database. *Nucleic Acids Res.* **36**, D674–D678 (2007).

36. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–80 (2004).

37. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35**, D198-201 (2007).

38. Magrane, M. & UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford).* **2011**, bar009 (2011).

39. Gallo Cassarino, T., Bordoli, L. & Schwede, T. Assessment of ligand binding site predictions in CASP10. *Proteins* **82 Suppl 2**, 154–63 (2014).

40. Roche, D. B., Brackenridge, D. A. & McGuffin, L. J. Proteins and their interacting partners: An introduction to protein-ligand binding site prediction methods. *Int. J. Mol. Sci.* **16**, (2015).

41. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–9 (2005).

42. Oh, M., Joo, K. & Lee, J. Protein-binding site prediction based on three-dimensional

protein modeling. *Proteins* **77 Suppl 9**, 152–6 (2009).

43. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–95 (2010).

44. López, G., Ezkurdia, I. & Tress, M. L. Assessment of ligand binding residue predictions in CASP8. *Proteins* **77 Suppl 9**, 138–46 (2009).

45. Haas, J. *et al.* The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database (Oxford).* **2013**, bat031 (2013).

46. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–7 (2013).

47. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* **84 Suppl 1**, 4–14 (2016).

48. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–51 (1975).

49. Schmidt, T., Haas, J., Gallo Cassarino, T. & Schwede, T. Assessment of ligand-binding residue predictions in CASP9. *Proteins* **79 Suppl 1**, 126–36 (2011).

50. Feng, Z. *et al.* Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **20**, 2153–5 (2004).

51. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically

relevant ligand–protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2012).

52. Wu, Q., Peng, Z., Zhang, Y. & Yang, J. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.* **46**, W438–W442 (2018).

53. Feinstein, W. P. & Brylinski, M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *J. Cheminform.* **7**, 18 (2015).

Appendix 2

1. How to analyse ligand-binding site prediction results from CASP competitions

1. Get a list of CASPIDs to PDBIDs from the CASP website
   http://www.predictioncenter.org/casp11/targetlist.cgi

2. Download the target structures from CASP
   http://www.predictioncenter.org/download_area/CASP11/targets/casp11.targets_unsplitted.release11242014.tgz
   unzip the file tar -xvzf targets_unsplitted.release11242014.tgz
3. Use the list of PDBIDs to search the BioLiP database to determine which proteins have biologically relevant ligands
   a. Once you have the list of PDBIDs search the BioLiP database to see which proteins have biologically relevant ligands
   b. The BioLiP database is available to download from
      http://zhanglab.ccmb.med.umich.edu/BioLip/download.html

   c. grep -f list of PDBIDs PDBIDsinBioLip
      The command more PDBIDs | wc –l will determine how many experimental structures were released

4. Download these PDB files from the PDB
   wget http://www.rcsb.org/pdb/files/1A3H.pdb

5. Extract the ligand atoms HETATMS and attached them to the CASP PDB files from number 2
   a. grep "HETATM" 4rn3.pdb | grep -v "HOH" > HET_T0854.het
   b. The HETATOMS need to be attached to the CASP PDB files
   c. cat T0854.pdb HET_T0854.het > T0854.pdb

6. Use proLigContacts.jar to determine ligand binding site residues, using new CASP PDB files
   a. To run proLigContacts see the READ ME
      wget http://www.reading.ac.uk/bioinf/downloads/proLigContacts.sh
      wget http://www.reading.ac.uk/bioinf/downloads/proLigContacts.jar
      chmod +x proLigContacts.jar
      chmod +x proLigContacts.sh

7. Predictions need to be analysed using IntFOLD TS133 and McGuffin TS162
   a. Models need to be downloaded from CASP wget
      http://www.predictioncenter.org/download_area/CASP11/predictions/HETATM_models.tar

8. The IntFOLD TS133 models need to be assessed using proLigContacts.jar as in step 6

9. Once the predicted ligand binding site residues and observed predicted binding site residues have been determined, compare using BDT score

10. To run the BDT score see the READ ME
    a. http://www.reading.ac.uk/bioinf/downloads/README_BDT
    b. Download at http://www.reading.ac.uk/bioinf/downloads/bdt.sh

2. FunFOLD3 Downloadable Executable

A downloadable version of the FunFOLD3 method is available as an executable JAR file, which can be run locally. The executable has several dependencies and system requirements, which are briefly described below. The executable along with a detailed README file and example input and output data can be downloaded from the following location: http://www.reading.ac.uk/bioinf/downloads/

The system requirements are as follows:
1. A linux-based operating system such as Ubuntu.
2. A recent version of Java (http://www.java.com/getjava/).
3. A recent version of PyMOL (http://www.pymol.org).
4. The TM-align program (http://zhanglab.ccmb.med.umich.edu/TM-align/)
5. wget and ImageMagick installed system wide.
6. The CIF chemical components database file should be downloaded from here: ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif.
7. The BioLip databases containing ligand and receptor PDB files are also required. The databases need to be downloaded in two sections: firstly all annotations prior to 6/3/2013 can be downloaded from here for the receptor database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/receptor_2013-03-6.tar.bz2 (3.6 G) and from here for the ligand database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/ligand_2013-03-6.tar.bz2(438 M). The Text File of the BioLip annotations can be downloaded from here: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2. To update the databases to include annotations after 2013-03-6 it is recommended to download and use this perl script which will update the databases: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/download_all_sets.pl. The BioLip text file: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2 and all the weekly update text files should be concatenated to form a large text file containing all of the annotations. Furthermore, it is recommended to regularly update your BioLip and CIF databases. Additionally, a shell script is available as downloadBioLipdata.sh, which can be downloaded from here: http://www.reading.ac.uk/bioinf/downloads/, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edits the file paths for the location of the BioLip databases and the executable directory.

3. Predicted and observed structure files for CASP11 target T0807

Predicted structure is titled T0807.lig2.pdb and observed structure is T0807_HET.pdb, the MSE/MET residues are highlighted in yellow and only the relevant ATOMS are shown

T0807.lig2.pdb

```
ATOM   170  CB  TYR   23    58.000  49.930  15.259  1.00  2.26      C
ATOM   171  CG  TYR   23    59.011  50.998  15.543  1.00  2.26      C
ATOM   172  CD1 TYR   23    59.815  51.495  14.541  1.00  2.26      C
ATOM   173  CD2 TYR   23    59.138  51.537  16.805  1.00  2.26      C
ATOM   174  CE1 TYR   23    60.744  52.479  14.789  1.00  2.26      C
ATOM   175  CE2 TYR   23    60.064  52.524  17.064  1.00  2.26      C
ATOM   176  CZ  TYR   23    60.868  52.999  16.056  1.00  2.26      C
ATOM   177  OH  TYR   23    61.818  54.011  16.317  1.00  2.26      O
ATOM   392  CA  GLU   53    69.598  46.902  11.215  1.00  0.70      C
ATOM   393  C   GLU   53    68.986  47.309  12.521  1.00  0.70      C
ATOM   394  O   GLU   53    68.821  46.491  13.424  1.00  0.70      O
ATOM   395  CB  GLU   53    71.085  47.293  11.230  1.00  0.70      C
ATOM   396  CG  GLU   53    71.832  46.805   9.985  1.00  0.70      C
ATOM   397  CD  GLU   53    73.192  47.482   9.959  1.00  0.70      C
ATOM   398  OE1 GLU   53    73.217  48.729   9.787  1.00  0.70      O
ATOM   399  OE2 GLU   53    74.221  46.770  10.117  1.00  0.70      O1-
ATOM   673  CB  ALA   89    86.478  50.084   8.954  1.00  1.08      C
ATOM   674  N   ALA   90    83.495  50.558   9.564  1.00  1.04      N
ATOM   675  CA  ALA   90    82.381  50.112  10.340  1.00  1.04      C
ATOM   676  C   ALA   90    81.230  49.809   9.436  1.00  1.04      C
ATOM   677  O   ALA   90    80.495  48.851   9.674  1.00  1.04      O
ATOM   678  CB  ALA   90    81.919  51.150  11.374  1.00  1.04      C
ATOM   679  N   MET   91    81.031  50.621   8.379  1.00  0.71      N
ATOM   680  CA  MET   91    79.910  50.400   7.508  1.00  0.71      C
ATOM   826  C   ASP  108    69.220  46.438  -3.263  1.00  0.66      C
ATOM   827  O   ASP  108    69.684  47.124  -4.173  1.00  0.66      O
ATOM   828  CB  ASP  108    67.429  44.727  -3.727  1.00  0.66      C
ATOM   829  CG  ASP  108    67.296  43.262  -4.142  1.00  0.66      C
ATOM   830  OD1 ASP  108    68.014  42.849  -5.092  1.00  0.66      O
ATOM   831  OD2 ASP  108    66.492  42.527  -3.506  1.00  0.66      O1-
ATOM   832  N   MET  109    68.971  46.962  -2.040  1.00  0.50      N
ATOM   833  CA  MET  109    69.249  48.352  -1.778  1.00  0.50      C
ATOM   975  CG  GLU  125    84.187  56.026   1.852  1.00  1.22      C
ATOM   976  CD  GLU  125    85.647  56.456   1.907  1.00  1.22      C
ATOM   977  OE1 GLU  125    86.363  56.202   0.902  1.00  1.22      O
ATOM   978  OE2 GLU  125    86.068  57.030   2.947  1.00  1.22      O1-
ATOM   979  N   ALA  126    81.949  53.841   1.520  1.00  0.75      N
ATOM   980  CA  ALA  126    81.898  52.435   1.805  1.00  0.75      C
ATOM   981  C   ALA  126    80.813  51.802   0.981  1.00  0.75      C
ATOM   982  O   ALA  126    80.981  50.700   0.460  1.00  0.75      O
ATOM  1148  O   GLU  147    73.640  69.207   1.369  1.00  1.21      O
ATOM  1149  CB  GLU  147    71.714  71.284   2.939  1.00  1.21      C
ATOM  1150  CG  GLU  147    71.143  70.696   4.230  1.00  1.21      C
ATOM  1151  CD  GLU  147    72.212  69.834   4.887  1.00  1.21      C
ATOM  1152  OE1 GLU  147    73.369  69.830   4.388  1.00  1.21      O
ATOM  1153  OE2 GLU  147    71.882  69.173   5.908  1.00  1.21      O1-
ATOM   1154  N   ASP  148     72.050  68.116   2.534  1.00  0.84          N
ATOM   1155  CA  ASP  148     72.871  66.954   2.761  1.00  0.84          C
ATOM   1485  CG  LEU  188     67.281  60.629  -4.450  1.00  1.58          C
ATOM   1486  CD1 LEU  188     68.142  59.360  -4.395  1.00  1.58          C
ATOM   1487  CD2 LEU  188     67.560  61.584  -3.273  1.00  1.58          C
ATOM   1488  N   PRO  189     65.123  58.945  -6.755  1.00  0.93          N
ATOM   1489  CA  PRO  189     63.766  58.601  -6.436  1.00  0.93          C
ATOM   1490  C   PRO  189     63.384  58.764  -4.997  1.00  0.93          C
ATOM   1491  O   PRO  189     64.096  58.267  -4.122  1.00  0.93          O
ATOM   1659  C   ASP  211     39.123  63.771   3.658  1.00  2.03          C
ATOM   1660  O   ASP  211     38.518  63.547   2.611  1.00  2.03          O
ATOM   1661  CB  ASP  211     40.202  65.870   4.309  1.00  2.03          C
ATOM   1662  CG  ASP  211     40.209  66.566   2.960  1.00  2.03          C
ATOM   1663  OD1 ASP  211     39.153  66.540   2.275  1.00  2.03          O
ATOM   1664  OD2 ASP  211     41.272  67.138   2.597  1.00  2.03          O1-
ATOM   1665  N   ILE  212     39.983  62.891   4.200  1.00  1.69          N
ATOM   1666  CA  ILE  212     40.349  61.660   3.559  1.00  1.69          C
ATOM   2170  CB  LYS  278     61.710  68.817  12.821  1.00  6.56          C
ATOM   2171  CG  LYS  278     60.924  68.471  14.088  1.00  6.56          C
ATOM   2172  CD  LYS  278     61.579  67.452  15.015  1.00  6.56          C
ATOM   2173  CE  LYS  278     60.847  67.295  16.342  1.00  6.56          C
ATOM   2174  NZ  LYS  278     60.987  68.536  17.135  1.00  6.56          N1+
ATOM   2175  N   ASN  279     63.868  66.672  11.017  1.00  6.32          N
ATOM   2176  CA  ASN  279     65.122  65.984  11.028  1.00  6.32          C
ATOM   2177  C   ASN  279     66.137  66.830  11.708  1.00  6.32          C
TER
HETATM 4516  P2B NAP  285     48.830  47.414  12.414  1.00 42.98           P
HETATM 4517  PA  NAP  285     53.080  53.596   8.953  1.00 36.22           P
HETATM 4518  PN  NAP  285     55.818  53.098   9.912  1.00 39.54           P
HETATM 4519  C1B NAP  285     48.401  50.831  10.770  1.00 41.97           C
HETATM 4520  C1D NAP  285     60.077  52.556   7.643  1.00 30.93           C
HETATM 4521  N1A NAP  285     44.001  50.746   8.027  1.00 50.77           N
HETATM 4522  N1N NAP  285     60.571  53.621   8.584  1.00 32.67           N
HETATM 4523  O1A NAP  285     53.311  52.503   7.922  1.00 36.48           O1-
HETATM 4524  O1N NAP  285     55.838  51.654  10.351  1.00 36.89           O1-
HETATM 4525  O1X NAP  285     48.577  46.886  13.811  1.00 45.95           O
HETATM 4526  C2A NAP  285     44.205  51.231   9.302  1.00 44.20           C
HETATM 4527  C2B NAP  285     49.299  49.808  11.432  1.00 43.23           C
HETATM 4528  C2D NAP  285     59.833  51.158   8.264  1.00 34.76           C
HETATM 4529  C2N NAP  285     61.746  54.249   8.282  1.00 34.51           C
HETATM 4530  O2A NAP  285     52.805  54.966   8.442  1.00 23.10           O
HETATM 4531  O2B NAP  285     48.560  48.998  12.312  1.00 37.76           O
HETATM 4532  O2D NAP  285     60.083  50.071   7.417  1.00 30.65           O
HETATM 4533  O2N NAP  285     56.559  54.025  10.803  1.00 33.97           O
HETATM 4534  O2X NAP  285     50.292  47.003  12.271  1.00 40.71           O1-
HETATM 4535  C3B NAP  285     50.341  50.656  12.136  1.00 42.35           C
HETATM 4536  C3D NAP  285     58.357  51.198   8.600  1.00 23.11           C
HETATM 4537  C3N NAP  285     62.278  55.235   9.117  1.00 35.14           C
HETATM 4538  N3A NAP  285     45.444  51.144   9.888  1.00 47.51           N
HETATM 4539  O3  NAP  285     54.312  53.706   9.945  1.00 33.14           O
HETATM 4540  O3B NAP  285     49.997  50.928  13.486  1.00 39.71           O
HETATM 4541  O3D NAP  285     57.858  49.934   9.035  1.00 38.29           O
HETATM 4542  O3X NAP  285     47.889  46.951  11.309  1.00 35.79           O
HETATM 4543  C4A NAP   285      46.463  50.595   9.202  1.00 41.87           C
HETATM 4544  C4B NAP   285      50.365  51.969  11.339  1.00 35.15           C
HETATM 4545  C4D NAP   285      57.849  51.859   7.340  1.00 29.61           C
HETATM 4546  C4N NAP   285      61.587  55.646  10.277  1.00 34.11           C
HETATM 4547  O4B NAP   285      49.274  51.927  10.417  1.00 33.45           O
HETATM 4548  O4D NAP   285      58.791  52.912   7.169  1.00 39.86           O
HETATM 4549  C5A NAP   285      46.293  50.104   7.919  1.00 43.76           C
HETATM 4550  C5B NAP   285      51.753  52.178  10.735  1.00 28.06           C
HETATM 4551  C5D NAP   285      56.441  52.464   7.310  1.00 35.60           C
HETATM 4552  C5N NAP   285      60.368  55.013  10.565  1.00 31.88           C
HETATM 4553  O5B NAP   285      51.753  53.246   9.834  1.00 38.18           O
HETATM 4554  O5D NAP   285      56.331  53.407   8.381  1.00 45.08           O
HETATM 4555  C6A NAP   285      45.044  50.162   7.321  1.00 43.07           C
HETATM 4556  C6N NAP   285      59.881  54.009   9.741  1.00 34.06           C
HETATM 4557  N6A NAP   285      44.899  49.895   6.014  1.00 32.48           N
HETATM 4558  C7N NAP   285      63.531  55.911   8.650  1.00 29.20           C
HETATM 4559  N7A NAP   285      47.483  49.634   7.487  1.00 42.95           N
HETATM 4560  N7N NAP   285      64.042  55.647   7.456  1.00 34.89           N
HETATM 4561  O7N NAP   285      64.149  56.831   9.477  1.00 32.62           O
HETATM 4562  C8A NAP   285      48.369  49.795   8.464  1.00 36.25           C
HETATM 4563  N9A NAP   285      47.752  50.381   9.532  1.00 38.51           N
HETATM 4589  P2B NAP   286      48.372  47.203  12.302  0.87 56.79           P
HETATM 4590  PA  NAP   286     52.521  53.392   8.625  1.00 55.11           P
HETATM 4591  PN  NAP   286     54.999  52.504   9.836  1.00 52.75           P
HETATM 4592  C1B NAP   286      48.390  50.685  11.971  0.74 56.24           C
HETATM 4593  C1D NAP   286      59.957  52.362   7.341  1.00 47.03           C
HETATM 4594  N1A NAP   286      43.282  51.206  11.637  0.14 56.15           N
HETATM 4595  N1N NAP   286      60.638  53.416   8.186  0.55 46.11           N
HETATM 4596  O1A NAP   286      52.218  54.692   9.068  1.00 54.77           O1-
HETATM 4597  O1N NAP   286      55.740  53.298  10.813  0.37 52.55           O1-
HETATM 4598  O1X NAP   286      48.208  46.432  10.997  1.00 55.84           O
HETATM 4599  C2A NAP   286      44.124  51.376  12.753  0.58 55.89           C
HETATM 4600  C2B NAP   286      49.457  49.584  11.638  0.63 56.35           C
HETATM 4601  C2D NAP   286      59.625  51.072   8.016  1.00 47.61           C
HETATM 4602  C2N NAP   286      61.731  54.203   7.697  1.00 45.43           C
HETATM 4603  O2A NAP   286      52.143  52.813   7.279  1.00 55.36           O
HETATM 4604  O2B NAP   286      49.350  48.484  12.552  0.77 56.47           O
HETATM 4605  O2D NAP   286      60.293  50.007   7.381  0.85 47.37           O
HETATM 4606  O2N NAP   286      55.351  51.099   9.484  0.47 52.85           O
HETATM 4607  O2X NAP   286      48.010  46.564  13.620  1.00 54.99           O1-
HETATM 4608  C3B NAP   286      50.795  50.311  11.644  1.00 56.36           C
HETATM 4609  C3D NAP   286      58.110  50.935   7.975  1.00 48.69           C
HETATM 4610  C3N NAP   286      62.325  55.139   8.568  0.57 45.01           C
HETATM 4611  N3A NAP   286      45.496  51.196  12.697  1.00 55.69           N
HETATM 4612  O3  NAP   286     54.098  53.172   8.718  0.07 53.79           O
HETATM 4613  O3B NAP   286      51.450  50.214  12.907  1.00 55.87           O
HETATM 4614  O3D NAP   286      57.692  49.652   7.493  0.50 49.34           O
HETATM 4615  O3X NAP   286      49.724  46.518  12.298  1.00 56.47           O
HETATM 4616  C4A NAP   286      45.985  50.832  11.448  0.13 56.20           C
HETATM 4617  C4B NAP   286      50.489  51.815  11.496  1.00 56.82           C
HETATM 4618  C4D NAP   286      57.591  52.092   7.046  0.95 48.66           C
HETATM 4619  C4N NAP   286      61.843  55.308   9.946  1.00 44.89           C
HETATM 4620  O4B NAP   286      49.034  51.911  11.742  0.63 56.74           O
HETATM 4621  O4D NAP   286      58.738  52.939   6.777  0.41 47.93           O
HETATM 4622  C5A NAP   286      45.191  50.649  10.295  0.01 56.38           C
HETATM 4623  C5B NAP   286      50.576  52.364  10.023  0.55 55.96           C
HETATM 4624  C5D NAP   286      56.339  52.849   7.473  1.00 48.44           C
HETATM 4625  C5N NAP   286      60.754  54.510  10.383  0.78 44.94           C
HETATM 4626  O5B NAP   286      51.978  52.325   9.695  1.00 54.62           O
HETATM 4627  O5D NAP   286      56.252  53.013   8.940  0.42 51.15           O
HETATM 4628  C6A NAP   286      43.801  50.838  10.382  1.00 56.53           C
HETATM 4629  C6N NAP   286      60.136  53.559   9.522  1.00 45.32           C
HETATM 4630  N6A NAP   286      42.900  50.705   9.386  0.75 56.41           N
HETATM 4631  C7N NAP   286      63.503  56.008   8.079  0.47 44.52           C
HETATM 4632  N7A NAP   286      45.978  50.288   9.201  1.00 57.16           N
HETATM 4633  N7N NAP   286      63.976  55.912   6.866  1.00 42.78           N
HETATM 4634  O7N NAP   286      63.965  56.784   8.882  1.00 45.11           O
HETATM 4635  C8A NAP   286      47.193  50.268   9.713  0.63 56.39           C
HETATM 4636  N9A NAP   286      47.257  50.588  11.062  1.00 55.95           N
HETATM 4446  P2B NAP   287      48.752  47.496  12.256  1.00 31.77           P
HETATM 4447  PA  NAP   287      52.917  53.709   8.741  1.00 32.82           P
HETATM 4448  PN  NAP   287      55.577  53.511   9.923  1.00 32.70           P
HETATM 4449  C1B NAP   287      48.108  50.887  10.735  1.00 31.45           C
HETATM 4450  C1D NAP   287      59.889  52.688   7.389  1.00 28.62           C
HETATM 4451  N1A NAP   287      44.000  50.610   7.742  1.00 29.20           N
HETATM 4452  N1N NAP   287      60.385  53.845   8.209  1.00 28.60           N
HETATM 4453  O1A NAP   287      52.276  54.883   8.277  1.00 33.31           O1-
HETATM 4454  O1N NAP   287      55.453  52.056  10.131  1.00 33.36           O1-
HETATM 4455  O1X NAP   287      47.945  47.044  13.478  1.00 32.07           O
HETATM 4456  C2A NAP   287      44.100  51.061   9.075  1.00 29.16           C
HETATM 4457  C2B NAP   287      49.251  49.965  11.230  1.00 32.01           C
HETATM 4458  C2D NAP   287      59.686  51.404   8.120  1.00 29.07           C
HETATM 4459  C2N NAP   287      61.640  54.430   7.908  1.00 28.62           C
HETATM 4460  O2A NAP   287      53.336  52.552   7.870  1.00 32.91           O
HETATM 4461  O2B NAP   287      48.767  49.143  12.303  1.00 31.69           O
HETATM 4462  O2D NAP   287      60.028  50.304   7.302  1.00 28.32           O
HETATM 4463  O2N NAP   287      56.086  54.300  11.059  1.00 32.87           O
HETATM 4464  O2X NAP   287      48.051  47.230  10.962  1.00 30.63           O1-
HETATM 4465  C3B NAP   287      50.407  50.897  11.570  1.00 32.42           C
HETATM 4466  C3D NAP   287      58.205  51.412   8.472  1.00 29.10           C
HETATM 4467  C3N NAP   287      62.102  55.469   8.695  1.00 28.05           C
HETATM 4468  N3A NAP   287      45.297  51.071   9.772  1.00 29.58           N
HETATM 4469  O3  NAP   287      54.190  54.158   9.555  1.00 32.64           O
HETATM 4470  O3B NAP   287      50.524  51.052  12.982  1.00 33.07           O
HETATM 4471  O3D NAP   287      57.656  50.108   8.720  1.00 29.76           O
HETATM 4472  O3X NAP   287      50.202  47.050  12.363  1.00 31.46           O
HETATM 4473  C4A NAP   287      46.388  50.611   9.050  1.00 29.60           C
HETATM 4474  C4B NAP   287      50.034  52.322  11.064  1.00 32.28           C
HETATM 4475  C4D NAP   287      57.549  52.176   7.257  1.00 29.27           C
HETATM 4476  C4N NAP   287      61.326  55.968   9.833  1.00 28.18           C
HETATM 4477  O4B NAP   287      48.647  52.165  10.631  1.00 32.04           O
HETATM 4478  O4D NAP   287      58.622  53.013   6.726  1.00 28.45           O
HETATM 4479  C5A NAP   287      46.339  50.151   7.729  1.00 29.22           C
HETATM 4480  C5B NAP   287      50.664  52.773   9.680  1.00 33.16           C
HETATM 4481  C5D NAP   287      56.269  52.928   7.482  1.00 29.96           C
HETATM 4482  C5N NAP   287      60.062  55.359  10.102  1.00 28.31           C
HETATM 4483  O5B NAP   287      52.053  53.022   9.928  1.00 33.04           O
HETATM 4484  O5D NAP   287      56.437  53.776   8.623  1.00 31.64           O
HETATM 4485  C6A NAP   287      45.126  50.152   7.044  1.00 28.66           C
HETATM 4486  C6N NAP   287      59.569  54.295   9.306  1.00 28.40           C
HETATM 4487  N6A NAP   287      44.918  49.748   5.784  1.00 28.08           N
HETATM 4488  C7N NAP   287      63.463  56.084   8.360  1.00 28.40           C
HETATM 4489  N7A NAP   287      47.598  49.759   7.298  1.00 29.40           N
HETATM 4490  N7N NAP   287      64.172  55.656   7.338  1.00 28.47           N
HETATM 4491  O7N NAP   287      63.845  56.971   9.087  1.00 28.87           O
HETATM 4492  C8A NAP   287      48.346  49.984   8.350  1.00 29.72           C
HETATM 4493  N9A NAP   287      47.664  50.497   9.430  1.00 30.20           N
HETATM 4409  N   GLU   288    62.343  53.498   9.602  1.00 44.32           N
HETATM 4410  CA  GLU   288      62.398  54.456  10.665  1.00 42.32           C
HETATM 4411  C   GLU   288    62.929  55.795  10.146  1.00 43.26           C
HETATM 4412  O   GLU   288    62.241  56.785  10.312  1.00 45.81           O
HETATM 4413  CB  GLU   288      63.264  53.871  11.793  1.00 42.80           C
HETATM 4414  CG  GLU   288      63.165  54.578  13.130  1.00 41.81           C
HETATM 4415  CD  GLU   288      64.483  54.633  13.792  1.00 37.69           C
HETATM 4416  OE1 GLU   288      65.353  55.350  13.294  1.00 35.71           O
HETATM 4417  OE2 GLU   288      64.631  53.934  14.778  1.00 38.76           O
HETATM 4418  OXT GLU   288      63.995  55.975   9.539  1.00 42.82           O
HETATM 4817  P2B NAP   289      48.579  47.042  11.759  1.00 11.96           P
HETATM 4818  PA  NAP   289    52.818  53.223   8.615  1.00  8.70           P
HETATM 4819  PN  NAP   289    55.516  53.046   9.993  1.00  9.44           P
HETATM 4820  C1B NAP   289      48.255  50.486  10.368  1.00 12.08           C
HETATM 4821  C1D NAP   289      59.860  52.322   7.486  1.00  8.97           C
HETATM 4822  N1A NAP   289      44.043  50.537   7.446  1.00 11.88           N
HETATM 4823  N1N NAP   289      60.316  53.467   8.312  1.00  9.65           N
HETATM 4824  O1A NAP   289      52.651  54.664   8.449  1.00  9.12           O1-
HETATM 4825  O1N NAP   289      55.830  54.261  10.627  1.00 12.96           O1-
HETATM 4826  O1X NAP   289      50.025  46.545  11.677  1.00 11.96           O
HETATM 4827  C2A NAP   289      44.205  51.000   8.698  1.00 11.71           C
HETATM 4828  C2B NAP   289      49.291  49.520  10.920  1.00 12.87           C
HETATM 4829  C2D NAP   289      59.765  51.017   8.228  1.00  8.13           C
HETATM 4830  C2N NAP   289      61.600  53.961   8.081  1.00  9.64           C
HETATM 4831  O2A NAP   289      53.086  52.409   7.364  1.00  8.99           O
HETATM 4832  O2B NAP   289      48.713  48.662  11.937  1.00 12.87           O
HETATM 4833  O2D NAP   289      59.901  49.872   7.363  1.00  8.83           O
HETATM 4834  O2N NAP   289      56.094  51.740  10.381  1.00 11.84           O
HETATM 4835  O2X NAP   289      47.838  46.572  12.993  1.00 14.51           O1-
HETATM 4836  C3B NAP   289      50.428  50.394  11.391  1.00 15.67           C
HETATM 4837  C3D NAP   289      58.312  51.081   8.751  1.00  8.69           C
HETATM 4838  C3N NAP   289      62.058  55.040   8.844  1.00  8.80           C
HETATM 4839  N3A NAP   289      45.360  50.924   9.361  1.00 11.57           N
HETATM 4840  O3  NAP   289    54.108  53.025   9.485  1.00 12.18           O
HETATM 4841  O3B NAP   289      50.538  50.333  12.810  1.00 20.93           O
HETATM 4842  O3D NAP   289      57.781  49.789   9.160  1.00  8.43           O
HETATM 4843  O3X NAP   289      47.826  46.815  10.451  1.00 13.33           O
HETATM 4844  C4A NAP   289      46.453  50.320   8.729  1.00 10.74           C
HETATM 4845  C4B NAP   289      49.995  51.835  11.067  1.00 14.21           C
HETATM 4846  C4D NAP   289      57.604  51.618   7.459  1.00  9.42           C
HETATM 4847  C4N NAP   289      61.220  55.601   9.854  1.00  9.22           C
HETATM 4848  O4B NAP   289      48.989  51.692  10.128  1.00 11.95           O
HETATM 4849  O4D NAP   289      58.535  52.616   6.999  1.00  9.03           O
HETATM 4850  C5A NAP   289      46.323  49.847   7.438  1.00 10.48           C
HETATM 4851  C5B NAP   289      50.897  52.861  10.452  1.00 15.97           C
HETATM 4852  C5D NAP   289      56.239  52.223   7.565  1.00  8.53           C
HETATM 4853  C5N NAP   289      59.902  55.130  10.092  1.00  9.61           C
HETATM 4854  O5B NAP   289      51.574  52.606   9.355  1.00 12.40           O
HETATM 4855  O5D NAP   289      55.873  53.138   8.521  1.00 11.43           O
HETATM 4856  C6A NAP   289      45.116  49.929   6.776  1.00 10.56           C
HETATM 4857  C6N NAP   289      59.455  54.072   9.304  1.00 10.42           C
HETATM 4858  N6A NAP   289      44.879  49.494   5.555  1.00 11.68           N
HETATM 4859  C7N NAP   289      63.396  55.677   8.530  1.00  8.44           C
HETATM 4860  N7A NAP   289      47.484  49.292   7.011  1.00 10.29           N
HETATM 4861  N7N NAP   289      63.943  55.344   7.357  1.00  9.01           N
HETATM 4862  O7N NAP   289      63.917  56.444   9.327  1.00  9.04           O
HETATM 4863  C8A NAP   289      48.325  49.467   8.056  1.00  9.99           C
HETATM 4864  N9A NAP   289      47.685  50.090   9.109  1.00  9.61           N
HETATM 4865  C1  TES   290      62.209  53.252  16.939  1.00 24.00           C
HETATM 4866  C2  TES   290      61.560  54.243  15.964  1.00 24.69           C
HETATM 4867  C3  TES   290      62.223  55.569  15.996  1.00 24.73           C
HETATM 4868  O3  TES   290      61.637  56.621  15.921  1.00 26.11           O
HETATM 4869  C4  TES   290      63.663  55.544  16.083  1.00 25.00           C
HETATM 4870  C5  TES   290      64.386  54.438  16.380  1.00 24.51           C
HETATM 4871  C6  TES   290      65.882  54.530  16.406  1.00 24.61           C
HETATM 4872  C7  TES   290      66.477  54.033  17.718  1.00 24.36           C
HETATM 4873  C8  TES   290      65.913  52.645  18.031  1.00 24.57           C
HETATM 4874  C9  TES   290      64.381  52.580  18.031  1.00 23.76           C
HETATM 4875  C10 TES   290      63.733  53.081  16.732  1.00 24.54           C
HETATM 4876  C11 TES   290      63.786  51.195  18.377  1.00 25.17           C
HETATM 4877  C12 TES   290      64.339  50.711  19.728  1.00 24.81           C
HETATM 4878  C13 TES   290      65.865  50.727  19.744  1.00 25.69           C
HETATM 4879  C14 TES   290      66.359  52.138  19.399  1.00 24.87           C
HETATM 4880  C15 TES   290      67.852  52.126  19.728  1.00 25.35           C
HETATM 4881  C16 TES   290      67.912  51.272  21.008  1.00 25.72           C
HETATM 4882  C17 TES   290      66.493  50.670  21.174  1.00 25.73           C
HETATM 4883  O17 TES   290      66.465  49.376  21.788  1.00 27.78           O
HETATM 4884  C18 TES   290      66.469  49.657  18.813  1.00 26.08           C
HETATM 4885  C19 TES   290      64.027  52.115  15.566  1.00 24.27           C
HETATM 4809  P2B NAP   291      47.862  46.959  11.412  1.00 33.49           P
HETATM 4810  PA  NAP   291      52.729  53.384   8.631  1.00 30.11           P
HETATM 4811  PN  NAP   291      55.086  52.603  10.398  1.00 22.86           P
HETATM 4812  C1B NAP   291      48.221  50.349   9.783  1.00 31.30           C
HETATM 4813  C1D NAP   291      59.749  52.317   7.445  1.00 14.05           C
HETATM 4814  N1A NAP   291      43.972  50.303   6.918  1.00 31.63           N
HETATM 4815  N1N NAP   291      60.288  53.392   8.309  1.00 17.27           N
HETATM 4816  O1A NAP   291      52.339  54.768   8.273  1.00 30.65           O1-
HETATM 4817  O1N NAP   291      54.557  51.248  10.659  1.00 31.27           O1-
HETATM 4818  O1X NAP   291      49.180  46.301  11.548  1.00 34.88           O
HETATM 4819  C2A NAP   291      44.187  50.861   8.099  1.00 31.56           C
HETATM 4820  C2B NAP   291      48.943  49.216  10.493  1.00 34.03           C
HETATM 4821  C2D NAP   291      59.906  50.992   8.153  1.00  9.22           C
HETATM 4822  C2N NAP   291      61.549  53.946   8.046  1.00 15.48           C
HETATM 4823  O2A NAP   291      52.969  52.420   7.533  1.00 32.59           O
HETATM 4824  O2B NAP   291      48.060  48.553  11.421  1.00 29.26           O
HETATM 4825  O2D NAP   291      60.198  49.916   7.270  1.00  8.60           O
HETATM 4826  O2N NAP   291      55.561  53.396  11.546  1.00 33.61           O
HETATM 4827  O2X NAP   291      46.807  46.643  12.393  1.00 41.99           O1-
HETATM 4828  C3B NAP   291      50.082  49.905  11.226  1.00 33.29           C
HETATM 4829  C3D NAP   291      58.537  50.779   8.756  1.00 17.33           C
HETATM 4830  C3N NAP   291      62.042  54.967   8.851  1.00 15.25           C
HETATM 4831  N3A NAP   291      45.358  50.774   8.705  1.00 30.71           N
HETATM 4832  O3  NAP  291      54.000  53.497   9.623  1.00 26.62           O
HETATM 4833  O3B NAP   291      49.791  50.081  12.612  1.00 40.48           O
HETATM 4834  O3D NAP   291      58.309  49.388   9.002  1.00 18.83           O
HETATM 4835  O3X NAP   291      47.281  46.661   9.942  1.00 31.41           O
HETATM 4836  C4A NAP   291      46.370  50.115   8.124  1.00 30.33           C
HETATM 4837  C4B NAP   291      50.190  51.273  10.597  1.00 32.57           C
HETATM 4838  C4D NAP   291      57.650  51.364   7.685  1.00 16.79           C
HETATM 4839  C4N NAP   291      61.289  55.422   9.929  1.00 15.51           C
HETATM 4840  O4B NAP   291      49.171  51.404   9.590  1.00 29.69           O
HETATM 4841  O4D NAP   291      58.340  52.548   7.248  1.00 17.50           O
HETATM 4842  C5A NAP   291      46.199  49.505   6.891  1.00 30.44           C
HETATM 4843  C5B NAP   291      51.544  51.422   9.951  1.00 31.71           C
HETATM 4844  C5D NAP   291      56.247  51.643   8.202  1.00 22.46           C
HETATM 4845  C5N NAP   291      60.043  54.872  10.202  1.00 18.85           C
HETATM 4846  O5B NAP   291      51.571  52.800   9.573  1.00 32.69           O
HETATM 4847  O5D NAP   291      56.291  52.521   9.337  1.00 28.12           O
HETATM 4848  C6A NAP   291      44.945  49.617   6.280  1.00 30.03           C
HETATM 4849  C6N NAP   291      59.538  53.864   9.391  1.00 16.60           C
HETATM 4850  N6A NAP   291      44.670  49.130   5.077  1.00 33.92           N
HETATM 4851  C7N NAP   291      63.387  55.623   8.544  1.00 13.34           C
HETATM 4852  N7A NAP   291      47.351  48.928   6.561  1.00 32.23           N
HETATM 4853  N7N NAP   291      63.920  55.306   7.375  1.00 11.60           N
HETATM 4854  O7N NAP   291      63.916  56.391   9.341  1.00 16.00           O
HETATM 4855  C8A NAP   291      48.222  49.178   7.539  1.00 32.91           C
HETATM 4856  N9A NAP   291      47.629  49.902   8.493  1.00 31.70           N
HETATM 4857  C1  TES  292      64.218  55.021  13.948  1.00 30.63           C
HETATM 4858  C2  TES  292      63.898  55.110  12.451  1.00 28.75           C
HETATM 4859  C3  TES  292      63.280  53.858  11.947  1.00 29.08           C
HETATM 4860  O3  TES  292      63.609  53.348  10.901  1.00 26.61           O
HETATM 4861  C4  TES  292      62.263  53.273  12.791  1.00 31.33           C
HETATM 4862  C5  TES  292      62.122  53.591  14.104  1.00 30.77           C
HETATM 4863  C6  TES  292      61.070  52.893  14.919  1.00 31.61           C
HETATM 4864  C7  TES  292      61.604  52.345  16.238  1.00 34.80           C
HETATM 4865  C8  TES  292      62.328  53.469  16.988  1.00 38.01           C
HETATM 4866  C9  TES  292      63.461  54.088  16.161  1.00 32.96           C
HETATM 4867  C10 TES   292      62.988  54.659  14.814  1.00 32.11           C
HETATM 4868  C11 TES   292      64.304  55.171  16.869  1.00 35.54           C
HETATM 4869  C12 TES   292      64.824  54.638  18.213  1.00 40.80           C
HETATM 4870  C13 TES   292      63.680  54.114  19.080  1.00 44.71           C
HETATM 4871  C14 TES   292      62.929  53.017  18.312  1.00 43.01           C
HETATM 4872  C15 TES   292      62.018  52.363  19.352  1.00 43.73           C
HETATM 4873  C16 TES   292      62.886  52.367  20.624  1.00 46.73           C
HETATM 4874  C17 TES   292      64.137  53.216  20.274  1.00 47.71           C
HETATM 4875  O17 TES   292      64.700  53.920  21.390  1.00 55.21           O
HETATM 4876  C18 TES   292      62.736  55.242  19.537  1.00 37.85           C
HETATM 4877  C19 TES   292      62.096  55.897  15.035  1.00 29.20           C
HETATM 4334  PA  NAD   293      52.736  54.088   8.925  1.00 30.06           P
HETATM 4335  PN  NAD   293      55.409  53.234   9.656  1.00 65.04           P
HETATM 4336  C1B NAD   293      48.548  51.054  10.409  1.00 48.31           C
HETATM 4337  C1D NAD   293      59.740  52.278   7.177  1.00 77.91           C
HETATM 4338  N1A NAD   293      44.390  50.772   7.931  1.00 42.79           N
HETATM 4339  N1N NAD   293      60.370  53.333   8.020  1.00 54.39           N
HETATM 4340  O1A NAD   293      52.408  55.427   8.391  1.00 17.09           O1-
HETATM 4341  O1N NAD   293      56.387  53.817  10.617  1.00 63.94           O1-
HETATM 4342  C2A NAD   293      44.747  51.209   9.148  1.00 34.34           C
HETATM 4343  C2B NAD   293      49.316  49.954  11.127  1.00 66.65           C
HETATM 4344  C2D NAD   293      59.514  50.931   7.893  1.00 70.45           C
HETATM 4345  C2N NAD   293      61.564  53.994   7.798  1.00 50.88           C
HETATM 4346  O2A NAD   293      52.818  52.972   7.935  1.00 53.49           O
HETATM 4347  O2B NAD   293      48.465  49.269  12.023  1.00 63.29           O
HETATM 4348  O2D NAD   293      59.986  49.817   7.119  1.00 27.64           O
HETATM 4349  O2N NAD   293      55.079  51.790   9.858  1.00 47.22           O
HETATM 4350  C3B NAD   293      50.467  50.672  11.822  1.00 96.77           C
HETATM 4351  C3D NAD   293      57.994  50.914   8.151  1.00 77.99           C
HETATM 4352  C3N NAD   293      62.032  55.003   8.657  1.00 52.93           C
HETATM 4353  N3A NAD   293      45.954  51.234   9.706  1.00 25.73           N
HETATM 4354  O3  NAD   293      54.023  54.163   9.728  1.00 64.56           O
HETATM 4355  O3B NAD   293      50.289  50.719  13.235  1.00117.09           O
HETATM 4356  O3D NAD   293      57.463  49.592   8.125  1.00 66.31           O
HETATM 4357  C4A NAD   293      46.867  50.736   8.863  1.00 20.08           C
HETATM 4358  C4B NAD   293      50.483  52.084  11.240  1.00 75.97           C
HETATM 4359  C4D NAD   293      57.433  51.748   7.001  1.00 83.36           C
HETATM 4360  C4N NAD   293      61.271  55.370   9.779  1.00 46.02           C
HETATM 4361  O4B NAD   293      49.378  52.191  10.316  1.00 53.72           O
HETATM 4362  O4D NAD   293      58.463  52.737   6.714  1.00 92.76           O
HETATM 4363  C5A NAD   293      46.656  50.252   7.588  1.00 20.67           C
HETATM 4364  C5B NAD   293      51.767  52.433  10.521  1.00 66.82           C
HETATM 4365  C5D NAD   293      56.036  52.397   7.214  1.00 71.60           C
HETATM 4366  C5N NAD   293      60.052  54.704  10.012  1.00 43.26           C
HETATM 4367  O5B NAD   293      51.722  53.735   9.968  1.00 53.19           O
HETATM 4368  O5D NAD   293      55.969  53.447   8.256  1.00 64.18           O
HETATM 4369  C6A NAD   293      45.340  50.271   7.114  1.00 31.82           C
HETATM 4370  C6N NAD   293      59.651  53.702   9.110  1.00 50.77           C
HETATM 4371  N6A NAD   293      44.994  49.823   5.897  1.00 23.68           N
HETATM 4372  C7N NAD   293      63.351  55.756   8.425  1.00 34.12           C
HETATM 4373  N7A NAD   293      47.840  49.813   7.025  1.00 22.49           N
HETATM 4374  N7N NAD   293      63.844  55.775   7.170  1.00 23.23           N
HETATM 4375  O7N NAD   293      63.870  56.335   9.375  1.00 21.93           O
HETATM 4376  C8A NAD   293      48.729  50.049   7.952  1.00 25.41           C
HETATM 4377  N9A NAD   293      48.208  50.626   9.071  1.00 34.84           N
HETATM 4716  P2B NAP   294      48.212  50.347  14.132  1.00 52.55           P
HETATM 4717  PA  NAP   294      52.945  55.492  10.533  1.00 38.71           P
HETATM 4718  PN  NAP   294      55.489  54.299   9.812  1.00 38.53           P
HETATM 4719  C1B NAP   294      47.843  53.250  11.477  1.00 45.87           C
HETATM 4720  C1D NAP   294      59.696  55.267  11.960  1.00 40.35           C
HETATM 4721  N1A NAP   294      43.408  52.004   9.231  1.00 46.73           N
HETATM 4722  N1N NAP   294      60.360  54.972  10.646  1.00 38.42           N
HETATM 4723  O1A NAP   294      52.667  56.304  11.803  1.00 35.06           O1-
HETATM 4724  O1N NAP   294      55.098  52.806   9.825  1.00 39.69           O1-
HETATM 4725  O1X NAP   294      49.722  50.069  14.274  1.00 54.86           O
HETATM 4726  C2A NAP   294      43.637  52.694  10.435  1.00 45.92           C
HETATM 4727  C2B NAP   294      48.766  52.320  12.320  1.00 46.29           C
HETATM 4728  C2D NAP   294      58.803  56.467  11.914  1.00 41.92           C
HETATM 4729  C2N NAP   294      61.590  55.566  10.288  1.00 35.94           C
HETATM 4730  O2A NAP   294      52.470  56.207   9.273  1.00 33.80          O
HETATM 4731  O2B NAP   294      48.060  51.854  13.485  1.00 49.87           O
HETATM 4732  O2D NAP   294      59.481  57.626  12.365  1.00 45.57           O
HETATM 4733  O2N NAP   294      55.790  54.771   8.356  1.00 36.49          O
HETATM 4734  O2X NAP   294      47.529  50.518  15.446  1.00 53.49           O1-
HETATM 4735  C3B NAP   294      49.990  53.179  12.627  1.00 45.71           C
HETATM 4736  C3D NAP   294      57.594  56.113  12.759  1.00 43.67           C
HETATM 4737  C3N NAP   294      62.139  55.268   9.049  1.00 33.96          C
HETATM 4738  N3A NAP   294      44.910  52.944  10.932  1.00 45.25           N
HETATM 4739  O3  NAP   294     54.345  55.338  10.491  1.00 38.85          O
HETATM 4740  O3B NAP   294      49.841  53.902  13.859  1.00 43.20           O
HETATM 4741  O3D NAP   294      57.658  56.674  14.075  1.00 46.01           O
HETATM 4742  O3X NAP   294      47.555  49.387  13.154  1.00 51.60           O
HETATM 4743  C4A NAP   294      45.943  52.456  10.152  1.00 43.77           C
HETATM 4744  C4B NAP   294      50.042  54.277  11.530  1.00 44.79           C
HETATM 4745  C4D NAP   294      57.559  54.560  12.778  1.00 42.44           C
HETATM 4746  C4N NAP   294      61.473  54.356   8.106  1.00 33.60          C
HETATM 4747  O4B NAP   294      48.651  54.312  11.017  1.00 45.17           O
HETATM 4748  O4D NAP   294      58.843  54.109  12.280  1.00 43.09           O
HETATM 4749  C5A NAP   294      45.777  51.763   8.942  1.00 43.89          C
HETATM 4750  C5B NAP   294      50.898  53.963  10.244  1.00 43.45           C
HETATM 4751  C5D NAP   294      56.471  53.959  11.940  1.00 42.22           C
HETATM 4752  C5N NAP   294      60.240  53.775   8.500  1.00 34.84          C
HETATM 4753  O5B NAP   294      52.243  53.974  10.719  1.00 41.84           O
HETATM 4754  O5D NAP   294      56.627  54.535  10.607  1.00 40.22           O
HETATM 4755  C6A NAP   294      44.479  51.520   8.457  1.00 45.23          C
HETATM 4756  C6N NAP   294      59.659  54.069   9.768  1.00 36.99          C
HETATM 4757  N6A NAP   294      44.143  50.873   7.320  1.00 46.04          N
HETATM 4758  C7N NAP   294      63.463  55.925   8.673  1.00 32.29          C
HETATM 4759  N7A NAP   294      47.006  51.408   8.406  1.00 41.98          N
HETATM 4760  N7N NAP   294      64.057  56.738   9.501  1.00 28.90          N
HETATM 4761  O7N NAP   294      63.913  55.664   7.575  1.00 31.78          O
HETATM 4762  C8A NAP   294      47.859  51.891   9.286  1.00 43.44          C
HETATM 4763  N9A NAP   294      47.270  52.536  10.353  1.00 44.65           N
```

4. Observed protein structure: T0807_HET.pdb

Note only HETATOMS have been shown

```
REMARK  T0807
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) :   1.80
REMARK   3   FREE R VALUE                     :  0.22129
REMARK   3   MEAN B VALUE      (OVERALL, A**2) :  34.365
CRYST1   99.483   99.483   55.181  90.00  90.00 120.00 P 62
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE
ATOM/HETATM
HETATM  170  N   MSE A  24      51.978  14.364  -0.720  1.00 26.86           N
HETATM  171  CA  MSE A  24      50.991  13.632   0.065  1.00 26.08           C
HETATM  172  C   MSE A  24      49.998  14.604   0.639  1.00 27.41           C
HETATM  173  O   MSE A  24      50.254  15.807   0.817  1.00 27.05           O
HETATM  174  CB  MSE A  24      51.654  12.812   1.171  1.00 24.82           C
HETATM  175  CG  MSE A  24      52.569  11.728   0.594  1.00 26.73           C
HETATM  176 SE   MSE A  24      53.180  10.645   2.121  1.00 28.93          SE
HETATM  177  CE  MSE A  24      54.323   9.426   1.101  1.00 28.30           C
HETATM  392  N   MSE A  54      53.686  23.740   4.714  1.00 23.28           N
HETATM  393  CA  MSE A  54      53.309  24.052   3.342  1.00 24.06           C
HETATM  394  C   MSE A  54      52.548  22.939   2.648  1.00 24.64           C
HETATM  395  O   MSE A  54      51.705  23.215   1.780  1.00 25.02           O
HETATM  396  CB  MSE A  54      54.586  24.347   2.476  1.00 26.10           C
HETATM  397  CG  MSE A  54      54.312  24.434   0.978  1.00 30.15           C
HETATM  398 SE   MSE A  54      56.079  24.574   0.127  1.00 37.36          SE
HETATM  399  CE  MSE A  54      55.422  23.946  -1.631  1.00 39.96           C
HETATM  673  N   MSE A  91      55.960  32.391  15.454  1.00 27.81           N
HETATM  674  CA  MSE A  91      56.595  31.102  15.430  1.00 27.17           C
HETATM  675  C   MSE A  91      55.847  30.067  16.250  1.00 27.49           C
HETATM  676  O   MSE A  91      55.679  28.942  15.840  1.00 26.22           O
HETATM  677  CB  MSE A  91      57.999  31.261  15.980  1.00 28.69           C
HETATM  678  CG  MSE A  91      58.760  29.966  16.084  1.00 29.87           C
HETATM  679 SE   MSE A  91      59.346  29.292  14.359  1.00 31.30          SE
HETATM  680  CE  MSE A  91      61.142  30.045  14.322  1.00 32.79           C
HETATM  826  N   MSE A 109      62.958  17.773  14.896  1.00 22.27           N
HETATM  827  CA  MSE A 109      63.656  18.836  14.187  1.00 23.51           C
HETATM  828  C   MSE A 109      62.678  19.733  13.504  1.00 22.57           C
HETATM  829  O   MSE A 109      61.655  19.281  12.974  1.00 21.79           O
HETATM  830  CB  MSE A 109      64.533  18.091  13.197  1.00 25.51           C
HETATM  831  CG  MSE A 109      65.219  18.945  12.193  1.00 28.74           C
HETATM  832 SE   MSE A 109      66.871  19.558  13.016  1.00 35.13          SE
HETATM  833  CE  MSE A 109      67.723  17.806  13.495  1.00 33.55           C
HETATM  975  N   MSE A 127      63.497  29.918  17.605  1.00 27.43           N
HETATM  976  CA  MSE A 127      63.697  28.474  17.588  1.00 26.51           C
HETATM  977  C   MSE A 127      64.213  28.038  18.954  1.00 27.91           C
HETATM  978  O   MSE A 127      63.816  27.004  19.477  1.00 28.23           O
HETATM  979  CB  MSE A 127      64.725  28.059  16.545  1.00 26.20           C
HETATM  980  CG  MSE A 127      64.278  28.225  15.102  1.00 26.41           C
HETATM  981 SE   MSE A 127      62.778  26.975  14.774  1.00 28.83          SE
HETATM  982  CE  MSE A 127      62.895  27.001  12.844  1.00 24.89           C
HETATM 1148  N   MSE A 149      74.699  30.191   5.265  1.00 35.20           N
HETATM 1149  CA  MSE A 149      75.193  29.406   6.425  1.00 34.88           C
HETATM 1150  C   MSE A 149      76.242  30.160   7.216  1.00 37.39           C
HETATM 1151  O   MSE A 149      76.265  30.123   8.441  1.00 37.49           O
HETATM 1152  CB  MSE A 149      75.773  28.084   5.983  1.00 34.89           C
HETATM 1153  CG  MSE A 149      74.765  27.089   5.421  1.00 34.49           C
HETATM 1154 SE   MSE A 149      73.279  26.734   6.685  1.00 34.68          SE
HETATM 1155  CE  MSE A 149      74.183  26.267   8.364  1.00 34.22           C
HETATM 1484  N   MSE A 191      70.463  17.187   4.614  1.00 26.83           N
HETATM 1485  CA  MSE A 191      69.299  16.418   4.138  1.00 25.04           C
HETATM 1486  C   MSE A 191      69.183  16.702   2.670  1.00 24.92           C
HETATM 1487  O   MSE A 191      69.064  17.841   2.307  1.00 26.30           O
HETATM 1488  CB  MSE A 191      68.026  16.856   4.864  1.00 25.02           C
HETATM 1489  CG  MSE A 191      68.111  16.603   6.364  1.00 25.16           C
HETATM 1490 SE   MSE A 191      66.469  17.188   7.266  1.00 26.02          SE
HETATM 1491  CE  MSE A 191      65.278  15.740   6.723  1.00 24.88           C
HETATM 1659  N   MSE A 215      74.159  -0.335 -15.576  1.00 54.29           N
HETATM 1660  CA  MSE A 215      74.867  -0.454 -14.333  1.00 55.95           C
HETATM 1661  C   MSE A 215      73.970  -1.026 -13.282  1.00 53.60           C
HETATM 1662  O   MSE A 215      74.427  -1.813 -12.467  1.00 56.32           O
HETATM 1663  CB  MSE A 215      75.378   0.913 -13.935  1.00 58.37           C
HETATM 1664  CG  MSE A 215      76.243   0.823 -12.682  1.00 63.59           C
HETATM 1665 SE   MSE A 215      76.954   2.605 -12.153  1.00 70.46          SE
HETATM 1666  CE  MSE A 215      77.523   3.298 -13.919  1.00 69.74           C
HETATM 2170  N   MSE A 283      67.246  24.684  -6.383  1.00 39.49           N
HETATM 2171  CA  MSE A 283      66.110  24.653  -5.428  1.00 45.39           C
HETATM 2172  C   MSE A 283      64.981  24.319  -6.331  1.00 50.57           C
HETATM 2173  O   MSE A 283      65.045  24.576  -7.522  1.00 50.98           O
HETATM 2174  CB  MSE A 283      65.689  25.947  -4.717  1.00 48.12           C
HETATM 2175  CG  MSE A 283      66.745  26.547  -3.818  1.00 50.80           C
HETATM 2176 SE   MSE A 283      66.240  28.417  -3.402  1.00 68.58          SE
HETATM 2177  CE  MSE A 283      67.943  29.230  -2.796  1.00 65.89           C
HETATM 2185  PA  NAP A 301      61.935  12.650  -5.038  1.00 31.18           P
HETATM 2186  O1A NAP A 301      63.309  12.825  -5.631  1.00 29.86           O
HETATM 2187  O2A NAP A 301      61.870  12.104  -3.623  1.00 33.14           O
HETATM 2188  O5B NAP A 301      60.999  11.823  -6.050  1.00 32.15           O
HETATM 2189  C5B NAP A 301      59.638  11.514  -5.805  1.00 29.43           C
HETATM 2190  C4B NAP A 301      59.120  10.668  -6.952  1.00 31.90           C
HETATM 2191  O4B NAP A 301      59.910   9.478  -7.027  1.00 29.35           O
HETATM 2192  C3B NAP A 301      57.691  10.266  -6.699  1.00 31.89           C
HETATM 2193  O3B NAP A 301      56.840  10.444  -7.818  1.00 32.95           O
HETATM 2194  C2B NAP A 301      57.782   8.784  -6.399  1.00 30.63           C
HETATM 2195  O2B NAP A 301      56.709   8.012  -6.895  1.00 30.03           O
HETATM 2196  C1B NAP A 301      59.028   8.351  -7.081  1.00 32.13           C
HETATM 2197  N9A NAP A 301      59.763   7.252  -6.447  1.00 31.95           N
HETATM 2198  C8A NAP A 301      60.156   7.138  -5.143  1.00 31.71           C
HETATM 2199  N7A NAP A 301      60.862   6.062  -4.923  1.00 29.14           N
HETATM 2200  C5A NAP A 301      60.904   5.430  -6.148  1.00 32.14           C
HETATM 2201  C6A NAP A 301      61.470   4.211  -6.572  1.00 32.00           C
HETATM 2202  N6A NAP A 301      62.170   3.358  -5.778  1.00 32.97           N
HETATM 2203  N1A NAP A 301      61.309   3.881  -7.880  1.00 34.46           N
HETATM 2204  C2A NAP A 301      60.659   4.683  -8.735  1.00 30.14           C
HETATM 2205  N3A NAP A 301      60.070   5.848  -8.436  1.00 33.94           N
HETATM 2206  C4A NAP A 301      60.233   6.158  -7.110  1.00 31.79           C
HETATM 2207  O3  NAP A 301      61.244  14.101  -5.019  1.00 29.96           O
HETATM 2208  PN  NAP A 301      60.623  15.024  -3.871  1.00 30.09           P
HETATM 2209  O1N NAP A 301      59.365  14.375  -3.310  1.00 30.88           O
HETATM 2210  O2N NAP A 301      60.436  16.483  -4.394  1.00 29.63           O
HETATM 2211  O5D NAP A 301      61.805  15.202  -2.807  1.00 28.39           O
HETATM 2212  C5D NAP A 301      61.933  14.401  -1.622  1.00 29.11           C
HETATM 2213  C4D NAP A 301      61.369  15.166  -0.405  1.00 26.69           C
HETATM 2214  O4D NAP A 301      62.178  16.294  -0.048  1.00 25.53           O
HETATM 2215  C3D NAP A 301      59.994  15.759  -0.541  1.00 25.73           C
HETATM 2216  O3D NAP A 301      58.913  14.833  -0.593  1.00 25.27           O
HETATM 2217  C2D NAP A 301      59.957  16.709   0.640  1.00 27.78           C
HETATM 2218  O2D NAP A 301      59.745  15.961   1.862  1.00 23.52           O
HETATM 2219  C1D NAP A 301      61.379  17.266   0.570  1.00 26.20           C
HETATM 2220  N1N NAP A 301      61.496  18.505  -0.245  1.00 29.03           N
HETATM 2221  C2N NAP A 301      61.932  19.600   0.376  1.00 26.91           C
HETATM 2222  C3N NAP A 301      62.063  20.826  -0.278  1.00 27.45           C
HETATM 2223  C7N NAP A 301      62.538  22.026   0.459  1.00 29.64           C
HETATM 2224  O7N NAP A 301      62.446  23.153  -0.082  1.00 29.10           O
HETATM 2225  N7N NAP A 301      63.106  21.838   1.663  1.00 25.30           N
HETATM 2226  C4N NAP A 301      61.790  20.850  -1.655  1.00 26.27           C
HETATM 2227  C5N NAP A 301      61.366  19.691  -2.310  1.00 29.23           C
HETATM 2228  C6N NAP A 301      61.255  18.502  -1.596  1.00 28.64           C
HETATM 2229  P2B NAP A 301      55.605   7.331  -5.879  1.00 32.53           P
HETATM 2230  O1X NAP A 301      55.233   8.354  -4.859  1.00 30.76           O
HETATM 2231  O2X NAP A 301      54.440   6.914  -6.666  1.00 32.30           O
HETATM 2232  O3X NAP A 301      56.284   6.181  -5.163  1.00 33.49           O
HETATM 2233  C   ACT A 302      59.064  22.186  -0.877  1.00 36.10           C
HETATM 2234  O   ACT A 302      59.196  23.284  -1.495  1.00 33.49           O
HETATM 2235  OXT ACT A 302      59.498  22.041   0.318  1.00 32.47           O
HETATM 2236  CH3 ACT A 302      58.358  21.069  -1.608  1.00 35.96           C
```

5. Observed PDB structure files for CASP11 target T0863

File has been truncated to show just the HETATOMS

```
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE ATOM/HETATM

HETATM11080 NA    NA A 801     -23.927 -59.285 -21.780  1.00 12.66          NA
HETATM11081 CL    CL A 802     -20.283 -57.183 -19.864  1.00 29.81          CL
HETATM11082  CHA HEM A 803     -10.679 -61.506 -40.506  1.00 12.80           C
HETATM11083  CHB HEM A 803      -9.943 -65.374 -37.746  1.00 13.61           C
HETATM11084  CHC HEM A 803     -14.708 -65.999 -37.411  1.00 13.51           C
HETATM11085  CHD HEM A 803     -15.411 -62.179 -40.211  1.00 12.26           C
HETATM11086  C1A HEM A 803     -10.076 -62.388 -39.671  1.00 13.48           C
HETATM11087  C2A HEM A 803      -8.671 -62.448 -39.371  1.00 13.85           C
HETATM11088  C3A HEM A 803      -8.482 -63.583 -38.642  1.00 13.79           C
HETATM11089  C4A HEM A 803      -9.739 -64.190 -38.478  1.00 14.41           C
HETATM11090  CMA HEM A 803      -7.230 -64.093 -38.003  1.00 13.65           C
HETATM11091  CAA HEM A 803      -7.606 -61.438 -39.741  1.00 13.52           C
HETATM11092  CBA HEM A 803      -7.603 -60.140 -38.914  1.00 13.68           C
HETATM11093  CGA HEM A 803      -6.624 -59.197 -39.564  1.00 14.16           C
HETATM11094  O1A HEM A 803      -5.386 -59.536 -39.602  1.00 14.01           O
HETATM11095  O2A HEM A 803      -6.930 -58.053 -40.036  1.00 14.71           O
HETATM11096  C1B HEM A 803     -11.189 -65.874 -37.414  1.00 14.71           C
HETATM11097  C2B HEM A 803     -11.381 -67.106 -36.684  1.00 14.53           C
HETATM11098  C3B HEM A 803     -12.744 -67.241 -36.545  1.00 14.93           C
HETATM11099  C4B HEM A 803     -13.344 -66.152 -37.342  1.00 13.79           C
HETATM11100  CMB HEM A 803     -10.270 -67.915 -36.069  1.00 15.03           C
HETATM11101  CAB HEM A 803     -13.545 -68.323 -35.933  1.00 14.94           C
HETATM11102  CBB HEM A 803     -13.022 -69.494 -35.907  1.00 16.13           C
HETATM11103  C1C HEM A 803     -15.370 -64.988 -38.062  1.00 13.45           C
HETATM11104  C2C HEM A 803     -16.786 -64.739 -38.028  1.00 12.91           C
HETATM11105  C3C HEM A 803     -16.954 -63.587 -38.761  1.00 12.97           C
HETATM11106  C4C HEM A 803     -15.669 -63.129 -39.259  1.00 13.12           C
HETATM11107  CMC HEM A 803     -17.812 -65.497 -37.176  1.00 13.19           C
```

```
HETATM11108  CAC HEM A 803     -18.195 -62.901 -39.169  1.00 13.69           C
HETATM11109  CBC HEM A 803     -19.374 -63.500 -39.229  1.00 13.51           C
HETATM11110  C1D HEM A 803     -14.137 -61.764 -40.644  1.00 12.35           C
HETATM11111  C2D HEM A 803     -13.947 -60.592 -41.488  1.00 11.91           C
HETATM11112  C3D HEM A 803     -12.614 -60.432 -41.586  1.00 12.08           C
HETATM11113  C4D HEM A 803     -12.036 -61.447 -40.744  1.00 12.57           C
HETATM11114  CMD HEM A 803     -14.959 -59.866 -42.331  1.00 12.10           C
HETATM11115  CAD HEM A 803     -11.858 -59.510 -42.477  1.00 12.20           C
HETATM11116  CBD HEM A 803     -11.389 -58.169 -41.897  1.00 11.96           C
HETATM11117  CGD HEM A 803     -10.991 -57.146 -42.916  1.00 11.53           C
HETATM11118  O1D HEM A 803     -10.327 -56.143 -42.481  1.00 11.07           O
HETATM11119  O2D HEM A 803     -11.340 -57.228 -44.141  1.00 11.34           O
HETATM11120  NA  HEM A 803     -10.690 -63.476 -39.155  1.00 13.40           N
HETATM11121  NB  HEM A 803     -12.384 -65.401 -37.807  1.00 13.96           N
HETATM11122  NC  HEM A 803     -14.729 -64.008 -38.792  1.00 12.93           N
HETATM11123  ND  HEM A 803     -12.986 -62.179 -40.133  1.00 12.68           N
HETATM11124 FE   HEM A 803     -12.679 -63.727 -39.002  1.00 14.09          FE
HETATM11125  C   CYN A 804     -12.592 -62.336 -37.614  1.00 15.82           C
HETATM11126  N   CYN A 804     -12.182 -62.237 -36.426  1.00 15.78           N
HETATM11127  C1  MPD A 805     -21.435 -35.377 -42.546  1.00 40.96           C
HETATM11128  C2  MPD A 805     -21.088 -36.857 -42.681  1.00 39.54           C
HETATM11129  O2  MPD A 805     -21.079 -37.561 -41.407  1.00 36.46           O
HETATM11130  CM  MPD A 805     -22.142 -37.577 -43.459  1.00 38.33           C
HETATM11131  C3  MPD A 805     -19.654 -36.959 -43.220  1.00 42.56           C
HETATM11132  C4  MPD A 805     -19.239 -36.363 -44.596  1.00 43.88           C
HETATM11133  O4  MPD A 805     -20.050 -37.069 -45.524  1.00 46.78           O
HETATM11134  C5  MPD A 805     -19.293 -34.828 -44.781  1.00 43.24           C
HETATM11135  C1  MPD A 806      -1.538 -57.208 -31.129  1.00 43.02           C
HETATM11136  C2  MPD A 806      -0.230 -57.938 -31.407  1.00 44.51           C
HETATM11137  O2  MPD A 806       0.027 -57.855 -32.788  1.00 42.26           O
HETATM11138  CM  MPD A 806       0.926 -57.203 -30.737  1.00 45.08           C
HETATM11139  C3  MPD A 806      -0.256 -59.449 -31.123  1.00 47.35           C
HETATM11140  C4  MPD A 806      -0.577 -60.348 -32.339  1.00 47.33           C
HETATM11141  O4  MPD A 806      -0.593 -59.585 -33.517  1.00 46.79           O
HETATM11142  C5  MPD A 806       0.427 -61.443 -32.713  1.00 49.63           C
HETATM11143  CHA HEM B 801      -2.291 -44.158  24.591  1.00 12.59           C
HETATM11144  CHB HEM B 801       1.341 -42.748  21.834  1.00 13.08           C
HETATM11145  CHC HEM B 801      -0.632 -38.381  21.352  1.00 12.48           C
HETATM11146  CHD HEM B 801      -4.250 -39.852  24.223  1.00 12.56           C
HETATM11147  C1A HEM B 801      -1.200 -44.202  23.766  1.00 12.75           C
HETATM11148  C2A HEM B 801      -0.411 -45.369  23.441  1.00 12.43           C
HETATM11149  C3A HEM B 801       0.620 -44.965  22.681  1.00 12.69           C
HETATM11150  C4A HEM B 801       0.501 -43.516  22.577  1.00 12.98           C
HETATM11151  CMA HEM B 801       1.744 -45.752  22.074  1.00 12.34           C
HETATM11152  CAA HEM B 801      -0.589 -46.783  23.816  1.00 12.75           C
HETATM11153  CBA HEM B 801      -1.668 -47.466  23.012  1.00 13.19           C
HETATM11154  CGA HEM B 801      -1.930 -48.819  23.624  1.00 13.11           C
HETATM11155  O1A HEM B 801      -0.981 -49.659  23.697  1.00 12.89           O
HETATM11156  O2A HEM B 801      -3.141 -49.117  23.950  1.00 13.05           O
HETATM11157  C1B HEM B 801       1.161 -41.390  21.499  1.00 13.63           C
```

```
HETATM11158  C2B HEM B 801       2.036 -40.591  20.695  1.00 13.79           C
HETATM11159  C3B HEM B 801       1.493 -39.352  20.556  1.00 13.84           C
HETATM11160  C4B HEM B 801       0.219 -39.442  21.258  1.00 13.55           C
HETATM11161  CMB HEM B 801       3.320 -40.992  20.011  1.00 13.92           C
HETATM11162  CAB HEM B 801       1.893 -38.097  19.887  1.00 14.25           C
HETATM11163  CBB HEM B 801       3.132 -37.733  19.902  1.00 15.32           C
HETATM11164  C1C HEM B 801      -1.854 -38.406  22.027  1.00 12.30           C
HETATM11165  C2C HEM B 801      -2.867 -37.381  21.986  1.00 11.26           C
HETATM11166  C3C HEM B 801      -3.962 -37.808  22.776  1.00 11.25           C
HETATM11167  C4C HEM B 801      -3.531 -39.117  23.296  1.00 12.03           C
HETATM11168  CMC HEM B 801      -2.747 -36.136  21.116  1.00 11.44           C
HETATM11169  CAC HEM B 801      -5.260 -37.180  23.180  1.00 10.71           C
HETATM11170  CBC HEM B 801      -5.369 -35.829  23.200  1.00 10.44           C
HETATM11171  C1D HEM B 801      -3.958 -41.132  24.609  1.00 12.35           C
HETATM11172  C2D HEM B 801      -4.766 -41.885  25.556  1.00 12.15           C
HETATM11173  C3D HEM B 801      -4.203 -43.113  25.638  1.00 12.32           C
HETATM11174  C4D HEM B 801      -3.070 -43.076  24.729  1.00 12.41           C
HETATM11175  CMD HEM B 801      -6.006 -41.475  26.296  1.00 11.91           C
HETATM11176  CAD HEM B 801      -4.646 -44.262  26.507  1.00 12.53           C
HETATM11177  CBD HEM B 801      -5.472 -45.343  25.944  1.00 12.52           C
HETATM11178  CGD HEM B 801      -6.020 -46.308  26.947  1.00 12.58           C
HETATM11179  O1D HEM B 801      -6.437 -47.463  26.479  1.00 12.95           O
HETATM11180  O2D HEM B 801      -6.130 -46.017  28.152  1.00 12.60           O
HETATM11181  NA  HEM B 801      -0.649 -43.093  23.231  1.00 12.68           N
HETATM11182  NB  HEM B 801       0.120 -40.685  21.844  1.00 13.51           N
HETATM11183  NC  HEM B 801      -2.276 -39.429  22.800  1.00 12.05           N
HETATM11184  ND  HEM B 801      -2.936 -41.826  24.147  1.00 12.74           N
HETATM11185 FE   HEM B 801      -1.471 -41.282  23.029  1.00 14.24          FE
HETATM11186 NA    NA B 802     -11.116 -34.514   5.635  1.00 13.26          NA
HETATM11187 CL    CL B 803     -10.971 -38.874   3.738  1.00 31.84          CL
HETATM11188  C   CYN B 804      -2.649 -41.945  21.510  1.00 16.68           C
HETATM11189  N   CYN B 804      -2.833 -42.666  20.470  1.00 17.78           N
HETATM11190  C1  MPD B 805     -11.073  -4.222   1.524  1.00 44.03           C
HETATM11191  C2  MPD B 805      -9.714  -3.743   0.982  1.00 48.25           C
HETATM11192  O2  MPD B 805      -9.631  -4.172  -0.409  1.00 47.44           O
HETATM11193  CM  MPD B 805      -9.835  -2.218   0.776  1.00 51.52           C
HETATM11194  C3  MPD B 805      -8.386  -4.082   1.849  1.00 46.84           C
HETATM11195  C4  MPD B 805      -8.446  -4.797   3.258  1.00 47.03           C
HETATM11196  O4  MPD B 805      -7.264  -5.012   4.041  1.00 40.73           O
HETATM11197  C5  MPD B 805      -9.406  -4.134   4.246  1.00 49.52           C
HETATM11198  C1  MPD B 806       4.593 -22.740 -42.044  1.00 32.12           C
HETATM11199  C2  MPD B 806       4.714 -24.256 -41.904  1.00 33.56           C
HETATM11200  O2  MPD B 806       5.348 -24.470 -40.638  1.00 27.94           O
HETATM11201  CM  MPD B 806       3.392 -24.970 -41.809  1.00 34.61           C
HETATM11202  C3  MPD B 806       5.503 -24.804 -43.078  1.00 35.18           C
HETATM11203  C4  MPD B 806       4.777 -24.726 -44.416  1.00 38.63           C
HETATM11204  O4  MPD B 806       3.966 -25.866 -44.504  1.00 41.30           O
HETATM11205  C5  MPD B 806       3.723 -23.649 -44.786  1.00 40.26           C
HETATM11206  C1  MPD B 807       1.589 -54.775  17.947  1.00 45.47           C
HETATM11207  C2  MPD B 807       2.349 -54.050  16.859  1.00 43.67           C
```

```
HETATM11208  O2  MPD B 807       2.566 -52.694  17.116  1.00 43.71           O
HETATM11209  CM  MPD B 807       3.751 -54.630  16.848  1.00 46.21           C
HETATM11210  C3  MPD B 807       1.714 -54.026  15.478  1.00 43.93           C
HETATM11211  C4  MPD B 807       0.304 -54.546  15.373  1.00 40.10           C
HETATM11212  O4  MPD B 807       0.301 -55.883  15.748  1.00 40.44           O
HETATM11213  C5  MPD B 807      -0.719 -53.887  16.276  1.00 36.58           C
HETATM11214  P   PO4 B 808       7.117 -52.019  38.412  1.00 53.67           P
HETATM11215  O1  PO4 B 808       6.444 -52.053  37.032  1.00 51.08           O
HETATM11216  O2  PO4 B 808       6.145 -52.673  39.377  1.00 57.19           O
HETATM11217  O3  PO4 B 808       7.474 -50.618  38.830  1.00 40.70           O
HETATM11218  O4  PO4 B 808       8.355 -52.923  38.440  1.00 56.58           O
```

Analysis of CASP11, CASP12 and CASP13

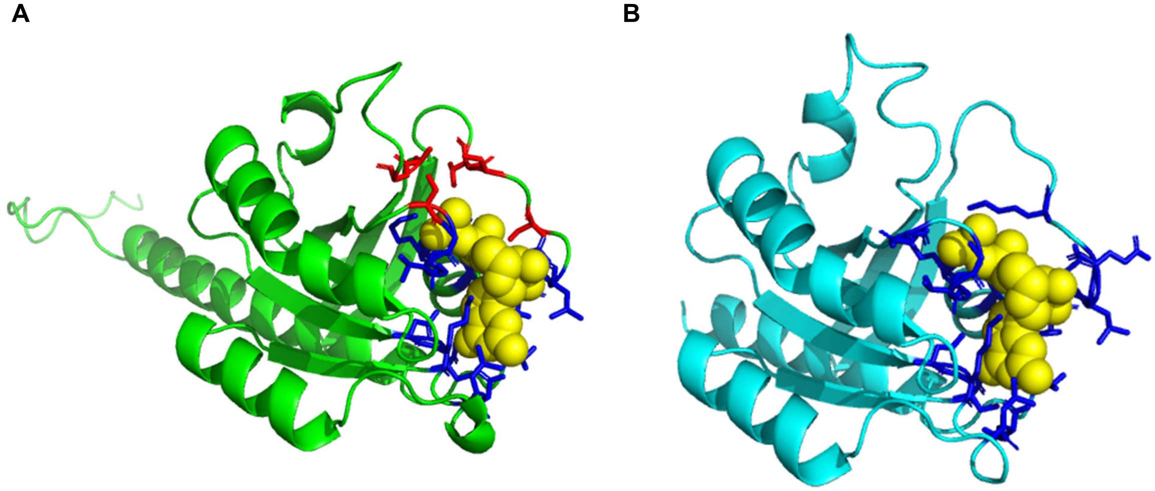


Figure S.1. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0798 (PDB ID 4ojk). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the guanosine-5'-diphosphate (GDP) ligand is shown as a sphere and coloured yellow. BDT score of 0.797 and MCC score of 0.753. **(B)** The observed ligand binding site residues shown as sticks for T0798 (PDB ID 4ojk), with binding site residues coloured in blue and the ligand GDP coloured yellow

The fourth CASP11 target is cGMP Dependent Protein Kinase II from *Rattus norvegicus.* (CASP 11 T0798 and PDB ID 4ojk). There were a total of nine incorrect predictions, which consisted of three underpredictions (ALA 12, TYR 32 and ARG 32) and six overpredictions (VAL 13, GLU 33, ILE 35, MET 36, ARG 61, GLU 62). A majority of under- and over-predictions were caused by extension of the ligand binding site, due to having a large ligand; guanosine-5'-diphosphate (GDP) with a large binding site.

Cyclic GMP-dependent protein kinases (PKG) are key mediators of the nitric oxide/cGMP signaling pathway and play a key role in the regulation of cardiovascular and neuronal functions. Type II is a membrane-anchored PKG.(Seifried, Schultz and Gohla, 2013)

A TM-score of 0.88190 was obtained, showing a high level of molecular similarity between the observed and predicted molecular structures. The number of residues obtained for the observed protein model, were 172 residues compared to 198 residues for the predicted protein model. Despite the difference in the number of residues, there was still a high level of molecular similarity. This was because the predicted protein model had extra residues; which made the α helix of the model longer and added a flexible loop. The rest of the predicted protein model aligns with the observed protein model.

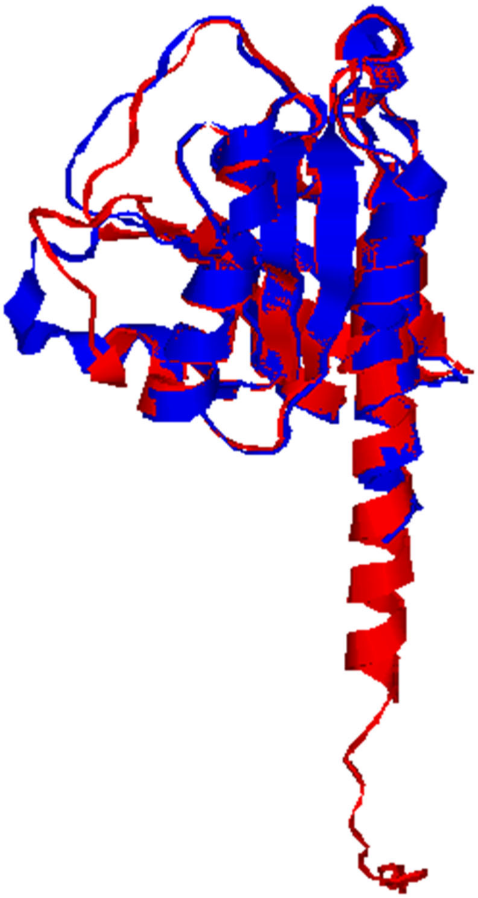


Figure S.2. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP11 target T0798 (PDB ID 4ojk). The structure in blue is the observed structure for T0798 and the structure in red is the predicted structure from InFOLD3(McGuffin *et al.*, 2015)

\

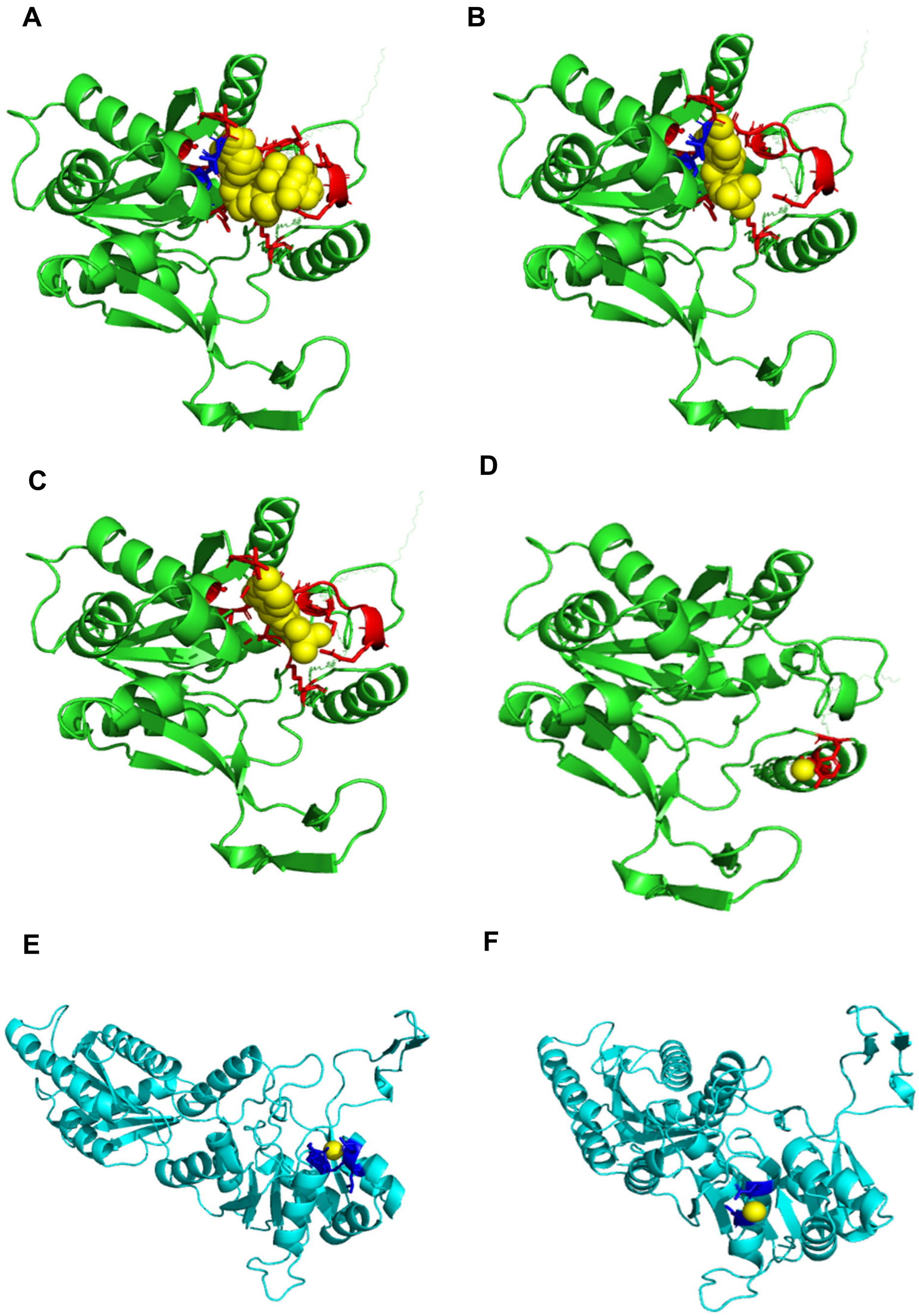


Figure S.3. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0783 (PDB ID 4cvh). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the cytidine-5'-triphosphate (CTP) ligand is shown as a sphere and coloured yellow. BDT and MCC score of 0.17 and 0.21, respectively and compared against the MG ligand due to correctly predicted residues against this observed ligand **(B)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the carbon (C) ligand is shown as a sphere and coloured yellow. BDT and MCC score of 0.17 and 0.21, respectively and compared against the MG ligand due to correctly predicted residues against this observed ligand **(C)** Predicted ligand binding site residues shown as sticks with and incorrect predictions in red, the cytidine-5'-monophosphate (C5P) ligand is shown as a sphere and coloured yellow. No BDT or MCC score could be calculated at the predicted ligand does not match the observed ligand **(D)** Predicted ligand binding site residues shown as sticks with and incorrect predictions in red, the copper (CU) ligand is shown as a sphere and coloured yellow. No BDT or MCC score could be calculated at the predicted ligand does not match the observed ligand **(E)** The observed ligand binding site residues shown as sticks for T0783 (PDB ID 4cvh), with binding site residues coloured in blue and the magnesium (MG) ligand coloured yellow **(F)** The observed ligand

binding site residues shown as sticks for T0783 (PDB ID 4cvh), with binding site residues coloured in blue and the chlorine (CL) ligand coloured yellow

The fifth CASP11 target is human isoprenoid synthase (CASP 11 T0783 and PDB ID 4cvh). There were only two correct predictions, which were THR 85 and ARG 86 (Figure S.3A and B). These correct predictions were on two (CTP and C) of the four predicted ligands. Refer to Figure S.3 for the location and labels for the associated predicted residues for this protein.

Isoprenoid synthase utilises the magnesium as a cofactor for the dissociation of the enzyme to the tetra anion.(Casteel *et al.*, 2010) The FunFOLD3 server did not correctly predicted any of these ligands. According to the PDB entry for this protein, Isoprenoid synthase binds a total of three ligands the previously mentioned magnesium ligand as well as chlorine and ethylene glycol. FunFOLD3 correctly predicted more than one ligand, however not the correct ones. Despite the incorrect ligands, two of the ligands; CTP and C had two of the same ligand-binding site residues as MG and therefore BCC and MDT scores were compared against this observed ligand. Similar BDT and MCC scores were achieved for the two predicted ligands with 0.21 and 0.17, respectively for CTP ligand and 0.27 and 0.20, respectively for C ligand. This was compared against the MG ligand due to two of the same ligand binding site residues in the predicted ligands; CTP and C.

A TM-score of 0.545 was obtained, meaning generally the same fold of the protein in structural classification of proteins/CATH protein structure classification database. The number of residues for the observed protein model was 399 and for the predicted model was 411 residues.

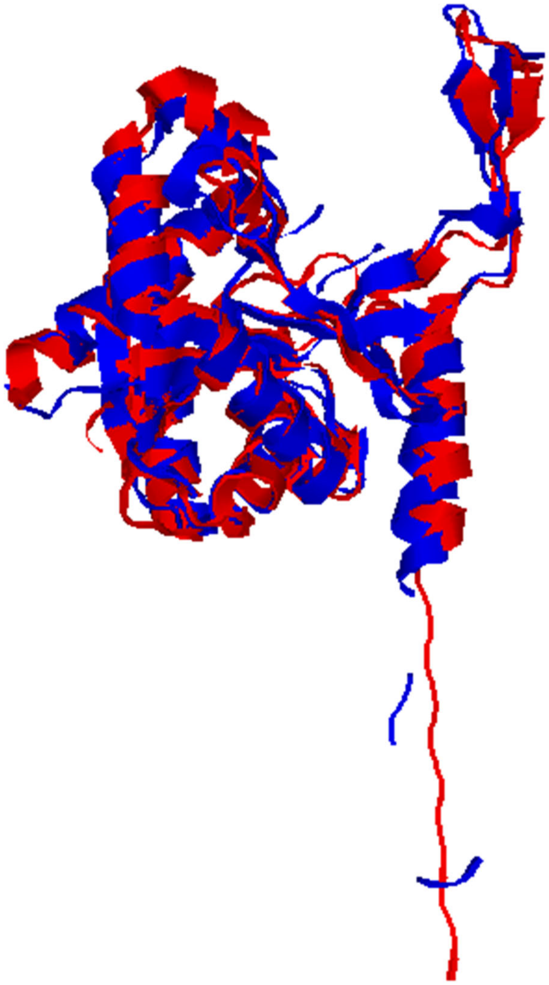


Figure S.4. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP11 target T0783 (PDB ID 4cvh). The structure in blue is the observed structure for T07983 and the structure in red is the predicted structure from InFOLD3(McGuffin *et al.*, 2015)

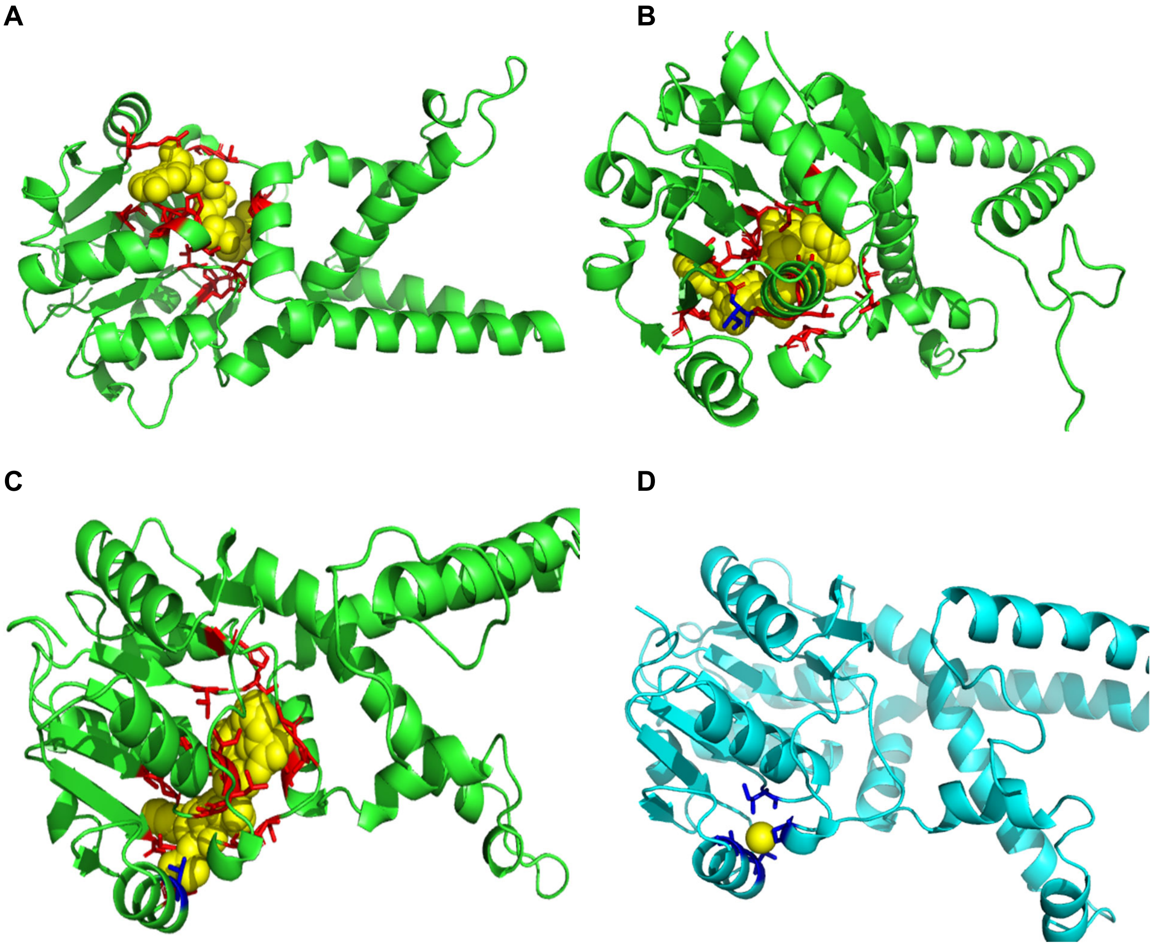


Figure S.5. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0813 (PDB ID 4wji). **(A)** Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the 1,4-dihydronicotinamide adenine dinucleotide (NAI) ligand is shown as a sphere and coloured yellow. BDT score of 0.11 and MCC score od -0.029 was achieved **(B)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and with incorrect predictions in red, the nicotinamide adenine dinucleotide (NAD) ligand is shown as a sphere and coloured yellow. BDT score of 0.086 and MCC score of 0.19 was achieved **(C)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and with incorrect predictions in red, the nicotinamide adenine dinucleotide phosphate (NAP) ligand is shown as a sphere and coloured yellow. BDT score of 0.2 and MCC score of 0.079 was achieved **(D)** The observed ligand binding site residues shown as sticks for T0813 (PDB ID 4wji), with binding site residues coloured in blue and the ligand MG coloured yellow

The sixth CASPP11 target is cyclohexadienyl dehydrogenase from *Sinorhizobium meliloti* (CASP 11 T0813 and PDB ID 4wji). There was only one correct prediction THR 42, as part of the prediction for NAD ligand (Figure S.5B).

Cyclohexadienyl dehydrogenase belongs to a family of enzymes in the tyrosine-pathway dehydrogenase in the TyrA protein family.(Park *et al.*, 2014) As with other enzymes, the magnesium ligand most likely acts as a cofactor for enzymatic reactions. The correct prediction was near the magnesium ligand. Upon investigating the PDB entry for this protein, there are a total of four ligands; NADP, tyrosine, chlorine and magnesium. FunFOLD3 predicted NAD and NAP as close ligands to NADP. As can be seen in Figure S.6, the incorrect predictions for FunFOLD3 are clustered around NADP and this could provide insight into why the incorrect predictions are in these locations. The tyrosine ligand is located directly next to NADP at residue 302 and could also be included in the prediction cluster. As a result, it would seem unreasonable to penalise FunFOLD3 for predicting this additional binding site, particularly as it is clearly clustered around an identifiable ligand.

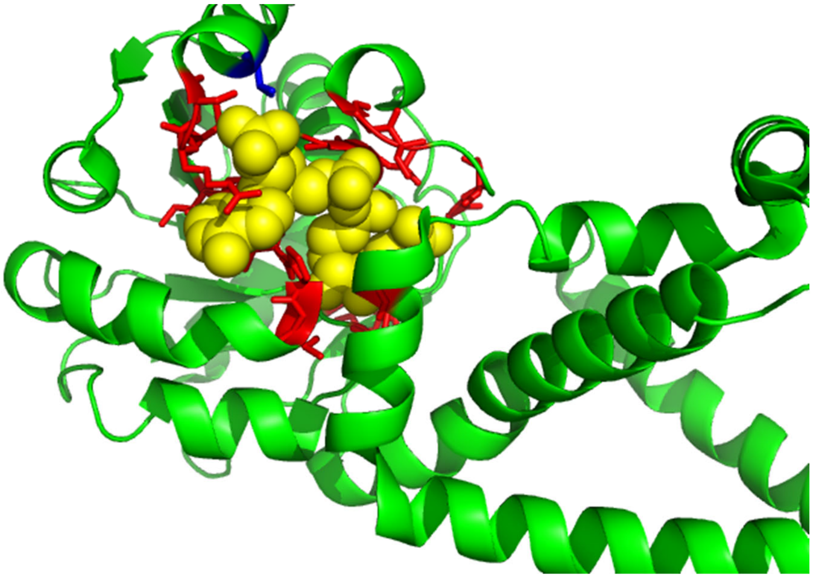


Figure S.6. NADP ligand bound to T0813 (PDB ID 4wji)
Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in *blue* and under- and over-predictions in *red*, the NADP ligand is shown as a sphere and coloured yellow

A TM-score of 0.856 was obtained suggesting a high level of structural similarity. The number of residues was close; with 302 residues for the observed protein model and 307 residues for the predicted protein model. This seems to be peculiar result, given that for the four top scoring BDT and MCC CASP11 targets with a prediction closer to perfect, as opposed to random there is also a high level of structural similarity. This shows the prediction of protein structure and ligand binding is multi-faceted. It is not just about matching the structure and assuming the ligand binding will also match. It is worth noting, that so far, that all the CASP11 targets with a close to perfect prediction also had high structural similarity, note structural similarity does not result in perfect predictions for BDT and MCC.

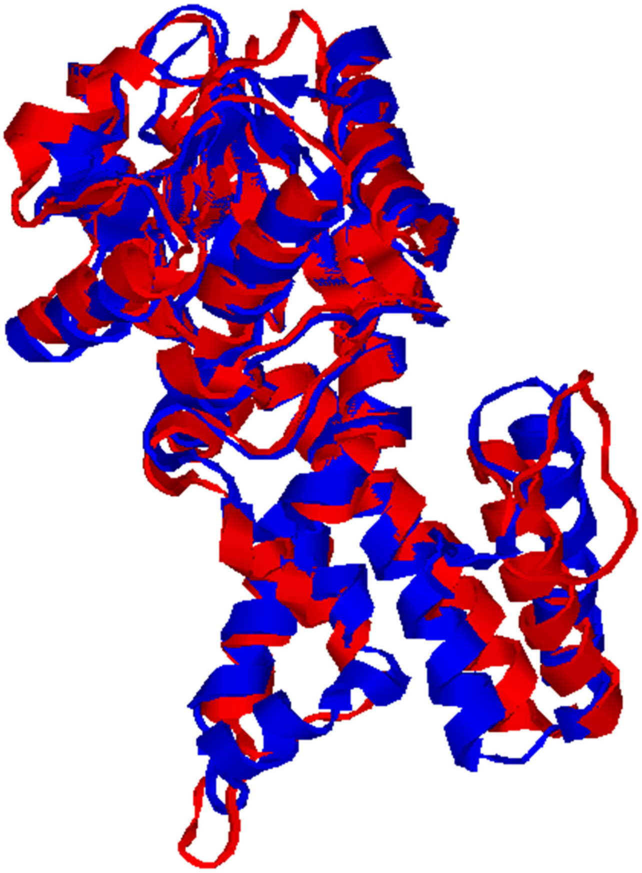


Figure S.7. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP11 target T0813 (PDB ID 4wji). The structure in blue is the observed structure for T0813 and the structure in red is the predicted structure from InFOLD3(McGuffin *et al.*, 2015)

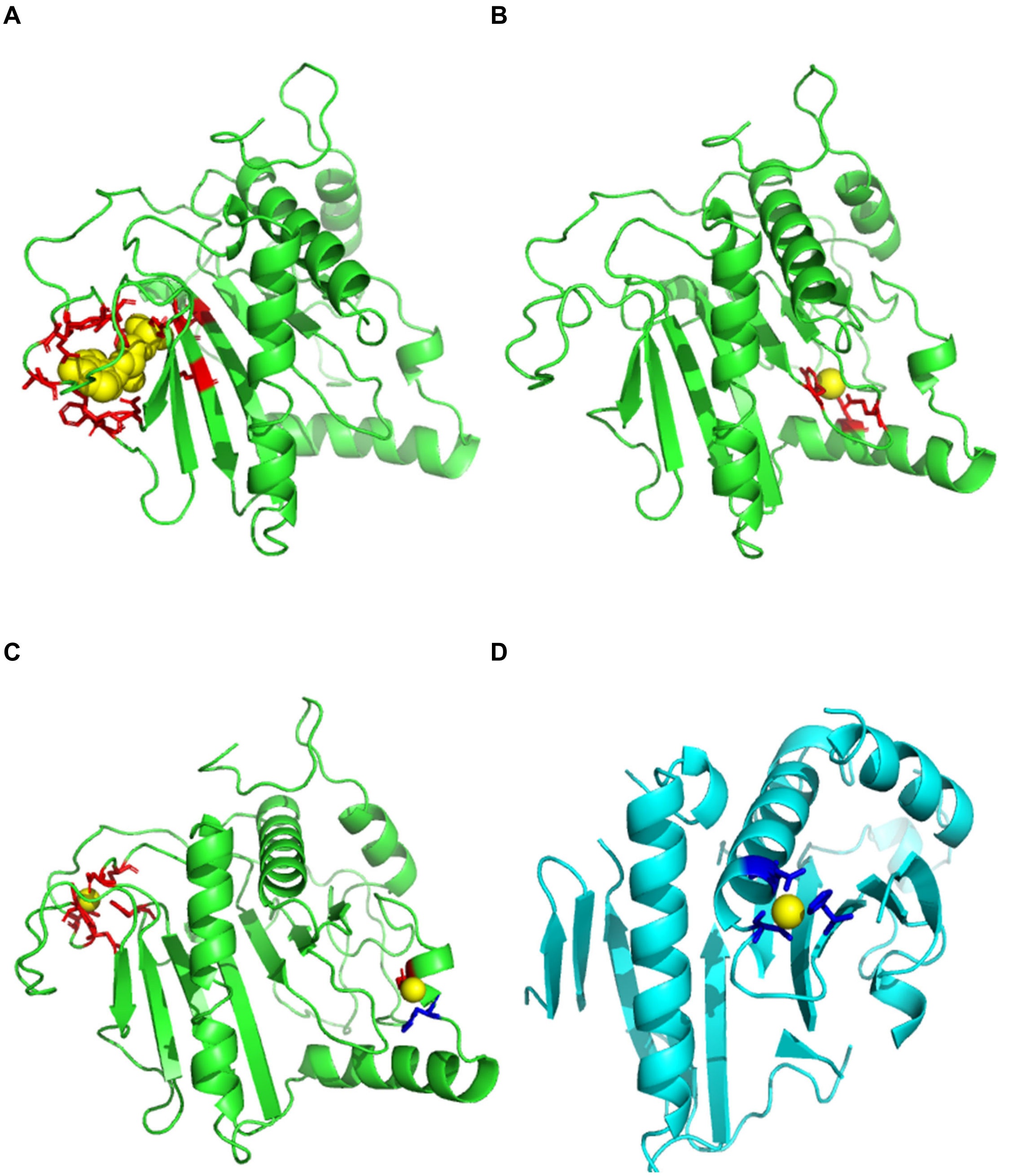


Figure S.8. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0786 (PDB ID 4qvu).
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the adenosine monophosphate (AMP) ligand is shown as a sphere and coloured yellow. No BDT and MCC scores were calculated as this ligand did not match the observed ligand **(B)** Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the iron (FE) ligand is shown as a sphere and coloured yellow. No BDT and MCC scores were calculated as this ligand did not match the observed ligand **(C)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and with incorrect predictions in red, the zinc (ZN) ligands are shown as a sphere and coloured yellow. BDT score of 0.0139 and MCC score of -0.014 was achieved for ZN ligand with no correct predictions and BDT score of 0.38 and MCC score of 0.40 for ZN ligand with one correct prediction **(D)** The observed ligand binding site residues shown as sticks for T0786 (PDB ID 4qvu), with binding site residues coloured in blue and the ligand ZN coloured yellow

The seventh CASP11 target is DUF4931 family protein (BCE0241) from *Bacillus cereus* (CASP 11 T0786 and PDB ID 4qvu). There was only one correct prediction; HIS 152.

DUF4931 binds a number of ligands; tetraethylene glycol, zinc and sodium. The FunFOLD3 server correctly identified the zinc ligand. An explanation for such a poor prediction is potentially FunFOLD3 has extended the binding site to attempt to capture the “binding” site for MSE and the two other ligands tetraethylene and sodium. For completeness, the predicted tetraethylene and sodium ligands have been illustrated in Figure S.8A to provide insight into the incorrect predictions. However, this has been missed Figure S.8B to focus on the zinc ligand binding site.

A TM-score of 0.679 was obtained; which although is within the high level of structural similarity, is still close to the lower end. The observed number of residues was 217 compared to 264 residues for the predicted protein model. The differences between the two structures are shown in Figure S9. As with other CASP11 targets (T0798 and T0854) the flexible loop on the predicted protein model is not fully aligned with the flexible loop on the observed model. Additionally, a part of the predicted protein model does not have an α helix which is present on the observed protein model and the predicted model either has an extra α helix or the α helix has not aligned with the rest of the molecule.

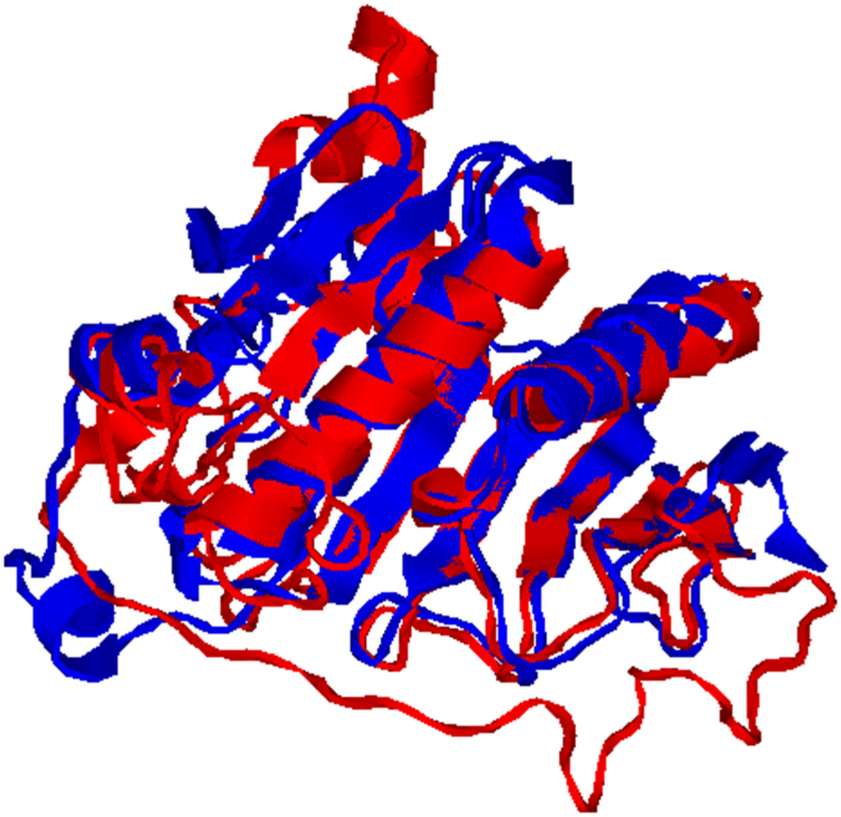


Figure S.9. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP11 target T0786 (PDB ID 4qvu). The structure in blue is the observed structure for T0786 and the structure in red is the predicted structure from IntFOLD3(McGuffin *et al.*, 2015)

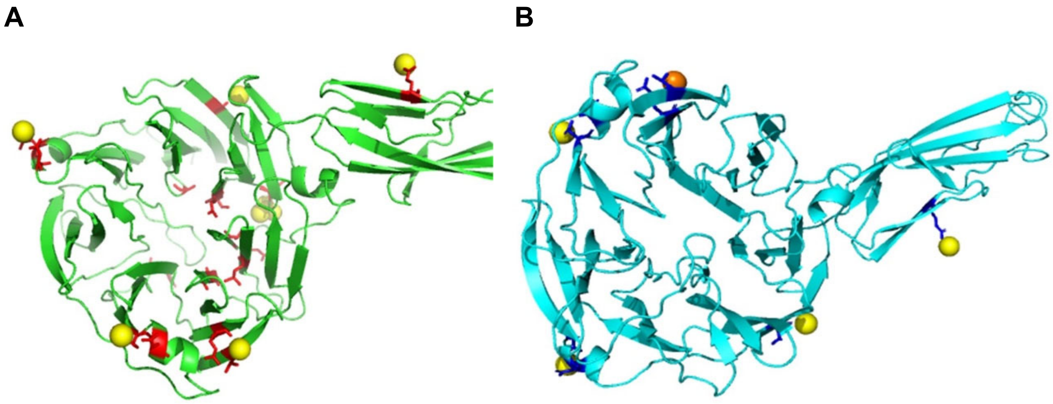


Figure S.10. Comparison of FunFOLD3 ligand binding site predictions for CASP 11 target T0845 (PDB ID 4r5o). **(A)** Predicted ligand binding site residues shown as sticks with correctly predicted binding site residues in blue and incorrect predictions in red, the chlorine (CL) ligand is shown as a sphere and coloured yellow. BDT score of 0.035 and MCC score of -0.02. **(B)** The observed ligand binding site residues shown as sticks for T0845 (PDB ID 4r5o), with binding site residues coloured in blue and the ligand CA coloured yellow and CL ligand coloured orange

The ninth CASP11 target is quinonprotein alcohol dehydrogenase-like protein (BT1487) from *Bacteroides thetaiotaomicron* (CASP 11 T0845 and PDB ID 4r5o). There were no correct predictions with this protein.

Alcohol dehydrogenase is responsible for the detoxification of ethanol to acetaldehyde and binds a number of ligands such as; calcium, chloride, polyethylene glycol and acetate. The FunFOLD3 server correctly identified calcium and chloride ligands. A explanation as to why the prediction is so poor; the server has tried to pick up all the ligands which are bound to the protein molecule, rather than having some specificity and selectivity.

A TM-score of 0.621 was obtained with 426 residues for the observed protein and 448 for the predicted protein. As Figure S.10 illustrates, the structural similarity is contained within some of the molecule. The predicted protein model has failed to form β sheets and flexible loops in a portion of the protein. The failure of the predicted protein molecule to form the entire molecule, may have contributed to no correct predictions being obtained.

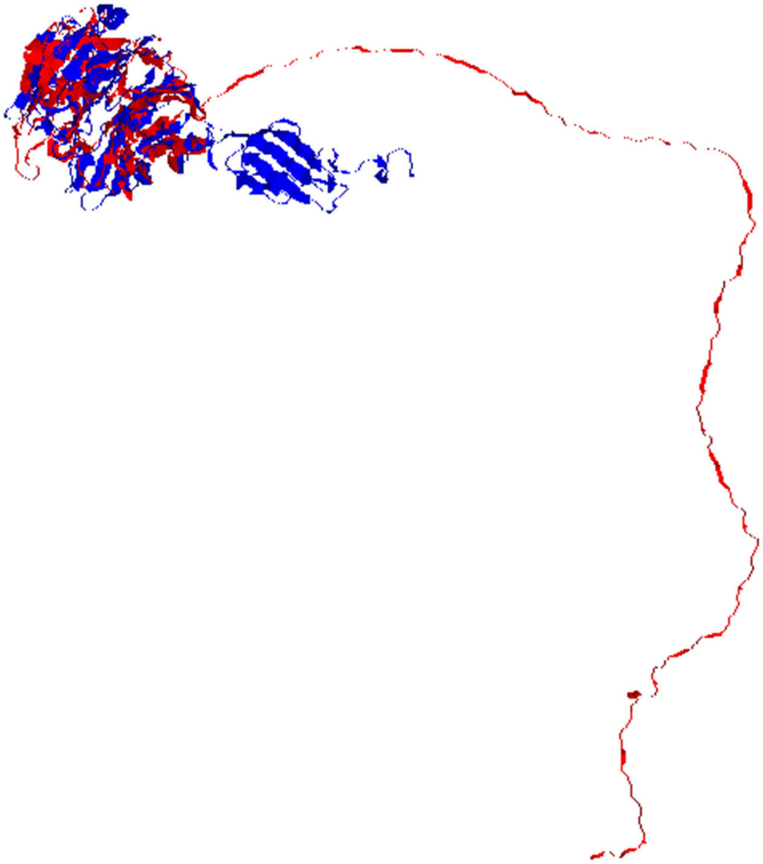


Figure S.11. Comparison of TMalign(Zhang and Skolnick, 2005) **structures for CASP11 target T0845 (PDB ID 4r5o)**
The structure in blue is the observed structure for T0845 and the structure in red is the predicted structure from InFOLD3

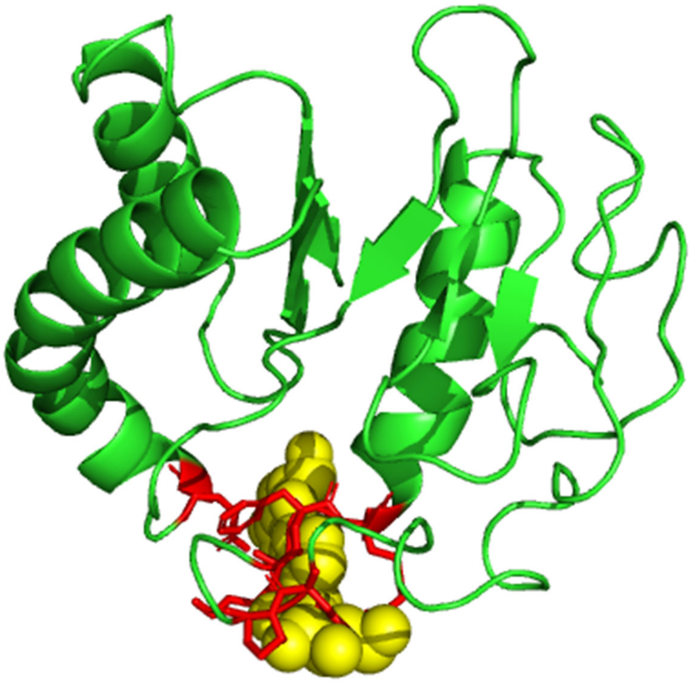


Figure S.12. FunFOLD3 ligand binding site predictions for CASP12 target T0868 (PDB ID 5j4a)
Predicted ligand binding site residues are shown as sticks and coloured red, the predicted ligand beta-d-glucose (BCG) is shown as spheres and coloured yellow.

The first predicted CASP12 target is CdiA-CT/CdiI-SU1 from *Burkholderia pseudomallei* (CASP ID T0868 and PDB ID 5j4a). As can be seen from Figure S.12, there is no image related to correct ligand-binding site residues. This is because, whilst FunFOLD3 predicted ligands on the predicted structure, there are no ligands associated with the observed protein and this is supported by the PDB entry. Further work will be done to determine the biological relevance of these predicted ligands.

CdiA-CT/CdiI-SU1 are responsible for bacterial contact-dependent growth inhibition (CDI) and these genes encoding CDI systems are distributed throughout α- β- γ-proteobacteria and are commonly found in human pathogens such as, enterohaemorrhagic Escherichia coli, Neisseria meningitidis, Pseudomonas aeruginosa and Burkholderia pseudomallei.(Lan *et al.*, 2016) As previously mentioned, T0868 is from Burkholderia pseudomallei upon researching the other closely aligned proteobacteria, such as CdiA-CT/CdiI toxin and immunity complex from Escherichia coli (PDB ID 4g6u) ligands are identified. The ligands identified with PDB ID 4g6u are YT3 (yttrium ion), ZN (zinc ion), ACT (acetate ion) and CL (chloride ion), FunFOLD3 identified BGC (beta-d-glucose) for T0868. As this wasn't identified in the functional text file, ZN and GDP might not have been biologically relevant ligands for this target. It could be likely that ZN is a biologically relevant ligand, based on sharing some commonality with PDB ID 4g6u additionally GDP also has a role as a ligand.

The TM-score for the predicted CASP12 target T0868 is 0.566 compared to the actual protein structure is, illustrating structural similarity, the TMalign structures are shown in Figure S.13. As mentioned previously, FunFOLD works on the basis that proteins with a similar structure will bind similar ligands and T0868 (PDB ID 5j4a) and 4g6u have a TM score of 0.46718 illustrating that the structures have a degree of homology. Thus, an explanation for why FunFOLD3 predicted ligands for this CASP12 target could be the incorrect template identification of a similar protein with biologically relevant ligands and this led to ligands being incorrectly identified. As a result of having no actual ligand binding residues, no MCC/BDT score can be obtained.

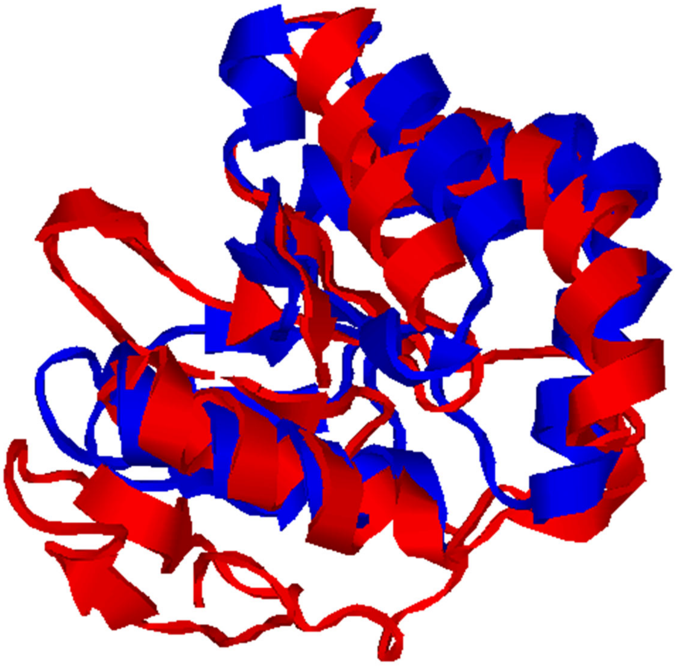


Figure S.13. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP12 target T0868 (PDB ID 5j4a) The structure in blue is the observed structure for T0868 and the structure in red is the predicted structure from IntFOL.D4. A TM-score of 0.566 was achieved for the protein structures. The score was normalised for the observed structure as it is the reference molecule

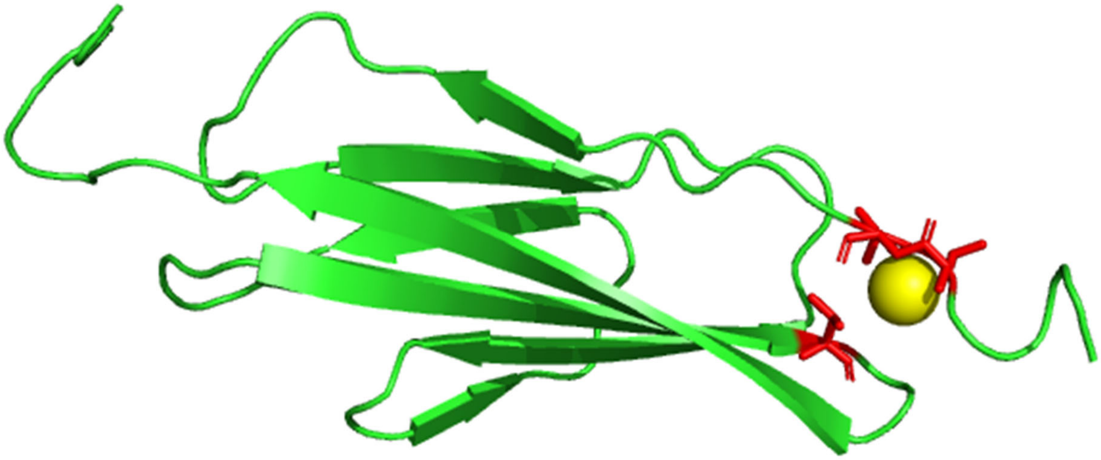


Figure S.14. FunFOLD3 ligand binding site predictions for CASP12 target T0872 (PDB ID 5jmb)
Predicted ligand binding site residues are shown as sticks and coloured red the predicted ligand calcium (CA) is shown as sphere and coloured yellow

The second predicted CASP12 target with predicted ligand-binding site residues is novel cellulases from *Bacteroides coprocola* (CASP ID T0872 and PDB ID 5jmb). As can be seen from Figure S.14, there is no image related to correct ligand-binding site residues. This is because, whilst FunFOLD3 predicted ligands, there are no ligands associated with the observed protein and this is further supported by the PDB entry for the target.

Bacteroides are a genus of gram-negative, anaerobic rod a bacterium, which are isolated from human faeces and is one of the predominant genera in human faeces.(Johnson *et al.*, 2016) rRNA gene sequence clone libraries have shown that may novel phyloptypes in the Bacteroides genus exist and up until 2015 *B.coprocola* was an unknown species.(Kitahara, 2005)

Further literature search on *B.coprocola* has shown that specific strains of this bacterium may be associated with type II diabetes and findings from a study investigating the gut metagenomes in type II diabetic patients, suggest a potential restorative influence of probiotic supplements could be further investigated.(Kitahara, 2005) Additionally, it is worth noting if there are any ligands with may be on benefit in these patients. Whilst no ligand has been identified on PDB, it could be due to the limited data available on this protein and this protein could be potentially studied further in docking experiments to provide insight into ligands, which bind and thus provide further insight into its function in type II diabetes. Data on this protein is so scarce, that on PDB the only publication related to this protein is from a group whom participated in CASP12 and to date, the publication has yet to be published.

The TM-score for the predicted CASP12 target T0868 compared the actual protein structure is 0.680 illustrating structural similarity, the TMalign superposition of observed and predicted structures is shown in Figure S.15. As a result of having no actual ligand binding residues, no MCC/BDT score can be obtained.

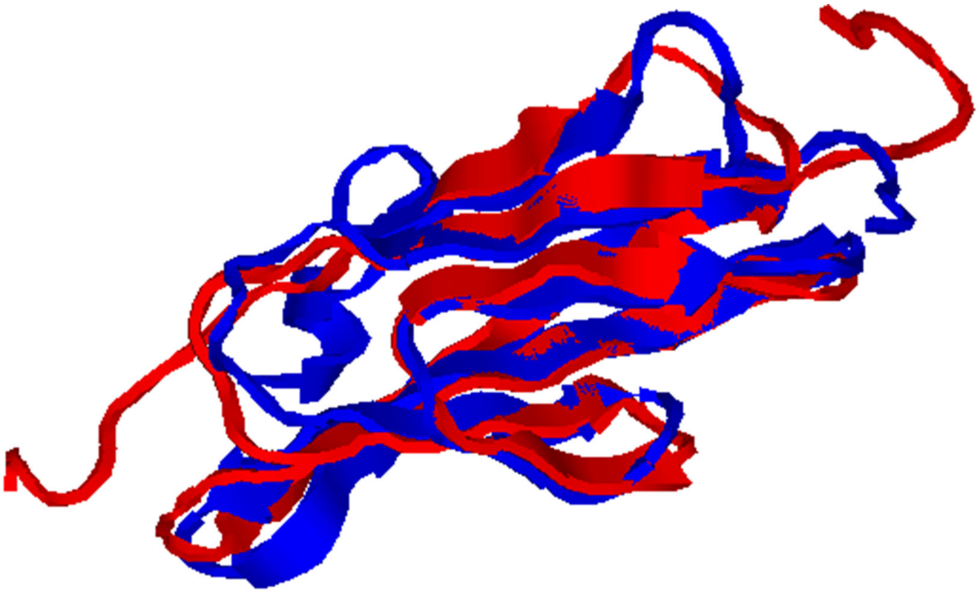


Figure S.15. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP11 target T0872 (PDB ID 5jmb) The structure in blue is the observed structure for T0872 and the structure in red is the predicted structure from IntFOLD4. A TM-align of 0.680 was achieved for the protein structures. The score was normalised for the observed structure as it is the reference molecule

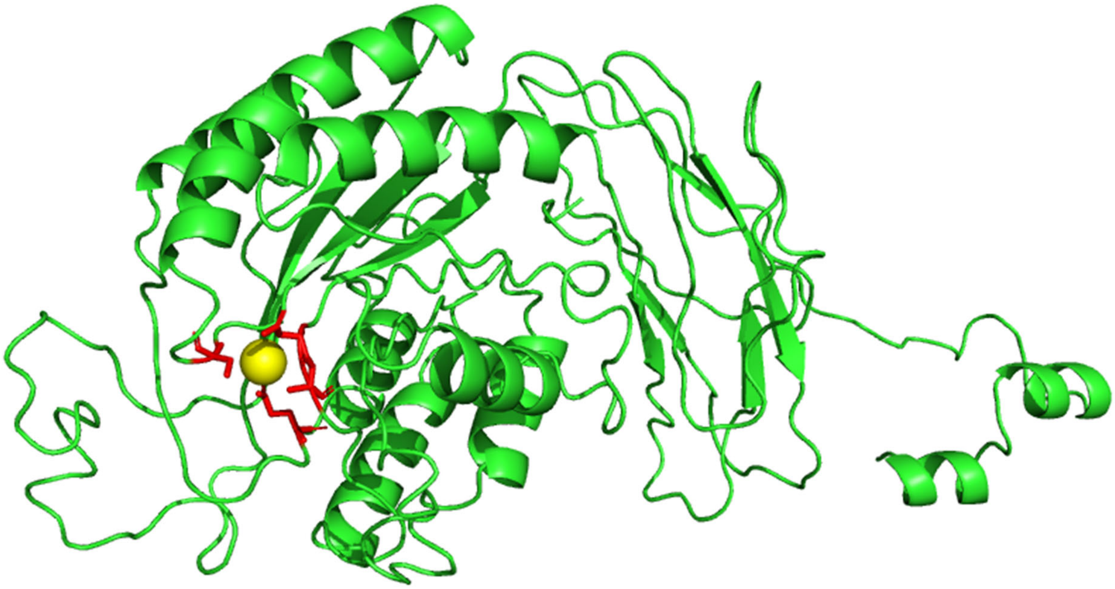


Figure S.16. FunFOLD3 ligand-binding site predictions for CASP12 target T0899
Predicted ligand-binding site residues are shown as sticks and coloured *red.* The predicted ligand magnesium (MG), is shown as sphere and coloured *yellow*

The third predicted CASP12 target with predictions is an uncharacterised protein from the gene Bd0412 and organism *Bdelllovibrio bacteriovorus.* As this CASP target was not associated with a PDB ID, information on observed ligand-binding residues is unable to be obtained. Information on UniProtKB(UniProt Consortium, 2019) demonstrates very little annotation with this protein, with only level one annotation, out of a possible five and the status is currently unreviewed. On UniProtKB,(UniProt Consortium, 2019) no GO terms are associated with this protein, so the function is unable to be determined. As one of the predicted ligands is magnesium, this can be used to establish what the function of the protein is based on other proteins that utilise magnesium as a ligand. Protein Tm1631 is from the hyperthermophilic organism (*Thermotoga maritima*) belongs to a domain of unknown function protein family. Ogrizek et al.,(Chen *et al.*, 2017) studied the role of magnesium ions in the protein and found that magnesium ions are required in the binding pockets to allow reactions to occur. Based on the location of the magnesium ion, in Figure S.16 it would be reasonable to assume that this is also the binding pocket of this protein and the magnesium ion has a role in enabling reactions. It is worth bearing in mind, that whilst the aim of FunFOLD3 for CASP competitions are prediction of ligand-binding sites, GO terms are also predicted and are presented in Table S.1 below. As can be seen in Table S.1, none of the GO terms are related to metal ion binding, which could suggest the primary function of this protein is not the binding of metal ions, but utilisation of metal ions to facilitate other actions of the protein. However, this would need to be validated using other methods. As no PDB ID is associated with this protein, the information provided by FunFOLD3 and UniProtKB are able to provide insights into the role and potentially function of the protein.

Table S.1. Predicted GO terms for CASP12 target T0899
The predicted GO terms for CASP12 target T0899 and their associated term domains and function are shown below. Biological process is coloured red, molecular function coloured green and cellular component coloured purple

GO term	GO term domain	Function
GO: 0001934	Biological process	positive regulation of protein phosphorylation
GO: 0001938	Biological process	positive regulation of endothelial cell proliferation
GO:0002576	Biological process	platelet degranulation
GO:0007155	Biological process	cell adhesion
GO:0007160	Biological process	cell-matrix adhesion
GO:0007229	Biological process	integrin-mediated signalling pathway
GO:0007275	Biological process	multicellular organism development
GO:0007411	Biological process	axon guidance
GO:0007596	Biological process	blood coagulation
GO:0010595	Biological process	positive regulation of endothelial cell migration
GO:0010745	Biological process	negative regulation of macrophage derived foam cell differentiation
GO:0010888	Biological process	negative regulation of lipid storage
GO:0014909	Biological process	smooth muscle cell migration
GO:0019048	Biological process	modulation by virus of host morphology
GO:0030168	Biological process	platelet activation
GO:0030949	Biological process	positive regulation of vascular endothelial growth factor receptor signalling pathway
GO:0032147	Biological process	activation of protein kinase activity
GO:0032369	Biological process	negative regulation of lipid transport
GO:0035295	Biological process	tube development
GO:0045124	Biological process	regulation of bone sorption
GO:0045715	Biological process	negative regulation of low-density lipoprotein particle receptor biosynthetic process
GO:0050731	Biological process	positive regulation of peptidyl-tyrosine phosphorylation
GO:0050748	Biological process	negative regulation of lipoprotein metabolic process
GO:0050900	Biological process	leukocyte migration
GO:0060055	Biological process	angiogenesis involved in wound healing
GO:0070527	Biological process	platelet aggregation
GO:0005886	Cellular component	plasma membrane

GO:0005887	Cellular component	integral component of plasma membrane
GO:0008305	Cellular component	integrin complex
GO:0016020	Cellular component	membrane
GO:0016021	Cellular component	integral component of membrane
GO:0031092	Cellular component	platelet alpha granule membrane
GO:0042470	Cellular component	melanosome
GO:0071062	Cellular component	alpha-beta3 integrin-vitronectin complex
GO:0003756	Molecular function	protein disulfide isomerase activity
GO:0004872	Molecular function	signalling receptor activity
GO:0005102	Molecular function	signalling receptor binding
GO:0005161	Molecular function	platelet-derived growth factor receptor
GO:0005515	Molecular function	protein binding
GO:0042802	Molecular function	identical protein binding
GO:0043184	Molecular function	vascular endothelial growth factor receptor 2 binding
GO:0050839	Molecular function	cell adhesion molecule binding

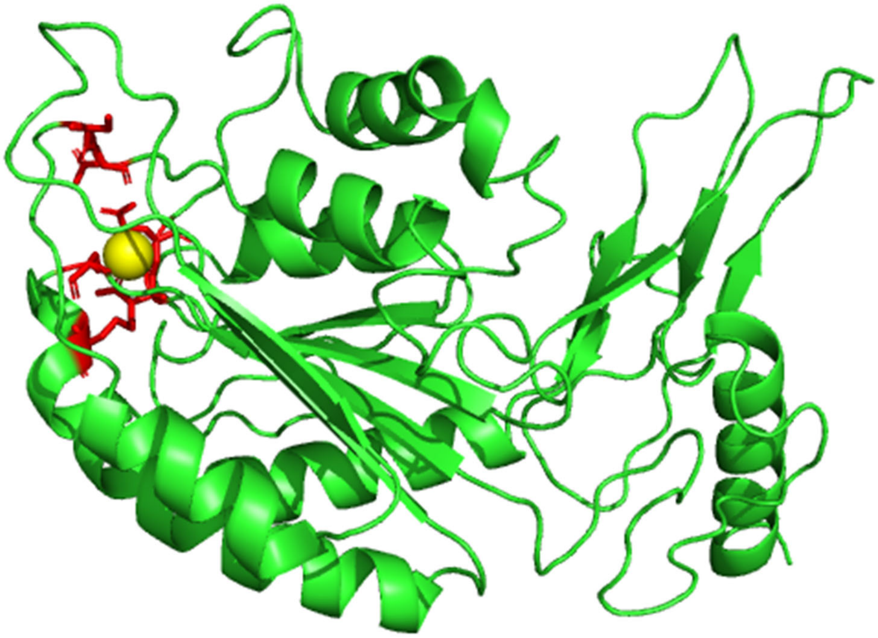


Figure S.18. FunFOLD3 ligand binding site predictions for CASP12 target T0901
Predicted ligand binding site residues are shown as sticks and coloured red with predicted ligand magnesium (MG) shown as spheres and coloured yellow

The fourth predicted CASP12 target is Bd3099 and is similar to T0899, in terms of organism (*Bdelllovibrio bacteriovorus*) and low/minimal level of annotation on UniProtKB.(UniProt Consortium, 2019) Once again, this protein has metal ions predicted with MG ligand. On PDB the enzymes that are associated with the MG ligand are; oxidoreductase (49 enzymes), hydrolase (39 enzymes), transferees (30 enzymes), lyases (nine enzymes), ligases (four enzymes) and isomerases (one enzyme). On the basis of limited information with this protein, it is impossible to determine which potential enzyme this protein could classify.(Ogrizek *et al.*, 2016) Further information around this protein, such as GO term annotations would be useful in determining its precise function. However, there were no predictions from FunFOLD3 related to this and this could be due to the limited information available in literature and available data in literature is required for BioLip annotations. Additionally, as no PDB ID is associated with this protein no further information can be provided.

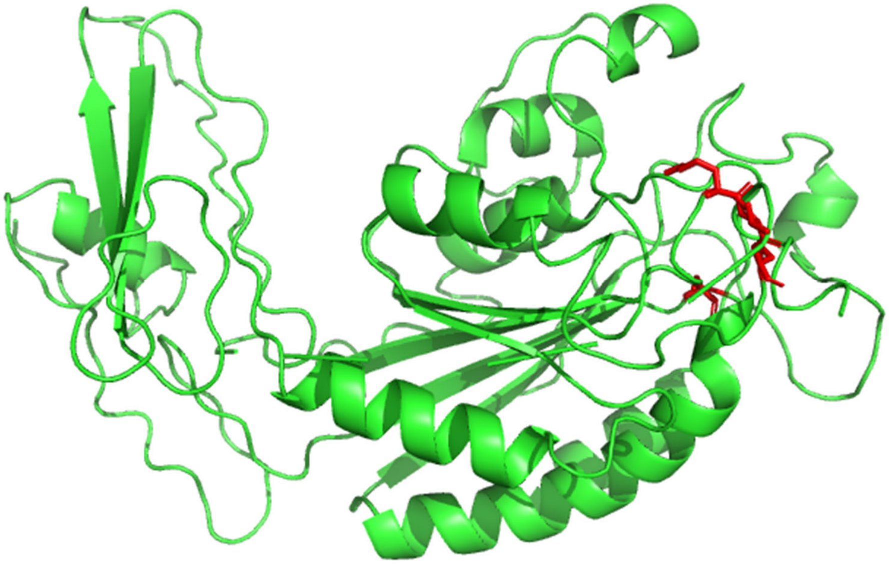


Figure S.18. FunFOLD3 ligand-binding predictions for CASP12 target T0905
Predicted ligand-binding site residues are shown as sticks and coloured *red* the predicted ligand glycine (GLY) has not been shown as it features as an amino acid throughout the structure of the protein

The fifth CASP12 target is Bd1483 (CASP ID T0905) and is similar to CASP targets T0899 and T0901, where there is low/minimal annotation available on this protein from UniProtKB(UniProt Consortium, 2019) and it is uncharacterised. Additionally, no PDB ID has been associated with the protein. However, the protein has been predicted but there is no evidence at protein, transcript or homology levels and this is also the same for CASP targets T0899 and T0901. Indeed, the results from the CASP12 experiment could be the level of evidence, which is available for these CASP targets. Interestingly, FunFOLD3 predicted the exact same GO terms for this CASP target as it did for T0899, however different ligands were predicted. Therefore, these two proteins could potentially share some structural homology and/or come from the same class of proteins just with different specific roles. It is highly likely that the predicted ligand, GLY, for T0905 is not biologically relevant due to not being presented in the structure.

One of the ways to determine structural homology is by using TMalign. CASP target T0899 contains 423 residues and T0905 contains 353 residues, thereby immediately demonstrating there is a difference between these two proteins. A TMscore of 0.719 is achieved for the two proteins, showing clear structural homology and potentially the same or similar function which adds further support to the same prediction of GO terms. The TMalign structures for these two CASP targets are illustrated in Figure S.19.

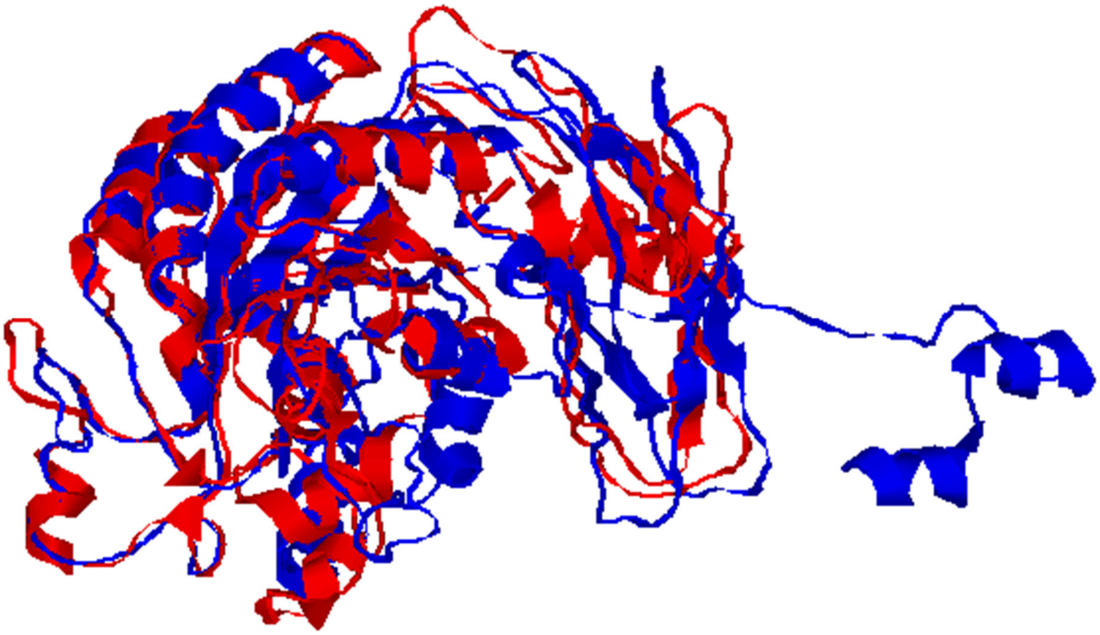


Figure S.19. Comparison of TMalign(Zhang and Skolnick, 2005) **superposition for CASP12 target T0899 and T0905** The structure in blue is the predicted structure for T0899 and the structure in red is T0905. The extra residues for T0899 can clearly be seen with an additional flexible loop and alpha helix. Both structures have been predicted by IntFOLD4

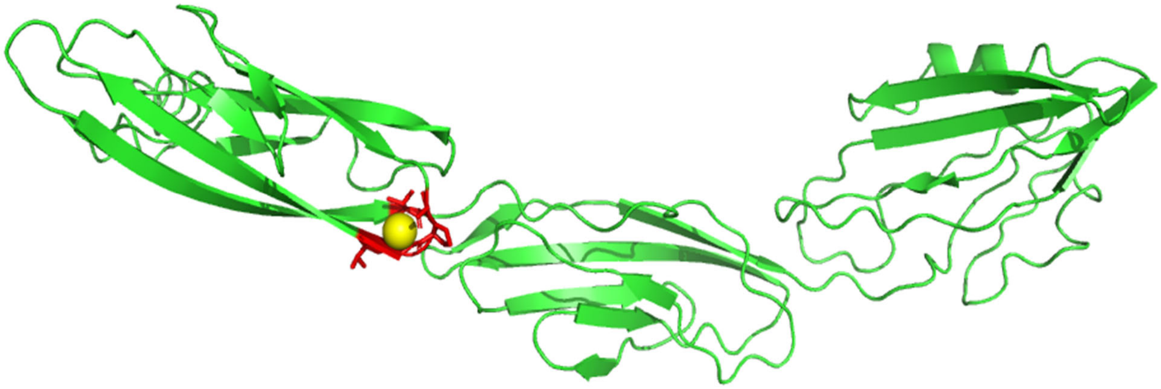


Figure S.20. FunFOLD3 ligand-binding predictions for CASP12 target T0907
Predicted ligand-binding site predictions are shown as sticks and coloured yellow with predicted ligand calcium (CA) shown as sphere and coloured red

The sixth predicted CASP12 target is PorM_NB-SU1 (CASP ID T0907) and as can be seen from Figure S.20, there is no image related to correct ligand-binding site residues. This is because, whilst FunFOLD3 predicted ligands, there are no ligands associated with this protein in the observed structure or from the PDB entry for the target, as is with CASP target T0868 and T0892.

A search on UniProtKB and PubMed has yielded no information on this specific protein. However, there is information linked to SU1 genes and proteins with SU1 in the submitted name.(Burley *et al.*, 2017) There are several proteins associated with the SU1 gene and these proteins are but not limited to sucrase, debranching enzyme, isoamylase, bacterium SU1 (bacterial small subunit ribosoml RNA), achromobacter sp. SU1 small subunit ribosomal RNA (16S), proteasome subunite beta type and RNA binding protein. The most interesting proteins are Cytochrome P450-SU1, which has eight annotations, one of these annotations is GO:0046872 and is metal ion binding, specifically cation binding and the ancestor chart is illustrated in Figure S.21, additionally baculoviral IAP repeat-containing protein 5.2-A also has this associated annotation.

Currently, the only information available for this specific protein is from the ligand predictions made by FunFOLD3. The predicted ligand is calcium, which is also located on the central part of the molecule. Currently, this does not reveal much information on the function of the protein. The prediction of GO terms would be the ideal starting point in order to provide further insight into the function of this protein.

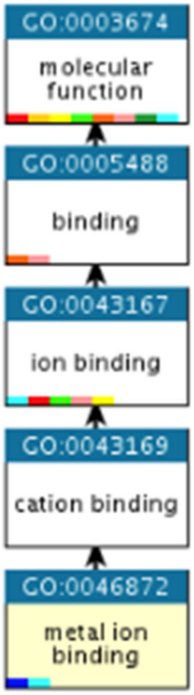


Figure S.21. GO ancestor chart for metal ion binding associated with Cytochrome P450-SU1 and baculoviral IAP repeat-containing protein 5.2-A
Chart demonstrates that metal ion binding is part of molecular function and as CASP12 target T0907 has a calcium ligand associated with it, there is a clear role of metal on binding with this protein. Figure created using(*QuickGO*, 2017)

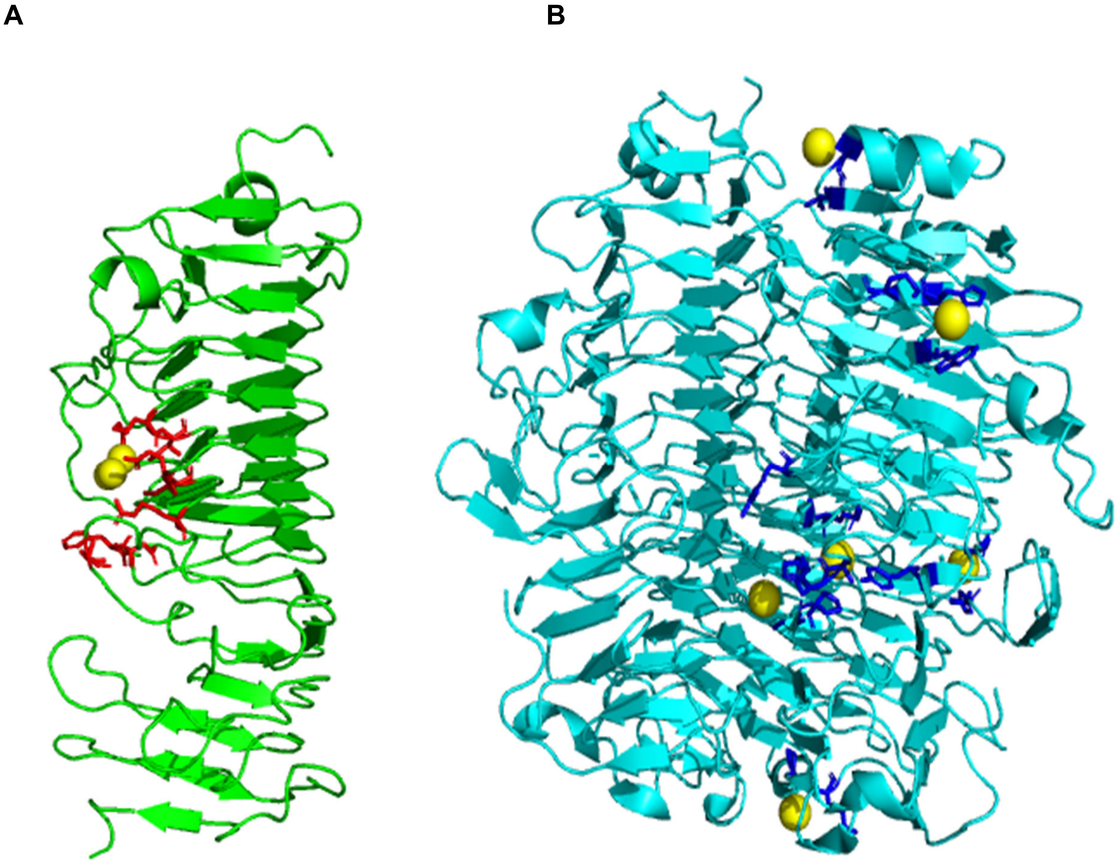


Figure S.22. FunFOLD3 ligand-binding site predictions for CASP12 target T0909 (PDB ID 5g5n)
(A) Predicted ligand binding site residues shown as sticks and coloured red and the predicted ligand CO3 shown as sphere and coloured yellow. MCC score of **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in blue the calcium (CL) ligands are shown as sphere and coloured yellow in multiple locations within the protein structure. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target

The seventh predicted CASP12 target is LH3 hexon-interlacing capsid protein (CASP T0909 and PDB ID 5g5n), as can be seen from Figure S.22 predicted and observed ligand-binding site residues were obtained. The latter of which was obtained from the PDB file associated with the entry and the CASP12 organisers did not release an observed structure. The protein is classified as a viral protein with three chains and the associated observed CL ligand present in all three chains. Additionally, methyl mercury ion and glycerol are also associated with the PDB entry. Literature information associated with the PDB entry, states a stable fragment of the snake adenovirus (SnAdV-1) LH3 protein was crystallised with a methylmercury chloride derivative and a total of fourteen chloride ions were modelled within the protein structure(Menéndez-Conejero *et al.*, 2017) and this matches the number of chloride ions in the sequence when viewed using PyMOL. It is clear from the information in literature that the chloride ion has been used in the crystallisation of the protein structure so it could be argued whether chloride ligand is biologically relelvant.

Although there is a PDB entry with ligands, there are limited data on the protein related to its function. UniProtKB has no annotations related to the function.(Consortium, 2017) The GO terms predicted by FunFOLD3 are shown in Table S.2 below:

Table S.2. Predicted GO terms for CASP12 target T0909 (PDB ID 5g5n)
The predicted GO terms for CASP12 target T0909 and their associated term domains and function are shown below. Biological process is coloured red, molecular function coloured green and cellular component coloured purple

GO term	GO term domain	Function
GO:0016740	Molecular function	transferase activity
GO:0016829	Molecular function	lyase activity
GO:0030570	Molecular function	pectate lyase activity
GO:0046872	Molecular function	metal ion binding
GO:0045490	Biological process	pectin catabolic process
GO:0005576	Cellular component	extracellular region

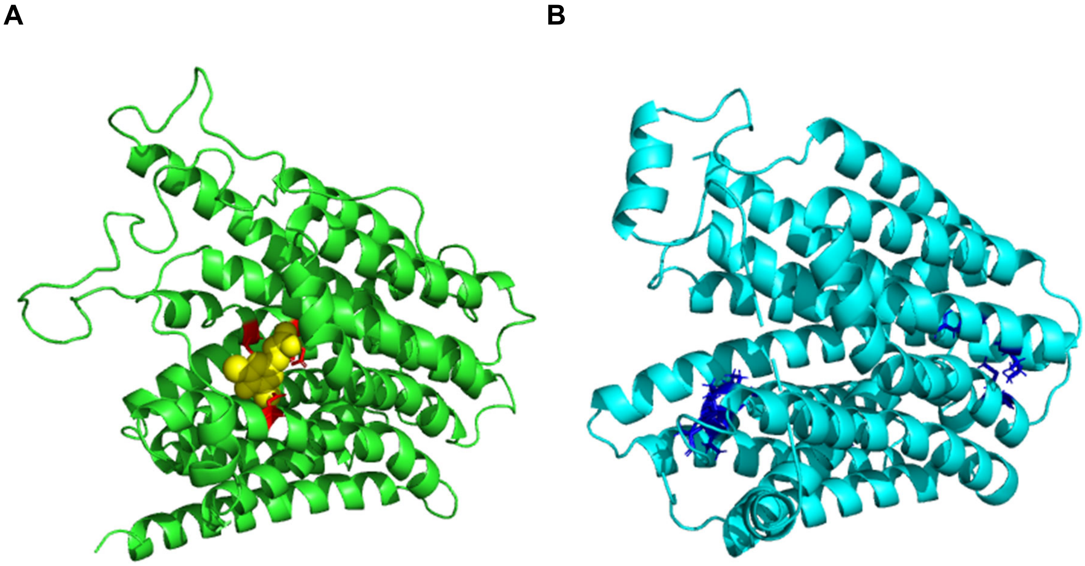


Figure S.23. FunFOLD3 ligand-binding predictions for CASP12 target T0911 (PDB ID 6e9n)
(A) Predicted ligand binding site residues shown as sticks incorrect predictions in *red*, the predicted ligand phosphate dibromotyrosine (DBY) shown as sphere and coloured *yellow.* **(B)** The observed ligand binding site residues shown as sticks for T0911, with binding site residues coloured in *blue*. The gluconic acid ligand has not been illustrated as it is not present in the sequence when in PyMOL

The eighth predicted CASP12 target is D-galactonate transporter from *Escherichia Coli* (CASP T0911). The PDB entry for this target classifies the protein as a membrane protein and in terms of ligands, nonyl beta D-glucopyranoside (BNG) which is present in chains A and B and D-gluconic acid (GCO) which is present in chain A. Literature information available on the protein, as associated with the PDB entry, identifies conserved residues in a pocket between sialin and vesicular glutamate transporters these include ARG47, ARG126 and GLU133. Of which, ARG126 was identified in the observed protein structure.(Leano *et al.*, 2019)

Several different entries exist for this protein on UniProtKB(Consortium, 2017) based on the residue length of 445 residues as obtained from CASP this short lists the protein to 36 entries out of a possible 2,652 entries. All of the 36 entries are unreviewed and the level of evidence is minimal with the protein predicted for the level of evidence available. This means there is no evidence at protein, transcript or homology levels. Despite the low level of annotation, there are GO terms associated with the molecular function and biological process of this protein. The predicted GO terms are shown in Table S.3 below:(Consortium, 2017)

Table S.3. Predicted GO terms for CASP12 target T0911
The predicted GO terms for CASP12 target T0911 and their associated term domains and function are shown below. Biological process is coloured red, molecular function coloured green and cellular component coloured purple

GO term	GO term domain	Function
GO:0022857	Molecular function	transmembrane activity
GO:0055085	Biological process	transmembrane transport
GO:0016020	Cellular component	membrane
GO:0016021	Cellular component	integral component of membrane

GO term 0022857 has the qualifier *enables* and 0055085 has the qualifier *involved in*, therefore this protein enables transmembrane transporter activity and is involved with transmembrane transport, which is also suggested in the name of the protein. FunFOLD3 predicted phosphate ion as a possible ligand but was not deemed to be the likely ligand. Literature evidence on phosphate ions suggests phosphate ions in the form of $HP0_4^{2-}$ and $H_2PO_4^-$ regulate the size of rapidly releasable intracellular calcium pool and is produced in the cytoplasm.(Consortium, 2017)

Figure S.24 below, shows the TMalign superposition of observed and predicted structures. A TM-score of 0.82921 was achieved between the structures, demonstrating good structural homology.

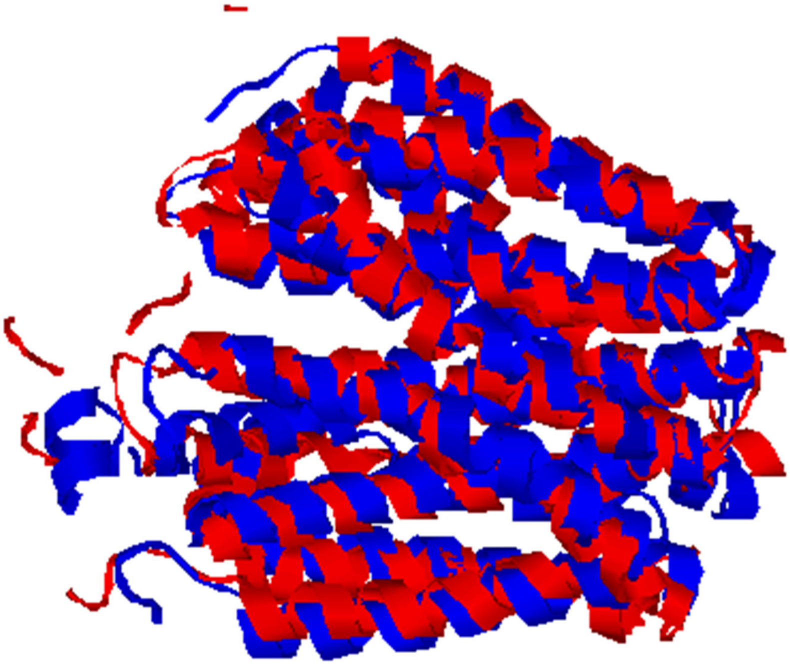


Figure S.24. Comparison of TM-align(Zhang and Skolnick, 2005) superposition for CASP12 target T0911
The structure in blue is the observed structure for T0911 and the structure in red is the predicted structure from IntFOLD4. A TM-score of 0.82921 was achieved. The score was normalised for the observed structure as it is the reference molecule.

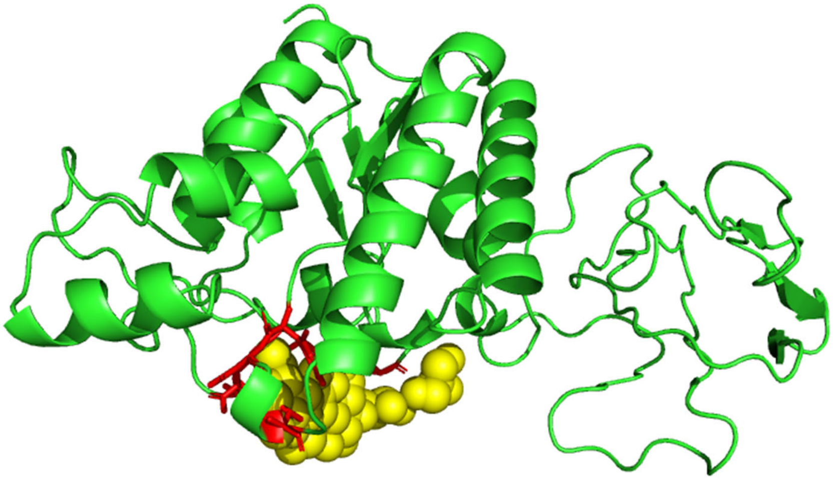


Figure S.25. FunFOLD3 ligand-binding site predictions for CASP12 target T0919
Predicted ligand binding site residues shown as sticks and coloured red with the predicted ligand BGC shown as sphere and coloured yellow. The observed structure was cancelled by CASP12 organisers

The twelfth CASP12 target is C-terminal part of Gp20 (T0919) and information available from UniProtKB identifies this protein as Phage protein. As with all the previous CASP12 targets, which have no PDB ID, there is limited data available on this protein, with no evidence at protein, transcript or homology levels. There is one annotation associated with this protein in cellular component, the GO term *is* 0019028 and the *qualifier is* part of meaning the protein is part of viral capsid.(Hergueta-Redondo *et al.*, 2014) In comparison, FunFOLD3 predicted the following GO terms as shown in Table S.4:

Table S.4. Predicted GO terms for CASP12 target T0919
The predicted GO terms for CASP12 target T0919 and their associated term domains and function are shown below. Molecular function coloured green

GO term	GO term domain	Function
GO:0006629	Molecular function	lipid metabolic process
GO:0016787	Molecular function	hydrolase activity
GO:0016788	Molecular function	hydrolase activity
GO:0046872	Molecular function	metal ion binding

GO:0046872 ties in with the prediction of a magnesium ligand, and the ancestor for this GO term is cation binding and magnesium is a cation. A literature search on portal ring protein Gp20 provides information of the function of Gp20 in the assembly of phage T4. Gp20 along with Gp40 and other unknown *E.Coli* proteins build a membrane-bound initiator complex and following on from this is involved in the next stage, which is assembly of a prohead. Gp20 plays a crucial role for the assembly process.(Consortium, 2017) As Gp20 is involved in the assembly of a membrane-bound initiator complex, the prediction of GO term 0006629 seems quite reasonable as the membrane layer will most likely be made of lipids. The formation of such a complex would most likely involve a reaction of which the ligand ion; magnesium may well act as a co-factor. As with all previous CASP12 with limited annotations, further investigation is required in order to determine if the GO terms predicted by FunFOLD3 are related to the function of this protein.

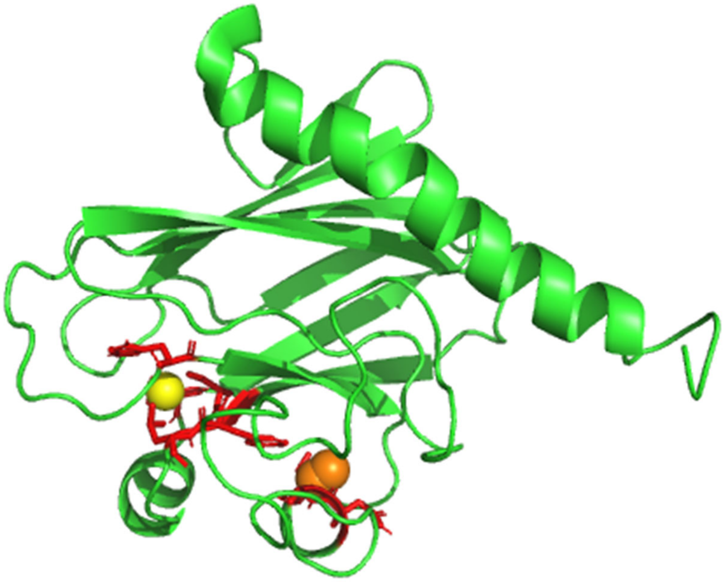


Figure S.26. FunFOLD3 ligand-binding site predictions for CASP13 target T0949
Predicted ligand binding site residues shown as sticks and coloured red with the predicted oxygen ligand shown as sphere and coloured orange and the predicted copper ligand also shown as sphere coloured yellow. No observed structure was released by CASP organisers so no comparisons can be made

The first CASP13 target with ligand binding site residue predictions and ligands is B7JAQ5_ACIF2. No PDB ID was associated with this protein and the CASP organisers did not release an observed structure following prediction. The UniProtKB(UniProt Consortium, 2019) entry identifies this protein as an uncharacterised protein with the lowest annotation score of one out of five meaning the denotation of the protein is predicted, meaning that there is no evidence at protein, transcript or homology levels. Furthermore, the homology model is PDB ID 6kol which is *roseiflexus castenholzil* and is classified as a metal binding protein with two ligands copper and chloride ion. *Roseiflexus castenholzil* contains a chloroflexus aurantiacus copper binding pocket. However, it is worth noting that this template was not identified by FunFOLD3, despite showing some structural homology which can be seen in Figure S.27 and an TMscore of 0.84753 was achieved, for normalisation of 6kol which is the reference structure in this instance. This clearly demonstrates a high level of molecular similarity between the predicted structure for T0949 and *roseiflexus castenholzil*. The total number of residues in *roseiflexus castenholzil* was 125 residues and in T0949 183 residues. The extra residues in T0949 can be seen in the alpha helix portion on the T0949 molecule.

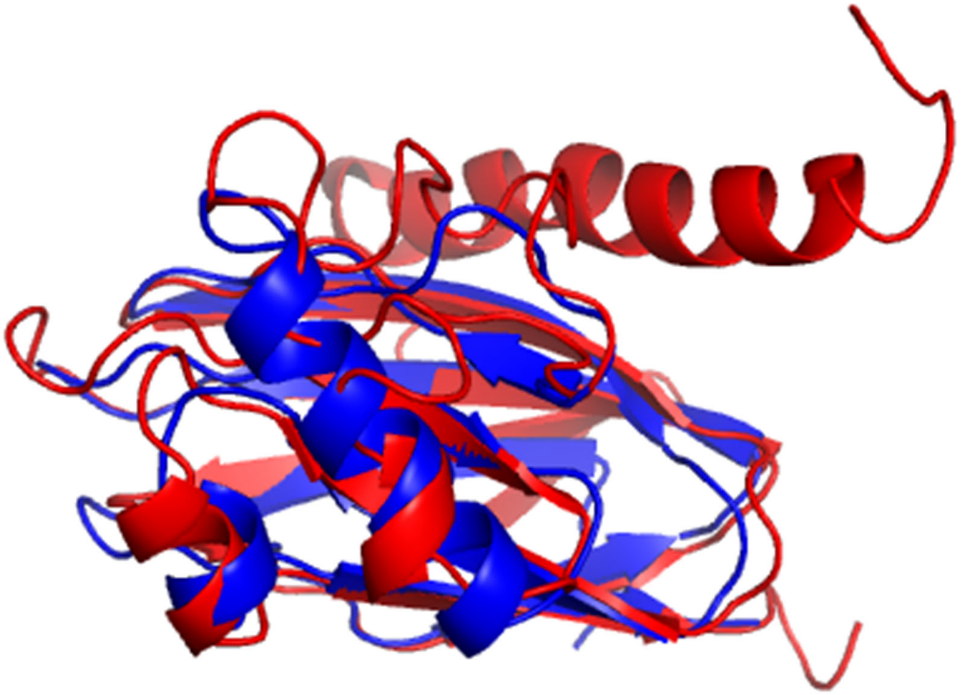


Figure S.27. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP13 target T0949 and *roseiflexus castenholzil* (PDB ID 6kol)
The structure in blue is the structure of 6kol and the predicted structure for CASP13 target T0949 in red.. A TM-align score of 0.84753 was achieved for the protein structures. This was normalised for *roseiflexus castenholzil* (PDB ID 6kol) as it is the reference molecule

The UniProtKB entry identifies *Acidithiobacillus ferrooxidans* as the organism as the source of the protein sequence.(UniProt Consortium, 2019) *Acidithiobacillus ferrooxidans* are microorganisms used for the industrial recovery of copper.(Valdés *et al.*, 2008)

The GO terms predicted by FunFOLD3 tie in with the function of *Acidithiobacillus ferrooxidans* in the following ways; copper ion binding (GO:005507) which also relates to metal ion binding (GO:0046872). In the process of bioleaching *Acidithiobacillus ferrooxidans* plays a key role by reoxidising the Fe(II) to Fe(III) and thus supports the following GO terms predicted by FunFOLD3; oxidoreductase (GO:0016491) and oxidation-reductase process (GO:0855114).

Table S.5 below, shows the GO terms predicted by FunFOLD3.

Table S.5. Predicted GO terms for CASP13 target T0949
The predicted GO terms for CASP13 target T0949 and their associated term domains and function are shown below. Biological process is coloured red, molecular function coloured green and cellular component is coloured purple

GO term	GO term domain	Function
GO: 0005507	Molecular function	copper ion binding
GO:0009055	Molecular function	electron transfer activity
GO:0016491	Molecular function	oxidoreductase activity
GO:0046872	Molecular function	metal ion binding
GO:0050421	Molecular function	nitrite reductase (NO-forming) activity
GO:0052716	Molecular function	hydroquinone:oxygen oxidoreductase activity
GO:0004129	Molecular function	cytochrome-c oxidase activitY
GO:0005509	Molecular function	calcium ion binding
GO:0050304	Molecular function	nitrous-oxide reductase activity
GO:0020037	Molecular function	heme binding
GO: 0030435	Biological process	sporulation resulting in formation of a cellular spore
GO: 0055114	Biological process	oxidation-reduction process
GO: 0006807	Biological process	nitrogen compound metabolic process
GO: 0019333	Biological process	denitrification pathway
GO: 0042128	Biological process	nitrate assimilation
GO:0022900	Biological process	electron transport chain
GO:0017000	Biological process	antibiotic biosynthetic process
GO:0030245	Biological process	cellulose catabolic process
GO:0046274	Biological process	lignin catabolic process
GO:0022904	Biological process	respiratory electron transport chain
GO:0042597	Cellular Component	periplasmic space
GO:0043245	Cellular Component	extraorganismal space
GO:0016020	Cellular Component	membrane
GO:0005886	Cellular Component	plasma membrane
GO:0016021	Cellular Component	integral component of membrane
GO:0070469	Cellular Component	respirasome
GO:0009279	Cellular Component	cell outer membrane

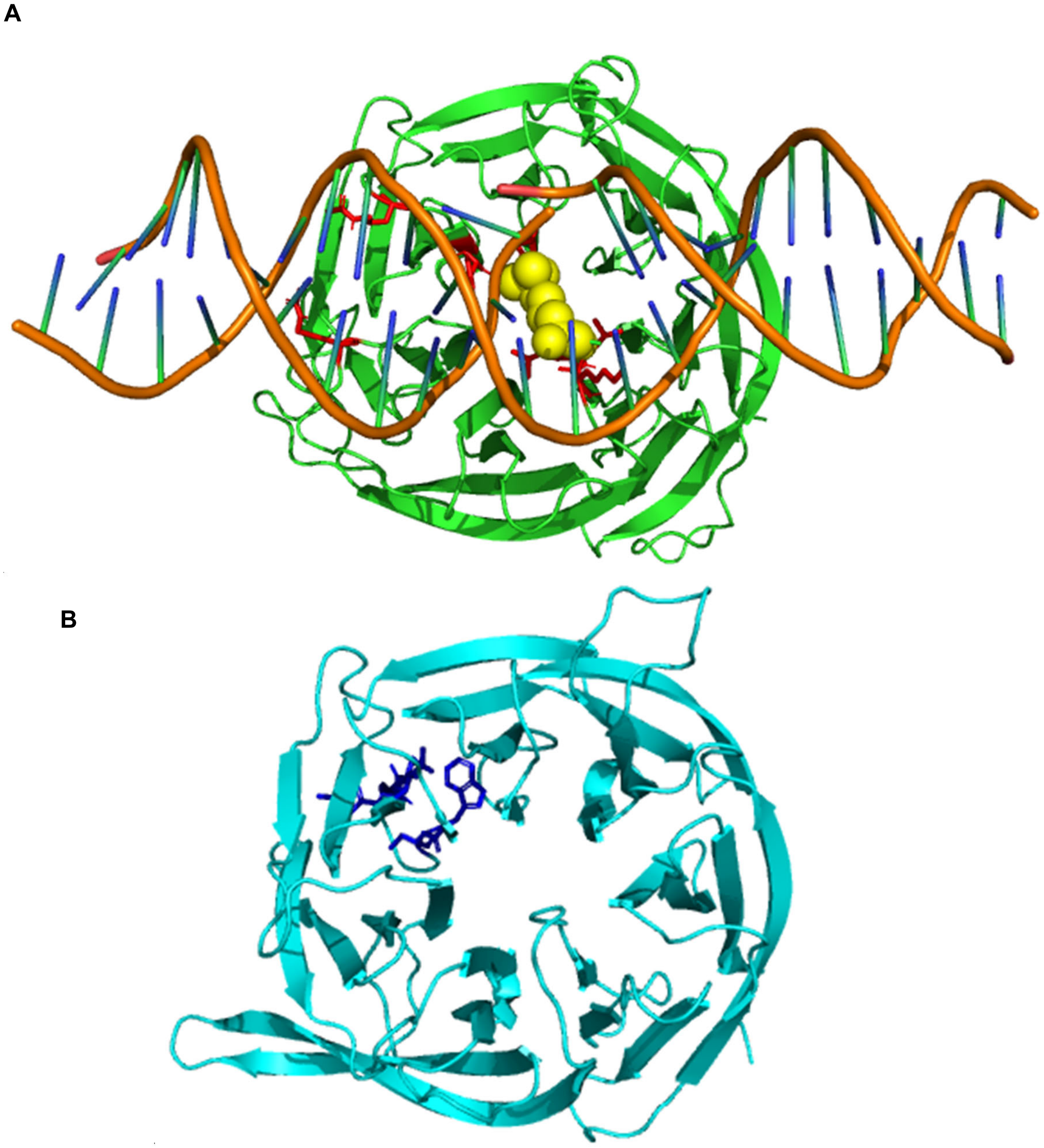


Figure S.28. FunFOLD3 ligand-binding site predictions for CASP13 target T0954 (PDB ID 6cvz)
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted ligand lysine shown as sphere and coloured yellow and the thymidine-5'-monophosphate (DT) ligand shown as double helix **(B)** The observed ligand binding site residues shown as sticks for T0954 with binding site residues coloured in blue the magnesium ligand is not shown as it is not present in the PDB file

The third CASP13 target was RFWD3_HUMAN and is classified as a transferase protein. The UniProtKB and PDB entry expands the protein to E3 ubiquitin-protein ligase RFWD3.(UniProt Consortium, 2019) The human genome is frequently exposed to insults form the environment and endogenous sources and DNA lesions resulting from these insults need to be repaired in order to maintain genetic stability.(Elia *et al.*, 2015) The co-ordination of the repair is done by a network known as DNA damage response (DDR). Regulation of DDR is mediated by a host of protein modifications, among which ubiquitination plays a key role.(Elia *et al.*, 2015) Ubiquitin-dependent signalling regulates the double-strand response of DNA.(Elia *et al.*, 2015) As can be seen in Figure S.28A DNA is one of the ligands that has been predicted by FunFOLD3. In contrast, the observed ligand is magnesium however, this does fit in with the molecular function of the target as per the UniProtKB(UniProt Consortium, 2019) entry and shown in Table S.6 below.

Table S.6 below shows the molecular function and biological processes associated with RFWD3 as per the UniProtKB(UniProt Consortium, 2019) entry. FunFOLD3 predictions which match the are denoted by ticks or if closely related to the GO term, the ancestor term is given

Table S.6. Predicted GO terms for CASP13 target T0954 (PDB ID 6cvz)
The GO terms for CASP13 target T0954 (PDB ID 6cvz) and their associated term domains and function as per the UniProtKB entry are shown below . The association with FunFOLD3 is denoted in the final column with exact matches denoted with a tick and related GO terms with the associated fucntion. Biological process is coloured red and molecular function

GO term	GO term domain	Function	FunFOLD3
GO:0097371	Molecular function	MDM2/MDM4 family protein binding	GO:0005515 (protein binding)
GO:0046872	Molecular function	metal ion binding	N/A
GO:0002039	Molecular function	p53 binding	GO:0005515 (protein binding)
GO:0061630	Molecular function	ubiquitin protein ligase activity	GO: 0004842 (ubiquitin-protein transferase activity)
GO:0006974	Biological process	cellular response to DNA damage stimulus	✓
GO:0031052	Biological process	chromosome breakage	N/A
GO:0000724	Biological process	double-strand break repair via homologous recombination	GO:0006281 (DNA repair)
GO:0036297	Biological process	interstrand cross-link repair	GO:0006281 (DNA repair)
GO:0031571	Biological process	mitotic G1 DNA damage checkpoint	N/A
GO:0016567	Biological process	protein ubiquitination	✓
GO:2000001	Biological process	regulation of DNA damage checkpoint	GO:0006974 (cellular response to DNA damage)
GO:0031297	Biological process	replication fork processing	N/A
GO:0010212	Biological process	response to ionising radiation	N/A

A TMscore of 0.87236 was obtained and this assumes the same fold within the two proteins.

The superposition of the structures is shown in Figure S.29 below.

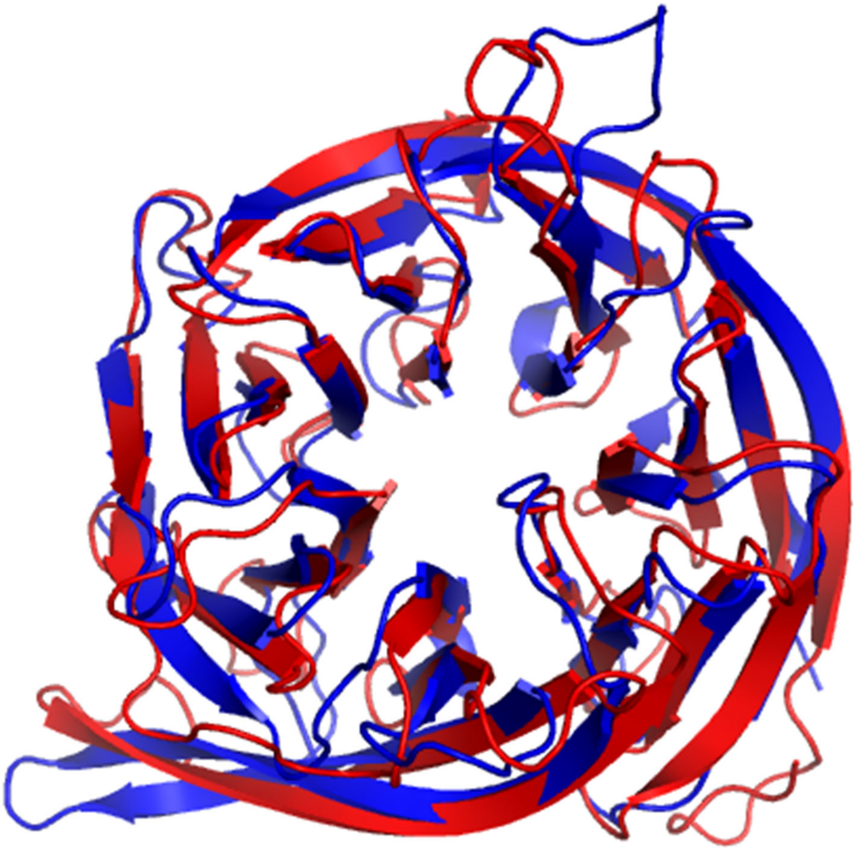


Figure S.29. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP13 target T0954 (PDB ID 6cvz) The structure in blue is the observed structure for 6f45 and the predicted structure for CASP13 target T0954 in red. A TM-align of 0.87236 was achieved for protein structures. The score was normalised for PDB ID 6cvz as it is the reference molecule

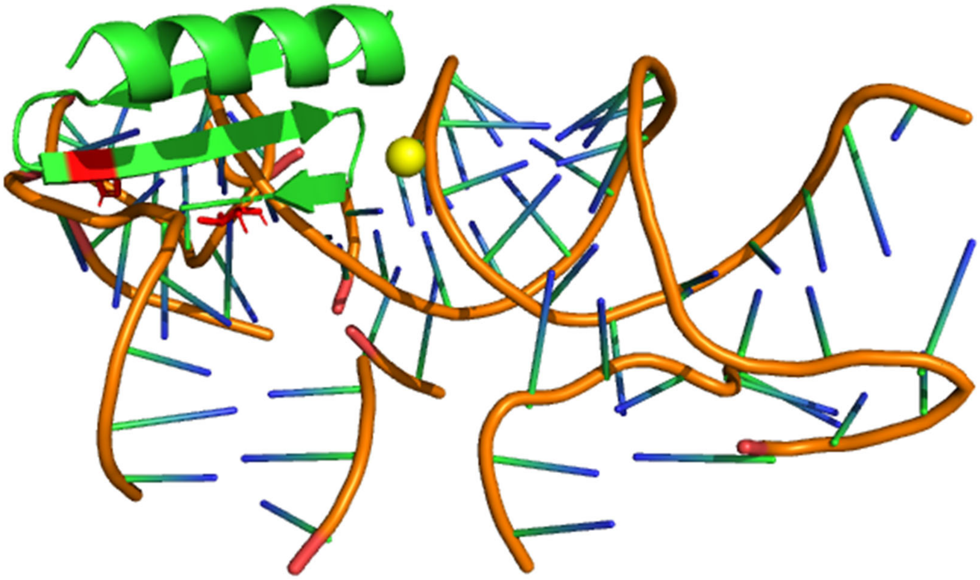


Figure S.30. FunFOLD3 ligand-binding site predictions for CASP13 target T0955 (PDB ID 5w9f)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted ligand ZN shown as sphere and coloured *yellow* and the RNA ligand shown as double helix. No biologically relevant ligands were found in the observed structure

The fourth predicted CASP13 target was gHEEE_02 and is classified as a de novo protein and is an synthetic construct. The protein is a mix of two peptides with divergent structures and sequences containing a mixed α/β topology with a helix packing against a three-stranded antiparallel β-sheet stabilised by three disulphide bonds.(Buchko *et al.*, 2018) As the protein has been designed computationally, there are no further data available the target and the PDB entry contains no information on ligands and no biologically relevant ligands were found on the observed structure using FunFOLD3.

The TMalign score for the predicted structure was 0.73628 showing good structural alignment between the predicted and observed protein structure. A superposition of the two structures are shown below in Figure S.31.

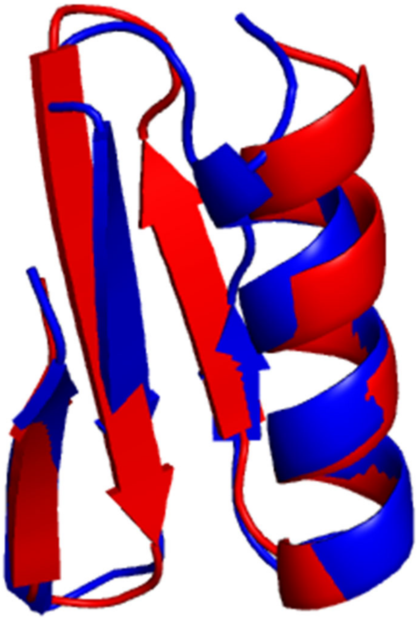


Figure S.31. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP13 target T0955 (PDB ID 5w9f) The structure in blue is the observed structure for PDB ID 5w9f and the predicted structure for CASP13 target T0955 in red. A TM-align of 0.73628 was achieved for protein structures. The score was normalised for PDB ID 5w9f as it is the reference molecule

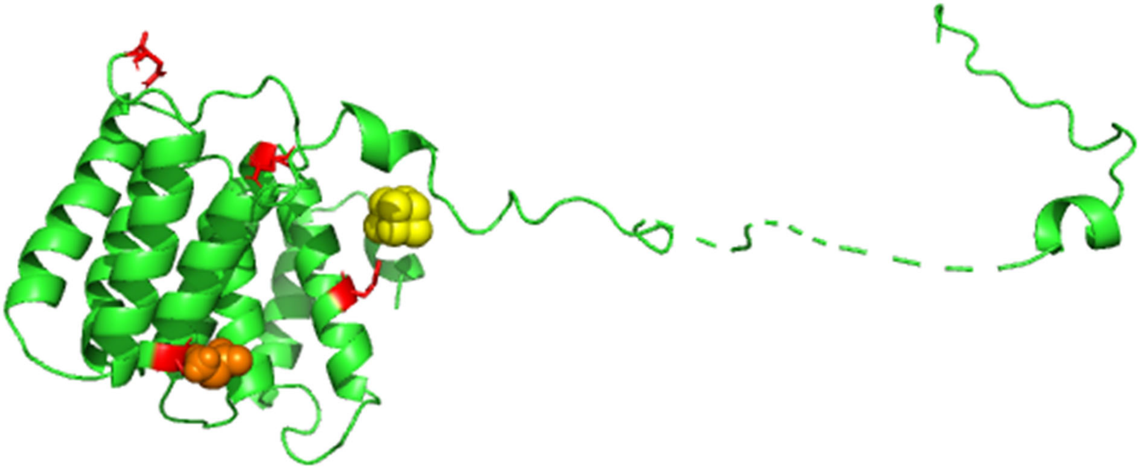


Figure S.32. FunFOLD3 ligand-binding site predictions for CASP13 target T0957s2 (PDB ID 6cp8)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted ligand iron/sulphur cluster (SF4) shown as sphere and coloured *yellow* and the alanine (ALA) ligand shown as sphere and coloured orange. No biologically relevant ligands were found in the observed structure

The fifth predicted CASP13 target was CdiA_CdiI-CPX200209, a contact-dependent growth inhibition toxin-immunity protein complex and is classified as a toxin/antitoxin. Contact-dependent growth inhibition (CDI) is a form of interbacterial competition meditated by CdiB-CdiA two partner secretion systems.(Gucinski *et al.*, 2019) CdiA effectors, bind to specific receptors on neighbouring bacteria and deliver C-terminal toxic domains to suppress target cell growth. Upon binding a specific receptor, CdiA transfers its C-terminal toxin domain (CdiA-CT) into the target bacterium through an incompletely understood translocation pathway.(Ruhe *et al.*, 2017) Therefore, based on this role it is expected that ligands and ultimately ligand-binding site residues would be predicted. The PDB entry for 6cp8 identifies glycerol and 4-(2-Hydroxyethyl)piperazin-1-ylethanesulphonic acid (EPE) as ligands and in comparison FunFOLD3 predicted SF4 and alanine.

As there were no biologically relevant ligands predicted in the observed structure and the identified ligand predicted by FunFOLD3 it is difficult to make any comparisons.

The TMscore for the predicted structure is 0.61518 showing good structural alignment between the predicted and observed protein structure. A superposition of the two structures are shown below in Figure S.33.

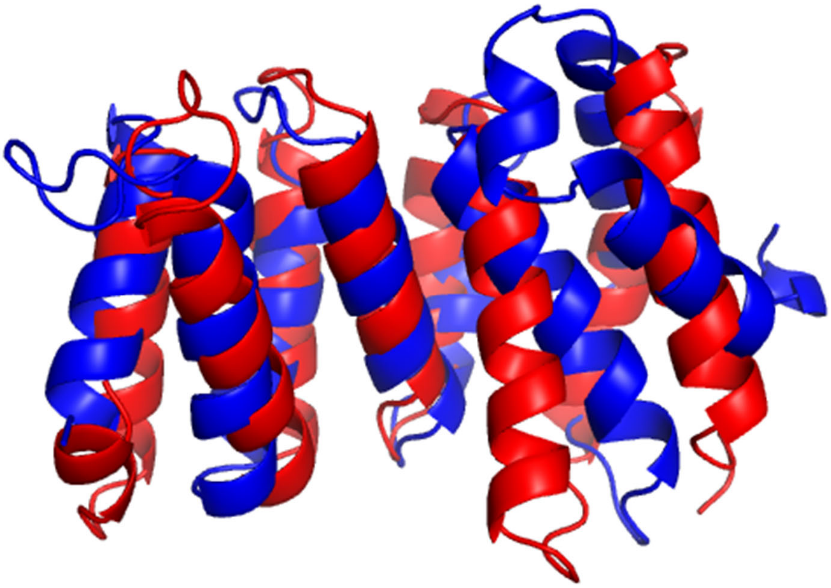


Figure S.33. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP13 target T0957s2 (PDB ID 6cp8) The structure in blue is the observed structure for PDB ID 6cp8 and the predicted structure for CASP13 target T0957s2 in red. A TM-align of 0.61518 was achieved for protein structures. The score was normalised for PDB ID 6cp8 as it is the reference molecule. The TMalign image is showing aligned portion of the protein molecules, therefore disordered regions have not been included

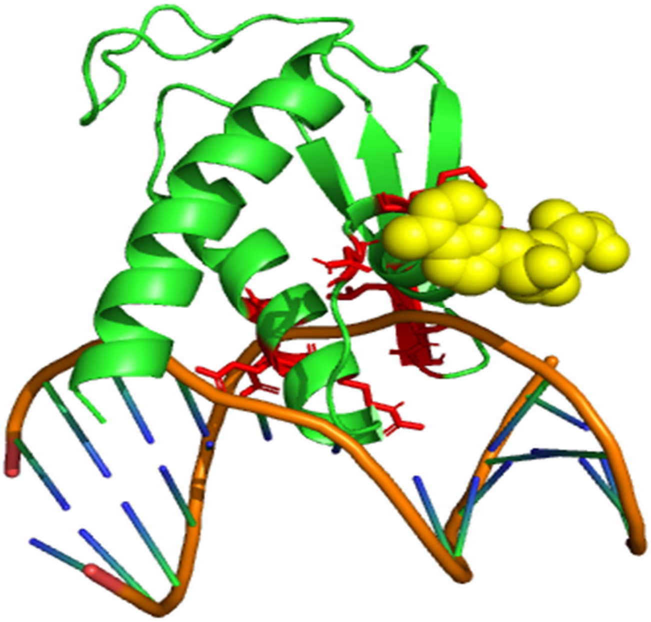


Figure S.34. FunFOLD3 ligand-binding site predictions for CASP13 target T0958 (PDB ID 6btc)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted s-adenosyl-l-homocysteine ligand (SAH) shown as sphere and coloured *yellow* and the DNA ligand shown as double-helix. No biologically relevant ligands were found in the observed structure

The sixth predicted CASP13 target was LP1413 and the PDB entry classifies the target as DNA binding protein, despite the classification, DNA and no other ligands are included in the PDB entry. Additionally, FunFOLD3 did not predict any biologically relevant ligands in the observed structure. However, DNA was predicted as a ligand in the predicted structure from the CASP competition and this fits in with the classification of the protein.

Literature information on LP1413 identifies the protein as a ssDNA-binding protein and is encoded by staphylococcal cassette chromosome (SCC) family.(Mir-Sanchis, Pigli and Rice, 2018) The naming of the protein means little protein with domain of unknown function 1413.(Mir-Sanchis, Pigli and Rice, 2018) Studies have found that it binds single stranded DNA with high affinity and weak affinity for double-stranded DNA, however experimental structure evidence shows LP1413 adopts a winged helix-turn-helix DNA binding motif and this suggests the protein can bind dsDNA, as shown in Figure S.34.(Mir-Sanchis, Pigli and Rice, 2018)

In addition to DNA, FunFOLD3 predicted *s*-adenosyl-L-homocysteine (SAH) as a ligand, SAH is a potent competitive inhibitor of *S*-adenosyl-l-methionine (AdoMet)-dependent methyltransferases, given the identification of a potential role of LP1413, it is difficult to determine the role of this ligand in relation to the protein function.

The TMscore for the predicted structure is 0.63081 showing good structural alignment between the predicted and observed protein structure. A superposition of the two structures are shown below in Figure S.35.

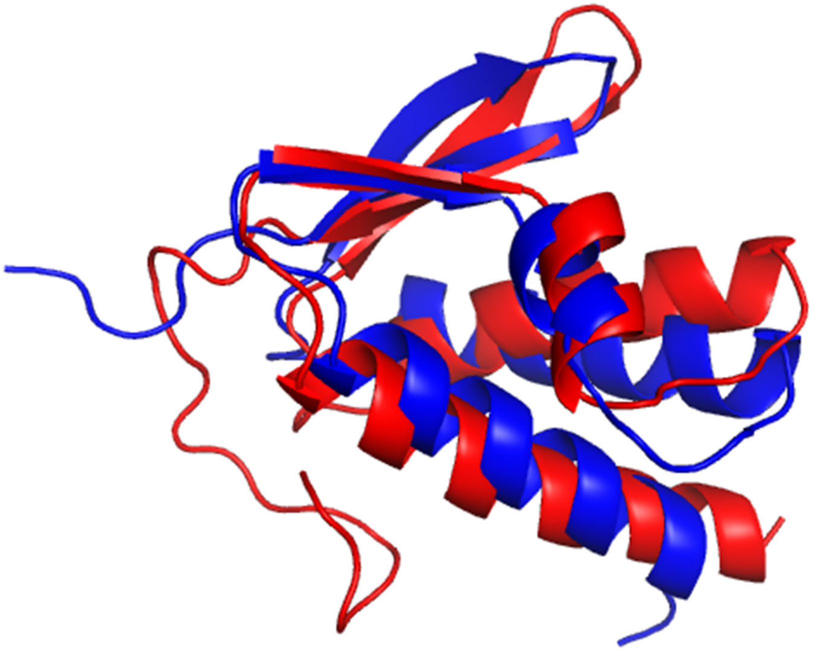


Figure S.35. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for CASP13 target T0958 (PDB ID 6btc) The structure in blue is the observed structure for PDB ID 6btc and the predicted structure for CASP13 target T0958 in red. A TM-align of 0.63081 was achieved for protein structures. The score was normalised for PDB ID 6btc as it is the reference molecule.

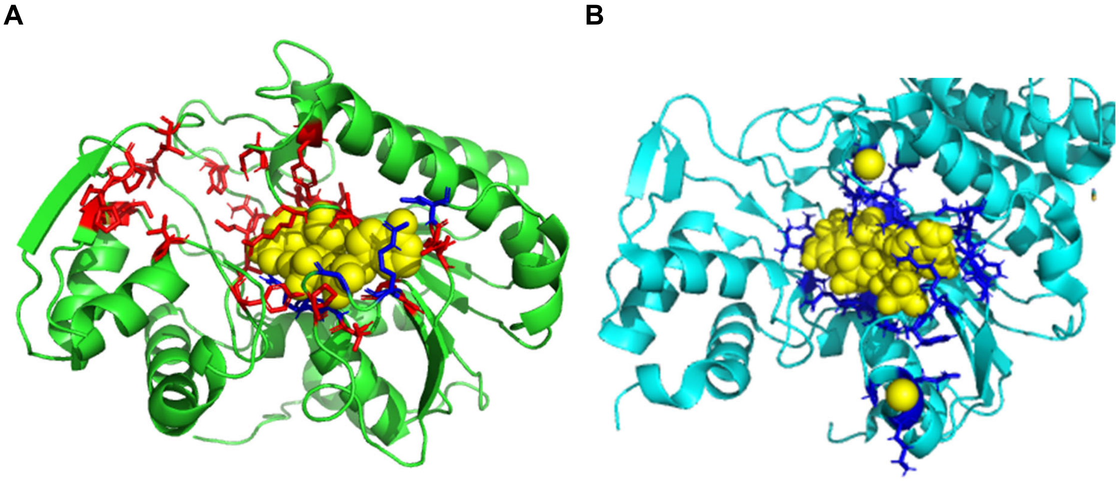


Figure S.36. FunFOLD3 ligand-binding site predictions for CASP13 target T0965 (PDB ID 6d2v)
(A) Predicted ligand binding site residues shown as sticks with incorrect predictions in red and correct predictions in blue**,** the predicted nicotinamide-adenine dinucleotide (NAD) ligand shown as sphere. A BDT score of 0.35 and an MCC score of 0.12 was achieved, respectively **(B)** The observed ligand binding site residues shown as sticks with binding site residues coloured in blue the NDP ligand is shown as the larger sphere and the CL ligands are shown as sphere and coloured yellow. The actual structure is a dimer, whereas the predicted structure is a homodimer. As a result of CASP13 organisers not releasing an observed structure, this is the structure as per the PDB entry for target

The eighth predicted CASP13 target is NADP dependent oxidoreductase, the PDB ID entry is the apo structure and has been used for the comparison against the predicted structure, due to no observed structure being released by the CASP13 organisers. The observed protein is a dimer with two chains and the PDB ID entry states NADPH dihydro-nicotinamide-adenine-dinucleotide phosphate (NDP), thiocyanate ion (SCN) and chloride (CL). FunFOLD3 predicted NAD as the biologically relevant ligand. Nicotinamide adenine dinucleotide is a coenzyme as well as a substrate for three classes of enzymes (sirtuin family deacetylases, poly(ADP)-ribosyl polymerase and cADP-ribose synthases.(Xiao *et al.*, 2018) NAD$^+$ can be reduced to NADH via dehydrogenases and can also be phosphorylated to NADP$^+$ via NAD$^+$ kinases.(Xiao *et al.*, 2018) The NAD+/NADH redox couple is known as a regulator of cellular energy metabolism e.g. glycolysis and mitochondrial oxidative phosphorylation.(Xiao *et al.*, 2018) In contrast, NADP+/NADPH is involved in maintaining redox balance and supporting the biosynthesis of fatty acids and nucleic acids.(Xiao *et al.*, 2018) Despite the close similarities between the two ligands, there are subtle differences. However, the FunFOLD3 prediction is close to the observed ligand. The BDT and MCC score was 0.35 and 0.12, respectively.

The TMscore for the predicted structure compared to the observed structure is 0.81556, demonstrating good structural alignment between the predicted structure and the structure obtained from PDB. A superposition of the two structures is seen below in Figure S.37.

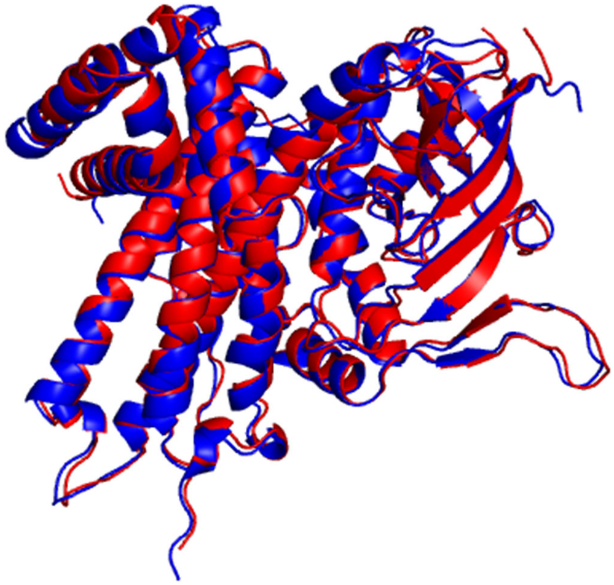


Figure S.37. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for T0965 and PDB ID 62dv
The structure in blue is the observed structure for PDB ID 62dv and the predicted structure for CASP13 target T0965 is in red . A TM-align of 0.81556 was achieved for protein structures. The score was normalised for PDB ID 62dv as it is the reference molecule,

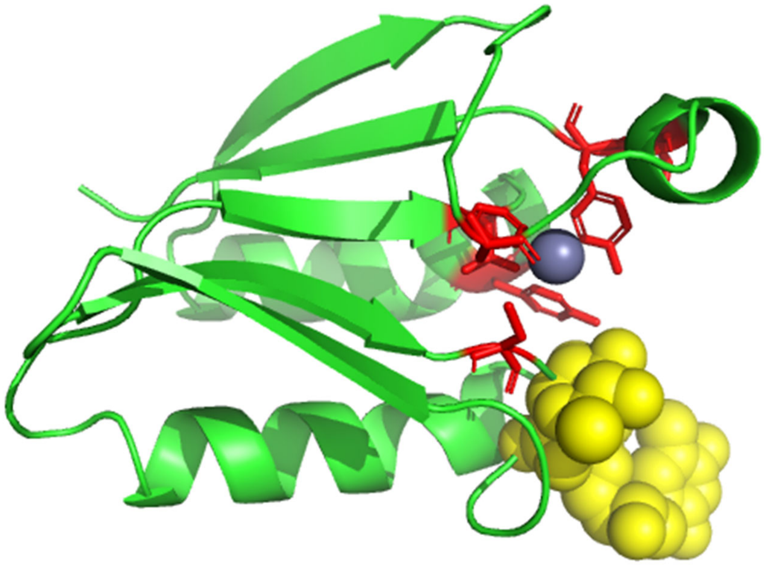


Figure S.38. FunFOLD3 ligand-binding site predictions for CASP13 target T0970 (PDB 6g57)
Predicted ligand binding site residues shown as sticks with predictions in red, the predicted ZN ligand shown as sphere and coloured blue and the dTDP-4-amino-4,6-dideoxyglucose (0FX) ligand shown as sphere and coloured yellow. No biologically relevant ligands were identified in the observed structure

The ninth predicted CASP13 target is Q6ZWB6 and the PDB entry classifies the protein target as human KCTD8 and is a ligase. The UniProtKB entry classifies the protein as BTB/POZ domain-containing protein KCTDB with experimental evidence at protein level,(UniProt Consortium, 2019) the highest level of evidence and demonstrates that there is clear experimental evidence for the existence of the protein and the criteria can include mass spectrometry, X-ray or NMR structure, this is supported by the data in the PDB entry. Additionally, the PDB entry has no ligands identified.

The UniProtKB entry states the function of the protein as an auxiliary subunit of GABA-B receptors that determine the pharmacology and kinetics of the receptor response. Increases agonist potency and markedly after the G-protein signalling of the receptors by accelerating onset and promoting desensitisation.(UniProt Consortium, 2019) The following GO – biological process terms were associated with the UniProtKB entry; GO:0051260 (protein homooligomerisation) and GO:0008277 (regulation of G protein-coupled receptor signalling pathway).

FunFOLD3 predicted the following GO terms listed below in Table S.7.

Table S.7. Predicted GO terms for CASP13 target T0970 (PDB ID 6g57)
The GO terms for CASP13 target T0970 (PDB ID 6g57) and their associated term domains and function are shown below. Biological process is coloured red and molecular function coloured green

GO term	GO term domain	Function
GO:0001510	Biological process	RNA methylation
GO:0032259	Biological process	methylation
GO:0005975	Biological process	carbohydrate metabolic process
GO:0006508	Biological process	proteolysis
GO:0006396	Biological process	RNA processing
GO:0008173	Molecular function	RNA methyltransferase activity
GO:0009020	Molecular function	tRNA (guanosine-2’-O-)-methyltransferase activity
GO:0016740	Molecular function	transferase activity
GO:0003824	Molecular function	catalytic activity
GO:0016810	Molecular function	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0004252	Molecular function	serine-type endopeptidase activity
GO:0003723	Molecular function	RNA binding
GO:0008168	Molecular function	Methyltransferase activity

As can be seen in Table S.7, the predicted function of the protein is related to RNA activity and is quite different to the function stated in the UniProtKB entry. The function as predicted by FunFOLD3 relates to the predicted zinc ligand as most RNA polymerases contain zinc and evidence suggests that zinc controls the stability of RNA polymerase.(Chanfreau, 2013) Additionally and directly related to the observed category of the protein, DNA ligases utilise zinc fingers to bind DNA containing secondary structures and to stimulate ligation of DNA strand breaks located near such structures.(Taylor, Whitehouse and Caldecott, 2000)

Figure S.39 below shows the TMalign superposition of observed and predicted structures from the PDB entry. A TM align score of 0.43860 was achieved, showing poor structural homology.

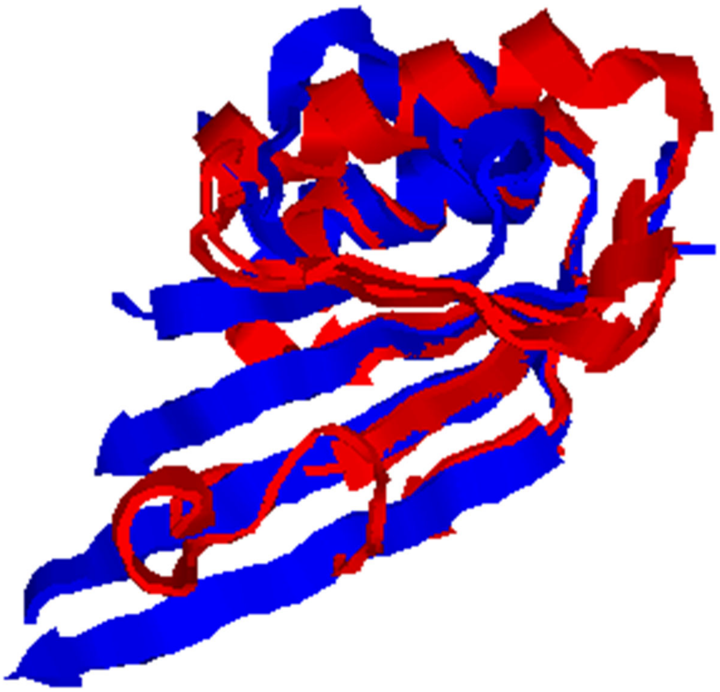


Figure S.39. Comparison of TMalign(Zhang and Skolnick, 2005) superposition for predicted and observed T0970
The structure in blue is the observed structure for T0970 and the predicted structure for CASP13 target T0970 is in red . A TM-align of 0.43860 was achieved for protein structures. The score was normalised for the observed T0970 target as it is the reference molecule

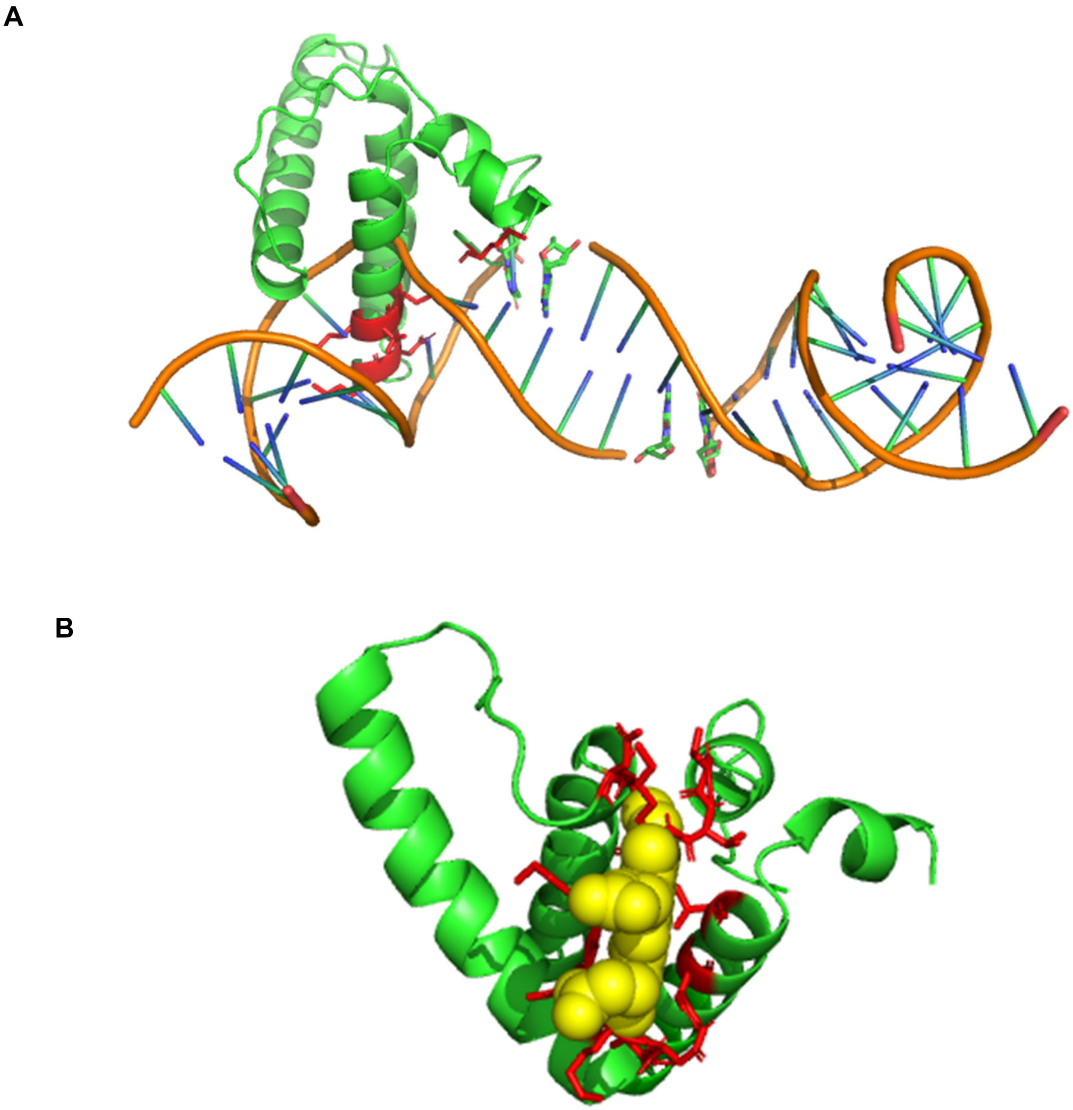


Figure S.40. FunFOLD3 ligand-binding site predictions for CASP13 target T0972
(A) Predicted ligand binding site residues shown as sticks with predictions in red, the predicted DNA ligand shown as double-helix **(B)** Predicted ligand binding site residues shown as sticks with predictions in red, the predicted HEM ligand is shown as sphere and coloured yellow. Structure was cancelled by CASP organisers so no observed structure has been released

The tenth predicted CASP13 target was T0972 and is classified as Q0P914_CAMJE. The structure was cancelled by the CASP13 organisers and no PDB entry is associated with this target therefore no comparisons can be made.

As per the UniProtKB entry the protein is uncharacterised and has the lowest annotation score of one out of five, denoting an entry with a rather basic annotation. Additionally, there is no evidence at protein, transcript or homology levels.(UniProt Consortium, 2019)

Similar proteins to Q0P914_CAMJE are invasion antigen CiaC and available literature describes the role of CiaC is to enable maximal invasion of host cells by *C.jejuni* and in part responsible for host cell cytoskeletal rearrangements that result in membrane ruffling.(Neal-McKinney and Konkel, 2012)

In terms of the predicted ligands, relaxation of DNA supercoiling leads to an increased ability of C.jejuni strains to penetrate human epithelial cells, hence potentially why DNA has been predicted as a ligand.(Scanlan *et al.*, 2017) No information in literature could be found about the role of haemoglobin.

Table S.8 below, shows the predicted GO term for T0972 which can serve to provide some potential insight into the function of T0972.

Table S.8. Predicted GO terms for CASP13 target T0972
The GO terms for CASP13 target T0972 (PDB ID 6g57) and their associated term domains and function are shown below. Biological process is coloured red, molecular function coloured green and cellular component is coloured purple

GO term	GO term domain	Function
GO:0006265	Biological process	DNA topological change
GO:0006259	Biological process	DNA metabolic process
GO:0022900	Biological process	electron transport chain
GO:0005694	Cellular Component	chromosome
GO:0042597	Cellular Component	periplasmic space
GO:0005746	Cellular Component	Mitochondrial respirasome
GO:0005524	Molecular function	ATP binding
GO:0003918	Molecular function	DNA topoisomerase type II
GO:0005506	Molecular function	iron ion binding
GO:0009055	Molecular function	electron transfer activity
GO:0020037	Molecular function	heme binding
GO:0003824	Molecular function	
GO:0004252	Molecular function	metal ion binding
GO:0003677	Molecular function	DNA binding

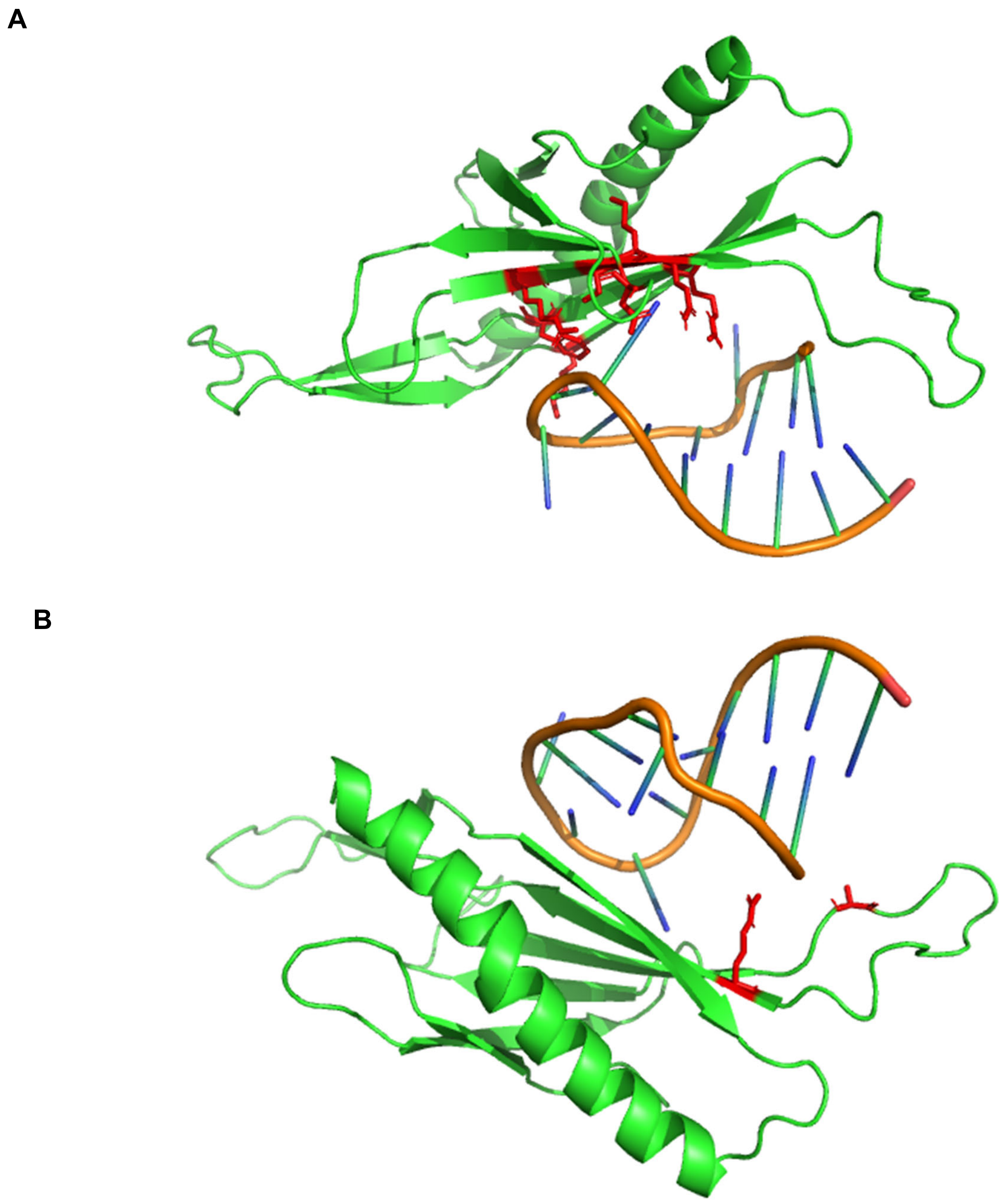


Figure S.41. FunFOLD3 ligand-binding site predictions for CASP13 target T0973 (PDB ID 6yfn)
(A) Predicted ligand binding site residues shown as sticks with predictions in red, the predicted RNA ligand shown as double-helix **(B)** Predicted ligand binding site residues shown as sticks with predictions in red, the predicted RNA ligand is shown as double-helix. Two different binding sites for the same ligand were identified for this target. No structure released by CASP organisers

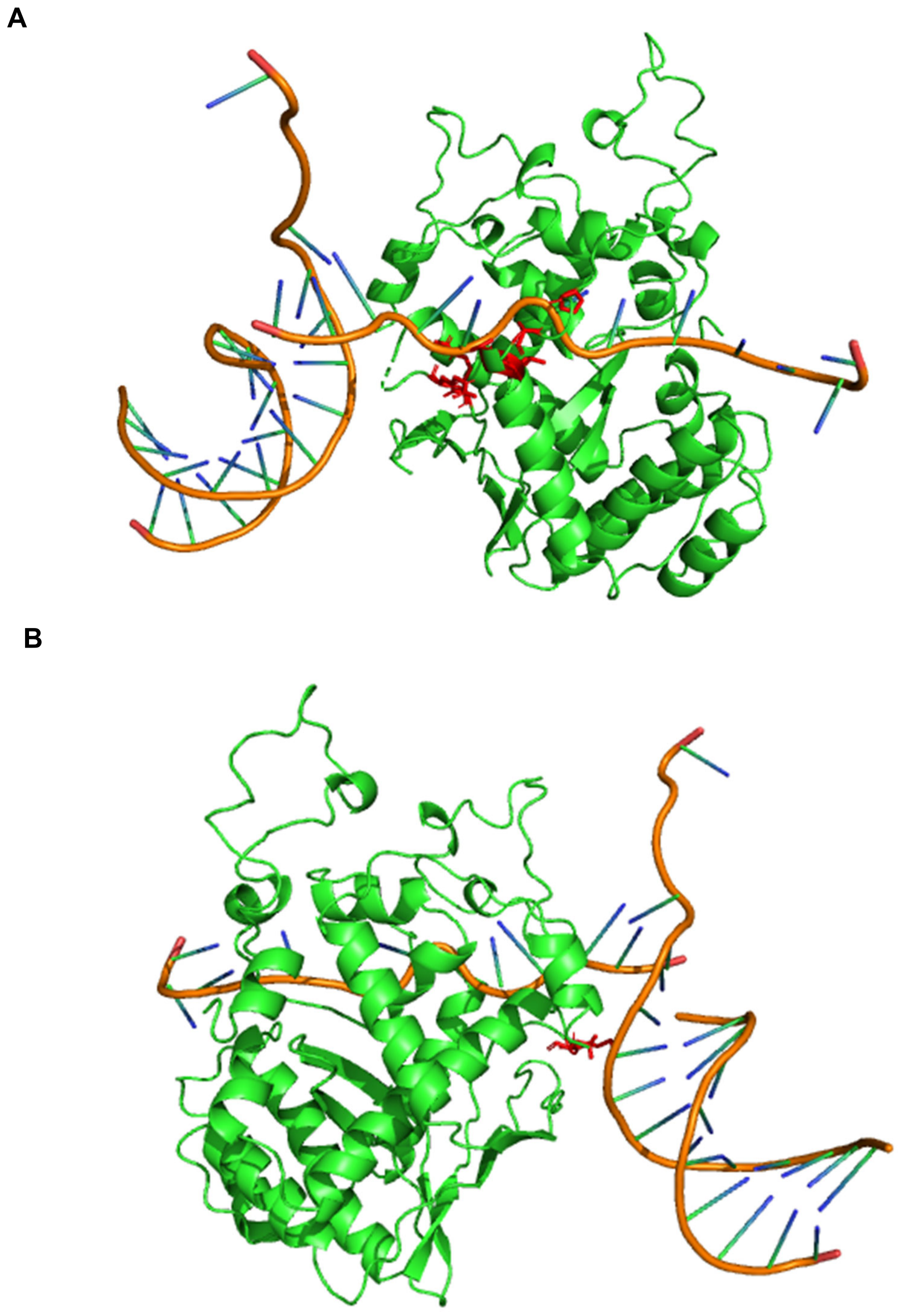
The eleventh CASP13 target is bacteriophage ESE058 coat protein and the associated PDB entry classifies the target as a virus like particle. No observed structure was released by CASP13 organisers and despite having a PDB entry, there is currently no pdb file available to download to enable comparisons between the predicted and observed structures. However, the PDB entry identifies calcium as a ligand. Furthermore, there is no entry for this protein on UniProtKB.

As a class of protein molecules, bacteriophages are composed of proteins that encapsulate DNA or RNA genome.(Liekniņa *et al.*, 2019) The single-stranded RNA (ssRNA) bacteriophages of the levivirdae family, as this target belows to, are a family of small viruses that infect a variety of gram-negative bacteria.(Liekniņa *et al.*, 2019) Based on this information, it would suggest that the RNA ligand predicted by FunFOLD3 is a rational conclusion. Additional information the ssRNA phage virus-like particles have found a variety of applications, mostly in the field of vaccine development where various antigens are presented onto the capsid surface to invoke a strong immune response.(Liekniņa *et al.*, 2019)

The GO terms predicted by FunFOLD3 are given below in table 20 and all the predictions relate to the available literature information on bacteriophages. In terms of similar templates which the target was modelled on were bacteriophage MS2 caspsid protein/RNA complex (PDB ID 1aq3), bacteriophage qbeta coat protein in complex with RNA operator hairpin (PBD ID 4l8h) and unusually MrkH, a novel c-di-GMP dependence transcription regulatory factor (PDB ID 5ejl)

Table S.9. Predicted GO terms for CASP13 target T0973
The GO terms for CASP13 target T0973 (PDB ID 6yfn) and their associated term domains and function are shown below. Molecular function coloured green and cellular component is coloured purple

GO term	GO term domain	Function
GO:0005694	Cellular Component	T=3 icosahedral viral capsid
GO:0019028	Cellular Component	viral capsid
GO:0019012	Cellular Component	virion
GO:0003723	Molecular function	RNA binding
GO:0005198	Molecular function	structural molecule activity


A
B

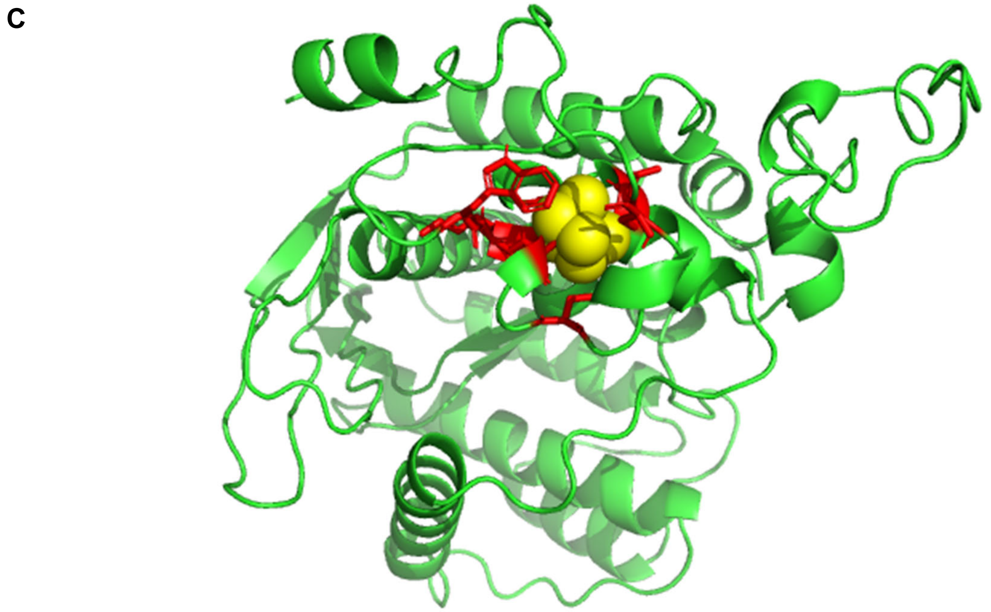


Figure S.42. FunFOLD3 ligand-binding site predictions for CASP13 target T0975
(A) Predicted ligand binding site residues shown as sticks with predictions in red, the predicted DNA ligand shown as double-helix**. (B)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted DNA ligand is shown as double-helix. Two different binding sites for the same ligand were identified for this target. **(C)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted ligand iron/sulphur cluster (SF4) shown as sphere and coloured *yellow*. No observed structure was released by CASP organisers

The thirteenth predicted CASP13 target is T0975 is EXO5 (saccharomyces cerevisiae exonuclease 5) and the UniProtKB entry classifies the function of the protein as a single-stranded (ssDNA) bidirectional exonuclease involved in DNA repair.(UniProt Consortium, 2019) Probably involved in DNA repair following ultraviolet (UV) irradiation and inter-strand cross-links (ICLs) damage.(UniProt Consortium, 2019) The protein is required for mitochondrial genome maintenance.(Sparks *et al.*, 2012)

Additionally, the protein has both 5’-3’ and 3’-5’ exonuclease activities with a strong preference for 5’ ends.(UniProt Consortium, 2019) This could potentially explain why two different binding site locations were predicted for the DNA molecule by FunFOLD3. EXO5 acts as a sliding exonuclease that loads ssDNA and then slides along the ssDNA prior to cutting,(Sparks *et al.*, 2012) once again, this could explain why the prediction DNA ligand is uncoiled in the prediction.

The protein has several co-factor binding sites of which includes a 4FE-4S cluster and also magnesium.(Sparks *et al.*, 2012) Available literature information about the protein, shows that that members of the EXO5 family share some common characteristics beyond that of primary amino acid sequence. They possess an iron-sulphur that is structurally important in linking the N terminus to the C terminus of the enzyme, thereby likely creating a cavity that may encircle the ssDNA.(Sparks *et al.*, 2012)

Table S.10 below, shows the GO terms associated with the protein as part of the UniProtKB entry. Additional GO terms predicted by FunFOLD3 are molecular function: GO:0016788 (hydrolase activity, acting on ester bonds), GO:0017111 (nucleoside-triphosphate activity), GO:0016787 (hydrolase activity), GO:0008854 (exodeoxyribonuclease V activity), GO:0005524 (ATP binding), GO:0005515 (protein binding), GO:0004527 (exonuclease activity), GO:0004519 (endonuclease activity), GO:0004518 (nuclease activity), GO:0004386 (helicase activity) and GO:0000166 (nucleotide binding). Biological process: GO:0000724 (double-strand break repair via homologous recombination), GO:0090305 (nucleic acid phosphodiester bond hydrolysis), GO:0006974 (cellular response to DNA damage stimulus), GO:0006310 (DNA recombination), GO:0006302 (double-strand break repair) and GO:0006281 (DNA repair). Cellular component GO:0009338 (exodeoxyribonuclease V complex).

Table S.10. Predicted GO terms for CASP13 target T0975
The GO terms for CASP13 target T0975 and their associated term domains and function as per the UniProtKB entry are shown below . The association with FunFOLD3 is denoted in the final column with exact matches denoted with a tick and related GO terms with the associated function. Biological process is coloured red and molecular function coloured green

GO term	GO term domain	Function	FunFOLD3
GO:0051539	Molecular function	4 iron, 4 sulfur clustering binding	✓ (additionally GO:0051536 iron-sulfur cluster binding)
GO:0003677	Molecular function	DNA binding	✓
GO:0046872	Molecular function	metal ion binding	✓
GO:0008310	Molecular function	single-stranded DNA 3'-5' exodeoxyribonuclease activity	N/A

GO:0045145	Molecular function	single-stranded DNA 5'-3' exodeoxyribonuclease activity	N/A
GO:0036297	Biological process	interstrand cross-link repair	N/A

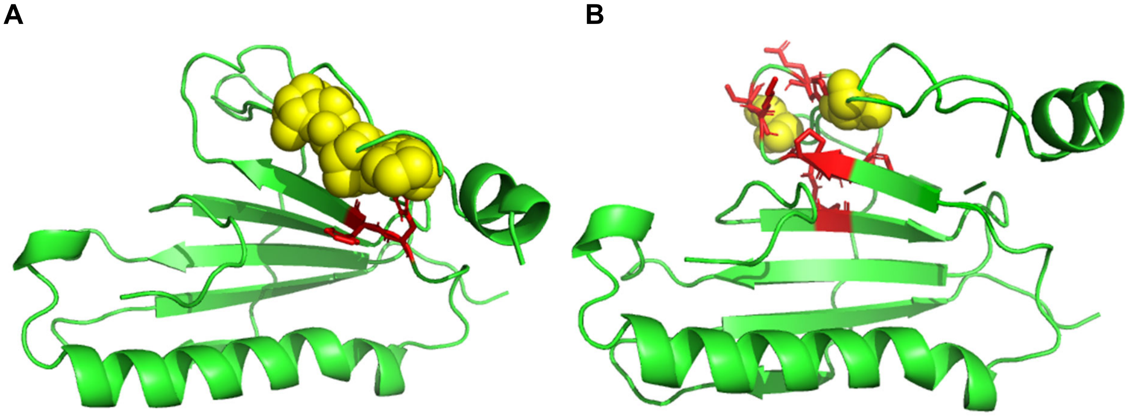


Figure S.43. FunFOLD3 ligand-binding site predictions for CASP13 target T0980s1 (PBD ID 6gnx)
(**A**) Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted proline (PRO) ligand shown as sphere and coloured *yellow* **(B)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted serine (SER) ligand is shown as sphere and coloured *yellow*. Two different binding sites for the same ligand were identified for this target. No observed structure was released by CASP organisers and no biologically relevant ligands were found in the PDB structure

The fourteenth predicted CASP13 target is Q3KP22-3; Q8NHR7 and the PDB entry associated with this target identifies the protein as structural protein. The UniProtKB entry states the protein as a membrane-anchored junction protein.(UniProt Consortium, 2019)

Additionally, the UniProtKB entry states the function of the proteins as a meiosis-specific telomere-associated protein involved in meiotic telomere attachment to the nucleus inner membrane, a crucial step for homologous pairing and synapsis.(UniProt Consortium, 2019) Q3KP22-3 is an isoform of the protein and is identified as isoform 1(UniProt Consortium, 2019) and differs from isoform 3 in terms of canonical sequence.(UniProt Consortium, 2019)

Component of the MAJIN-TERB1-TERB2 complex, which promotes telomere cap exchange by mediating attachment of the telomeric DNA to the inner nuclear membrane and replacement of the protective cap of telomeric chromosomes, in early meiosis, the MAJIN-TERB1-TERB2 complex associates with telomeric DNA and the shelterin/telosome complex.(UniProt Consortium, 2019)

The UniProtKB entry denotes the following GO terms as shown in Table S.11 which are associated with the protein and provide insight into the function of the protein. In comparison, FunFOLD3 predicted the following GO terms as shown in Table S.12.

Table S.11. Predicted GO terms for CASP13 target T0980s1 (PDB ID 6gnx)
The GO terms for CASP13 target T0980s1 and their associated term domains and function as per the UniProtKB entry are shown below. Biological process is coloured red and molecular function coloured green

GO term	GO term domain	Function
GO:0003677	Molecular function	DNA binding
GO:0070197	Biological process	meiotic attachment of telomere to nuclear envelope
GO:0045141	Biological process	meiotic telomere clustering
GO:0007129	Biological process	homologous chromosome pairing at meiosis

Table S.12. Predicted GO terms for CASP13 target T0980s1 (PDB ID 6gnx) as predicted by FunFOLD3
The GO terms for CASP13 target T0980s1 (PDB ID 6gnx and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0003723	Molecular function	RNA binding
GO:0004652	Molecular function	polynucleotide adenylyltransferase activity
GO:0016779	Molecular function	nucleotidyltransferase activity
GO:0004527	Molecular function	exonuclease activity
GO:0016787	Molecular function	hydrolase activity
GO:0005515	Molecular function	protein binding
GO:0008432	Molecular function	JUN kinase binding
GO:0042803	Molecular function	protein homodimerization activity
GO:0006259	Biological process	DNA metabolic process
GO:0043631	Biological process	RNA polyadenylation
GO:0031123	Biological process	RNA 3'-end processing
GO:0006351	Biological process	transcription, DNA-templated
GO:0001558	Biological process	regulation of cell growth
GO:0006281	Biological process	DNA repair
GO:0006302	Biological process	double-strand break repair
GO:0006355	Biological process	regulation of transcription, DNA-templated
GO:0006974	Biological process	cellular response to DNA damage
GO:0007049	Biological process	cell cycle
GO:0007094	Biological process	mitotic spindle assembly checkpoint
GO:0033188	Biological process	positive regulation of peptidyl-serine phosphorylation
GO:0042177	Biological process	negative regulation of protein catabolic process
GO:0042771	Biological process	DNA damage response, signal transduction resulting in transcription
GO:0045893	Biological process	Positive regulation of transcription, DNA-templated

GO:0051301	Biological process	cell division
GO:0000070	Biological process	mitotic sister chromatid segregation
GO:0000075	Biological process	cell cycle checkpoint
GO:0000087	Biological process	mitotic M phase
GO:0000236	Biological process	mitotic prometaphase
GO:0000278	Biological process	mitotic cell cycle
GO:0007093	Biological process	mitotic cell cycle checkpoint
GO:0031145	Biological process	anaphase-promoting complex-dependent catabolic process
GO:0043066	Biological process	negative regulation of apoptotic process
Go:0090267	Biological process	positive regulation of mitotic cell cycle spindle assemble checkpoint
GO:0005634	Cellular Component	nucleus
GO:0005654	Cellular Component	neoplasm
GO:0005737	Cellular Component	cytoplasm
GO:0005819	Cellular Component	spindle
GO:0005856	Cellular Component	cytoskeleton
GO:0016035	Cellular Component	zeta DNA polymerase complex
GO:0005680	Cellular Component	anaphase-promoting complex
GO:0000775	Cellular Component	chromosome centromeric region
GO:0000776	Cellular Component	kinetochore
GO:0000777	Cellular Component	condensed chromosome kinetochore
GO:0000922	Cellular Component	spindle pole
GO:0005694	Cellular Component	chromosome
GO:0005829	Cellular Component	cytosol
GO:0048471	Cellular Component	perinuclear region of cytoplasm
GO:0005643	Cellular Component	nuclear pore

Whilst the GO terms predicted by FunFOLD3 do not match the GO terms associated with the UniProtKB entry for the target. There are some similarities with the predictions. Namely around, mitotic functions (GO:0007093, GO:0000087, GO:0000236, GO:0000278) RNA binding (similarity to the DNA binding), nucleus (GO:0005634), the latter which matches the subcellular location of the protein as per UniProtKB.(UniProt Consortium, 2019) Membrane-anchored junction protein shows DNA-binding activity, possible for the stabilisation of telomere attachment on the nucleus inner membrane.(UniProt Consortium, 2019)

Figure S.44 below shows the TM-align superposition between the predicted model and observed structure from the PDB entry. A TM align score of 0.34981 was achieved, showing poor structural homology.

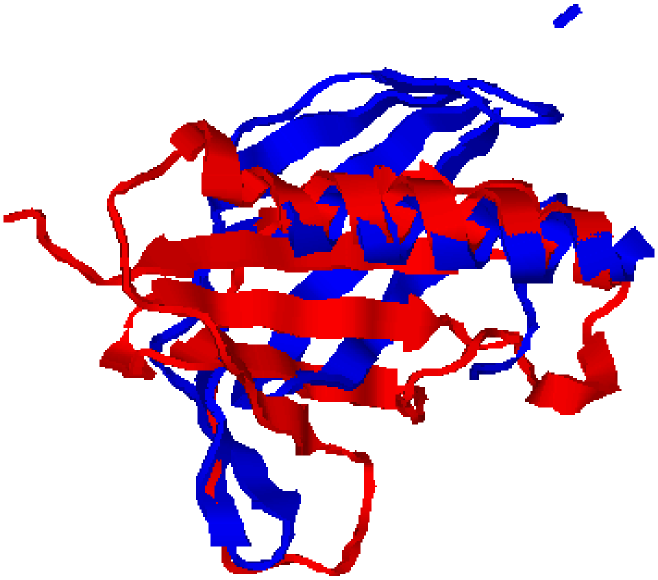


Figure S.44. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted T0980s1 and PDB ID 6gnx
The structure in blue is the observed structure for PDB ID 6gnx and the predicted structure for CASP13 target T0980s1 is in red. A TM-align of 0.34981 was achieved for protein structures. The score was normalised for PDB ID 6tri target as it is the reference molecule

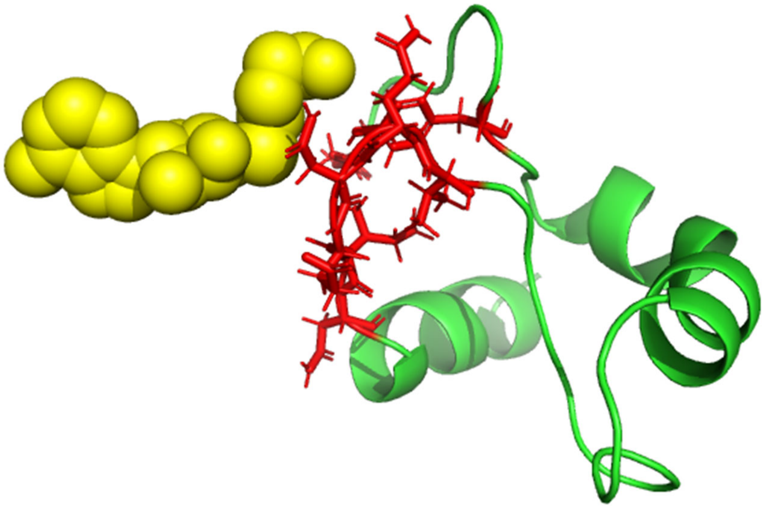


Figure S.45. FunFOLD3 ligand-binding site predictions for CASP13 target T0980s2 (PBD ID 6gnx)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted s-adenosyl-l-homocysteine (SAH) ligand shown as sphere and coloured *yellow***.** No observed structure was released by CASP organisers organisers and no biologically relevant ligands were found in the PDB structure

The fifteenth predicted CASP13 target is subunit 2 of Q3KP22-3;Q8NHR7. As with T0980s1, no biologically relevant ligands were found in the PDB structure. As T0980s2 is related to T0980s1, no further literature information is available on the target. However, FunFOLD3 predicted the following GO terms and are specific to this target. The predictions are different to the GO terms for T0980s1 and are not similar or complimentary to T0980s1.

Table S.13. Predicted GO terms for CASP13 target T0980s2 (PDB ID 6gnx) as predicted by FunFOLD3
The GO terms for CASP13 target T0980s2 (PDB ID 6gnx) and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0000166	Molecular function	nucleotide binding
GO:0000287	Molecular function	magnesium ion binding
GO:0004363	Molecular function	glutathione synthase activity
GO:0005524	Molecular function	ATP binding
GO:0016874	Molecular function	ligase activity
GO:0046872	Molecular function	metal ion binding
GO:0004324	Molecular function	Ferredoxin-NADP+ reductase activity
GO:0016491	Molecular function	oxidoreductase activity
GO:0050660	Molecular function	flavin adenine dinucleotide binding
GO:0050661	Molecular function	NADP binding
GO:0006750	Biological process	glutathione biosynthetic process
GO:0055114	Biological process	oxidation-reduction process
GO:0005829	Cellular Component	cytosol
GO:0009579	Cellular Component	thylakoid
GO:0016020	Cellular Component	membrane
GO:0030089	Cellular Component	phycobilisome
GO:0042651	Cellular Component	thylakoid membrane

Figure S.46 below shows the TM-align superposition between the predicted model and observed structure from the PDB entry. A TM-align score of 0.21242 was achieved, showing poor structural homology. This poor structural homology could also explain why the GO terms and ultimately function predictions are not aligned with the observed protein.

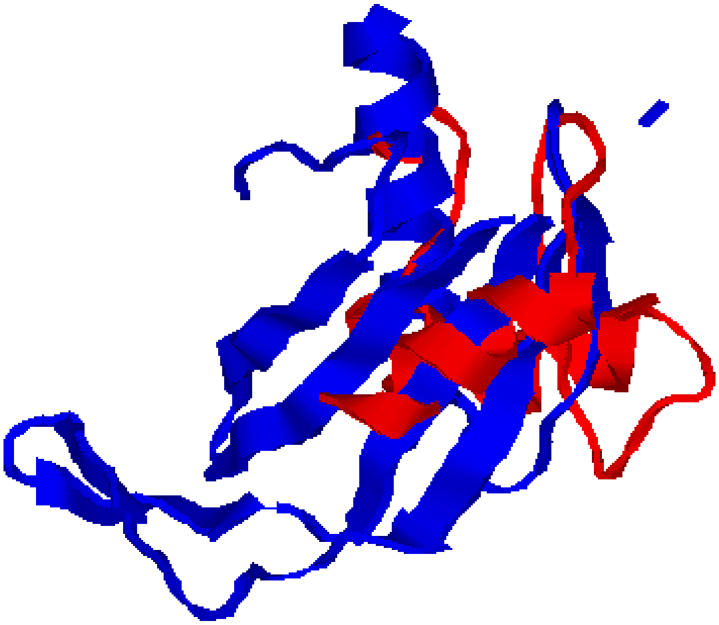


Figure S.46. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted T0980s2 and PDB ID 6gnx
The structure in blue is the observed structure for PDB ID 6gnx and the predicted structure for CASP13 target T0980s2 is in red. A TM-align of 0.21242 was achieved for protein structures. The score was normalised for PDB ID 6tri target as it is the reference molecule

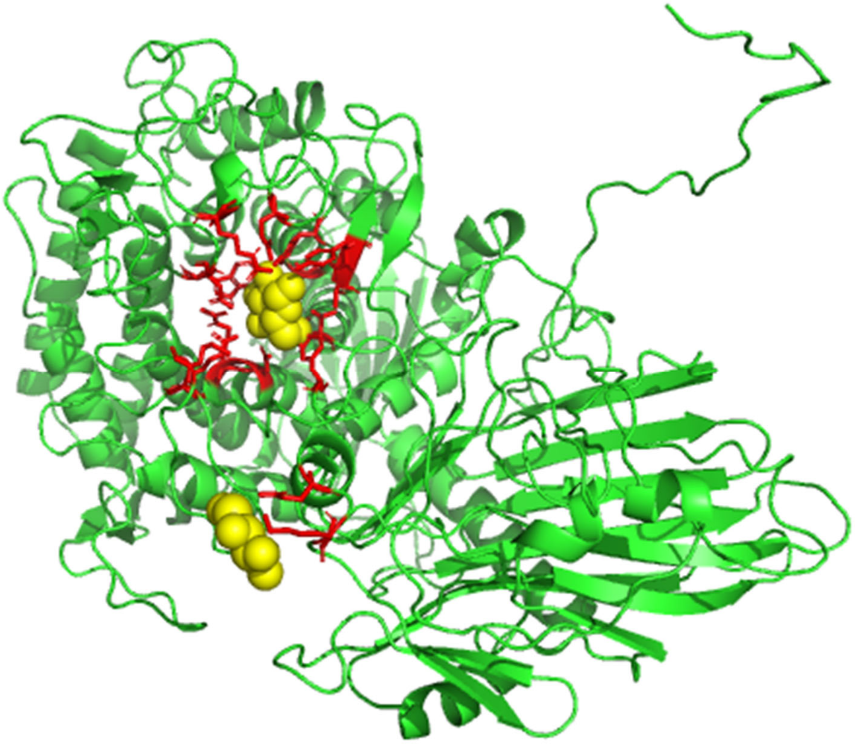


Figure S.47. FunFOLD3 ligand-binding site predictions for CASP13 target T0985
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted beta-D-glucopyranose (BGC) ligand shown as sphere and coloured *yellow***.** Two different binding sites for the same ligand were identified for this target. No observed structure was released by CASP organisers

The seventeenth predicted CASP13 target is ACL_1061, there was no structure released by the CASP13 organisers and no PDB ID is associated with the target. UniProtKB identifies the protein as glycol_hydro_36 domain-containing protein and ACL_1061 is the gene.(UniProt Consortium, 2019) In terms of annotation status, the protein has the lowest score of one out of five and is deemed ‘protein predicted’ and denotes there is no evidence at protein, transcript or homology levels.(UniProt Consortium, 2019)

In terms of function of the protein, the UniProtKB entry has transferase activity (GO:0016740) for the function of the protein. The GO terms predicted by FunFOLD3 are provided in Table S.14 below. The GO term related to transferase activity has also been predicted by FunFOLD3.

Table S.14. Predicted GO terms for CASP13 target T0985 as predicted by FunFOLD3
The GO terms for CASP13 target T0985 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0003824	Molecular function	catalytic activity
GO:0030246	Molecular function	carbohydrate binding
GO:0004348	Molecular function	glucosylceramidase activity
GO:0046872	Molecular function	metal ion binding
GO:0016740	Molecular function	transferase activity
GO:0016757	Molecular function	transferase activity, transferring glycosyl groups
GO:0047738	Molecular function	cellobiose phosphorylase activity
GO:0005975	Biological process	carbohydrate metabolic process
GO:0006665	Biological process	Sphingolipid metabolic process
GO:0016021	Cellular Component	integral component of membrane

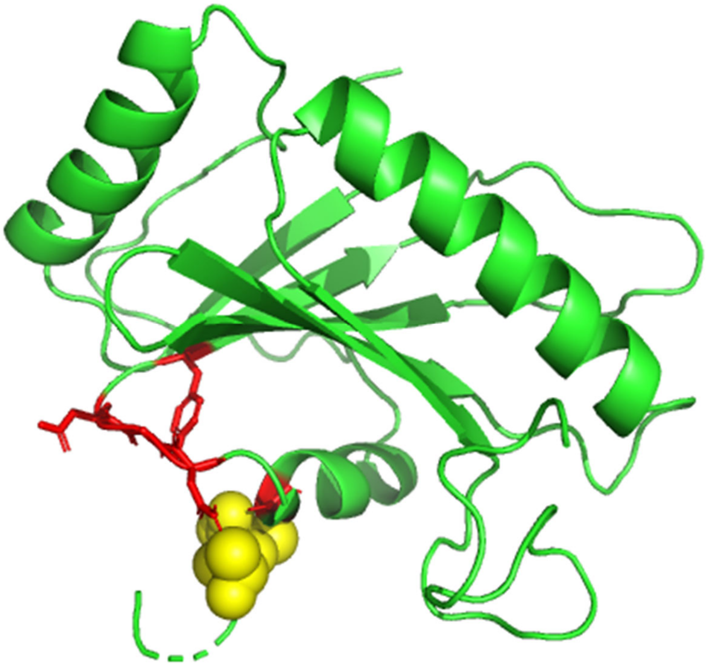


Figure S.48. FunFOLD3 ligand-binding site predictions for CASP13 target T0986s2 (PDB ID 6d7y)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted 2-hydroxybenzoic acid (DAL) ligand shown as sphere and coloured *yellow*. No observed structure was released by CASP organisers and no biologically relevant ligands were found in the PDB file

The eighteenth predicted CASP13 target is toxic C-terminal tip of CdiA and immune protein and is classified as classified as toxin as per the PDB entry and there is no ligand associated with the target. The UniProtKB entry provides quite extensive information into the function of Toxic CdiA proteins.(UniProt Consortium, 2019)

Toxin CdiA proteins are toxic component of a toxin-immunity protein module, which functions as a cellular contact-dependent growth inhibition (CDI) system.(UniProt Consortium, 2019) CDI modules allow bacteria to communicate with and inhibit the growth of closely related neighboring bacteria in a contact-dependent fashion.(UniProt Consortium, 2019) The CdiA protein is thought to be exported from the cell through the central lumen of CdiB, the other half of its two-partner system (TPS).(UniProt Consortium, 2019)

Finally, in terms of related GO terms; toxin activity (GO:0090729), cell adhesion (GO:0007155), and pathogenesis (GO:0009405), none of these terms were predicted by FunFOLD3.

Figure S.49 below is of the TM-align superposition of the predicted and observed protein structure, with the observed structure from the PDB entry for the target. A TM-align score of 0.31392 was achieved showing poor structural homology between the predicted and observed structure.

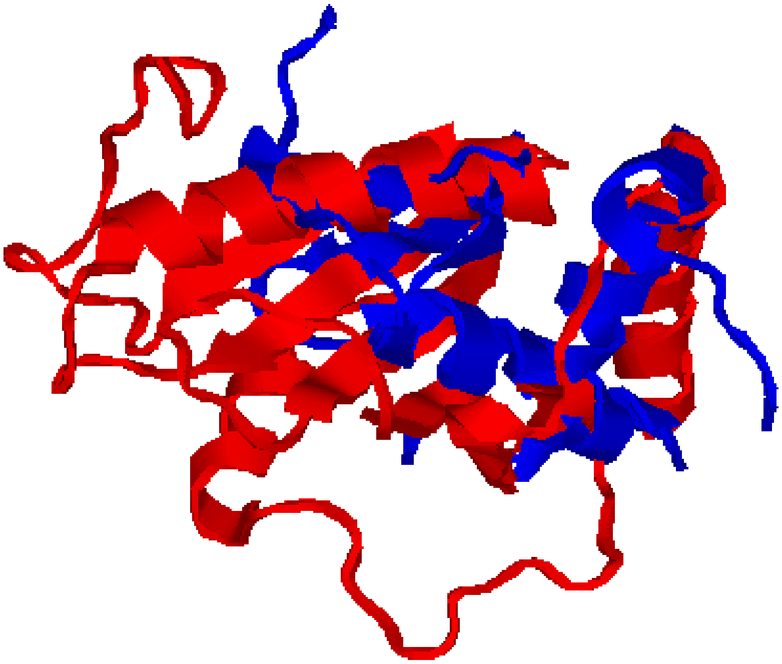


Figure S.49. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted T0986s2 and PDB ID 6d7y
The structure in blue is the observed structure for PDB ID 6d7y and the predicted structure for CASP13 target T0986s2 is in red. A TM-align of 0.31392 was achieved for protein structures. The score was normalised for PDB ID 6d7y target as it is the reference molecule

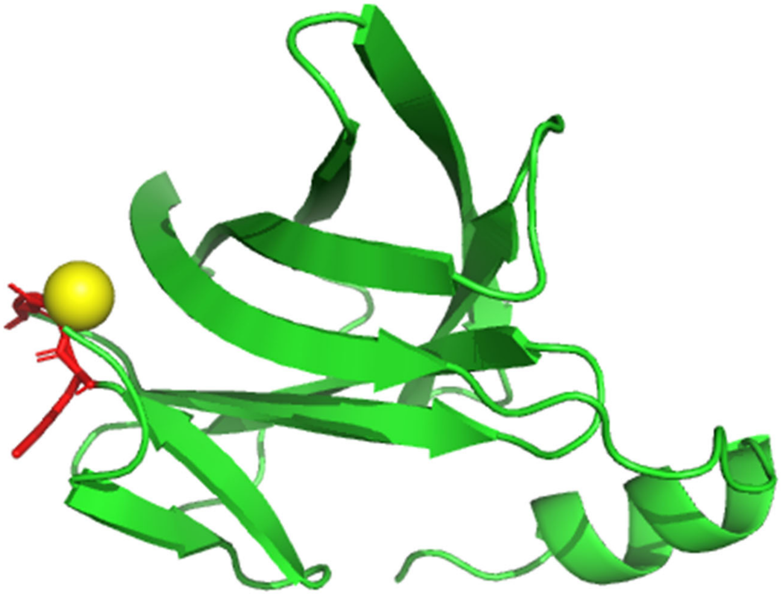


Figure S.50. FunFOLD3 ligand-binding site predictions for CASP13 target T0992
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted CA ligand shown as sphere and coloured *yellow*. No observed structure was released by CASP organisers and no PDB ID is associated with the target

The nineteenth predicted CASP13 target is Q6MKZ7, no structure was released by CASP13 organisers and no PDB ID is associated with the target. UniProtKB identifies the protein from gene Bd2229 and is an uncharacterised protein.(UniProt Consortium, 2019) In terms of status the protein has the lowest level of annotation and has been predicted, meaning there is no evidence at protein, transcript or homology levels.(UniProt Consortium, 2019)

As there is limited information available on the target, information related to FunFOLD3 will be presented to provide potential insight into the target. Table S.15, below shows the GO terms predicted by FunFOLD3 with their associated functions.

Table S.15. Predicted GO terms for CASP13 target T0992 as predicted by FunFOLD3
The GO terms for CASP13 target T0992 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0003824	Molecular function	catalytic activity
GO:0016837	Molecular function	carbon-oxygen lyase activity, acting on polysaccharides
GO:0016829	Molecular function	lyase activity
GO:0030246	Molecular function	carbohydrate binding
GO:0005975	Biological process	carbohydrate metabolic process
GO:0005576	Cellular Component	extracellular region

In terms of templates which the prediction has been based, the following information is available; 2qlf (caspase-7 and classified as hydrolase), 3jqw (collagenase and classified as cell adhesion), 3sik (bacillus hemophore IsdX1 and classified as transport protein), 3zm8 (mannanases and classified as hydrolase), 4ruw (endonuclease/exonuclease/phosphatase and classified as hydrolase), 5c2v (hydrazine synthase and classified as oxidoreductase), 5mnw (cinaciguat classified as a lyase), 5otl (CK2alpha classified as transferase) and 6dk4 (campylobacter jejuni peroxide and classified as metal transport).

In terms of the templates with the same ligand 3jqw, 3zm8 and 5c2v all have calcium as a ligand. The classification of each of the proteins is quite different so comparisons to a specific protein would be quite difficult. For example 3jqw, collagen-binding derived from collagenase G (s3b) in the presence of Ca^{2+} shows shortened hydrodynamic radius, better stability and more efficient substrate binding.(Bauer *et al.*, 2013) Additionally, X-ray crystal structures of s3b were solved in the presence of Ca^{2+} (holo) as well as in the abense of Ca^{2+} (apo) to show a secondary structure transformation of the linker at its N terminus.(Bauer *et al.*, 2013) In terms of 3zm8, β-mannanase are encountered as modular enzymes and some harbour carbohydrate binding modules (CBM), one of which is CBM35.(Couturier *et al.*, 2013) Calcium has been identified in the structure of CBM35 and is involved in carbohydrate recognition,(Bauer *et al.*, 2013) which ties in nicely with two of the GO terms predicted by FunFOLD3 (GO:0030246 and GO:0005975). There is no information in literature to relate the role of calcium to 5c2v.

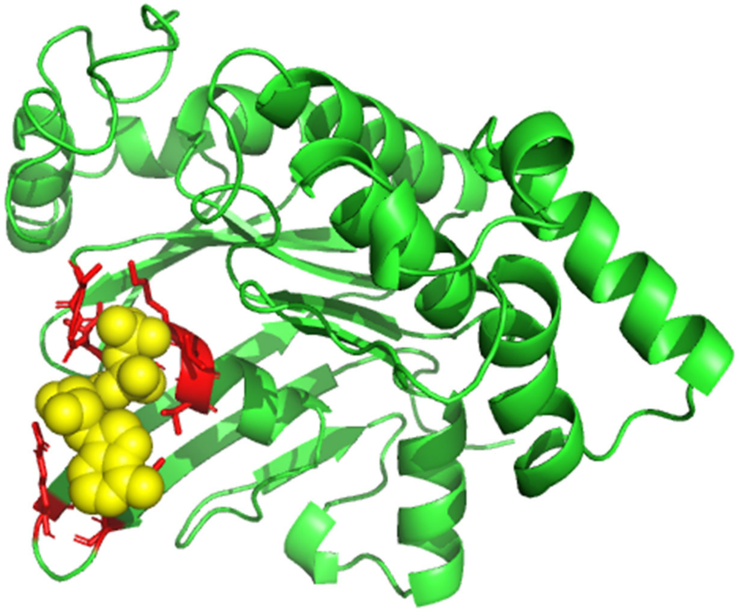


Figure S.51. FunFOLD3 ligand-binding site predictions for CASP13 target T0993s1 (PDB ID 6xbd)
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted adenosine-5-‘phosphate (ADP) ligand shown as sphere and coloured *yellow*. No observed structure was released by CASP organisers organisers and no biologically relevant ligands were found in the PDB file

The twentieth predicted CASP13 target is MiaFA and is classified as lipid transport as per the PDB entry. UniProtKB information around a similar protein, called intermembrane phospholipid transport system binding protein MiaD identifies the function as part of the ABC transporter complex MiaFEDB,which is involved in a phospholipid transport pathway that maintains lipid asymmetry in the outer membrane by retrograde trafficking of phospholipids from the outer membrane to the inner membrane(Malinverni and Silhavy, 2009),(Thong *et al.*, 2016) MiaD functions in substrate binding with strong affinity for phospholipids and modulates ATP hydrolytic activity of the complex.(Thong *et al.*, 2016) Based on this information from the UniProtKB entry, it seems reasonable as to why ATP was predicted as a ligand. However, no ligands are associated with the PDB entry, therefore the identification of ATP would be based on 1b0u, a transport protein and 1ji0 also a transport protein.

Figure S.51 below, shows the TMalign superposition between the predicted and observed protein structure, with the observed structure from the PDB entry for the target. A TM-align score of 0.24098 was achieved showing poor structural homology between the predicted and observed structure.

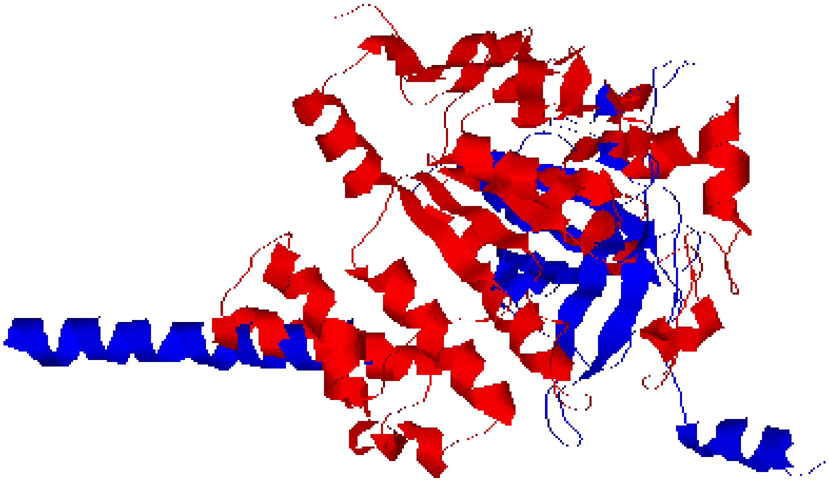


Figure S.52. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted T0993s1 and PDB ID 6xbd
The structure in blue is the observed structure for PDB ID 6xbd and the predicted structure for CASP13 target T0993s1 is in red. A TM-align of 0.24098 was achieved for protein structures. The score was normalised for PDB ID 6xbd target as it is the reference molecule

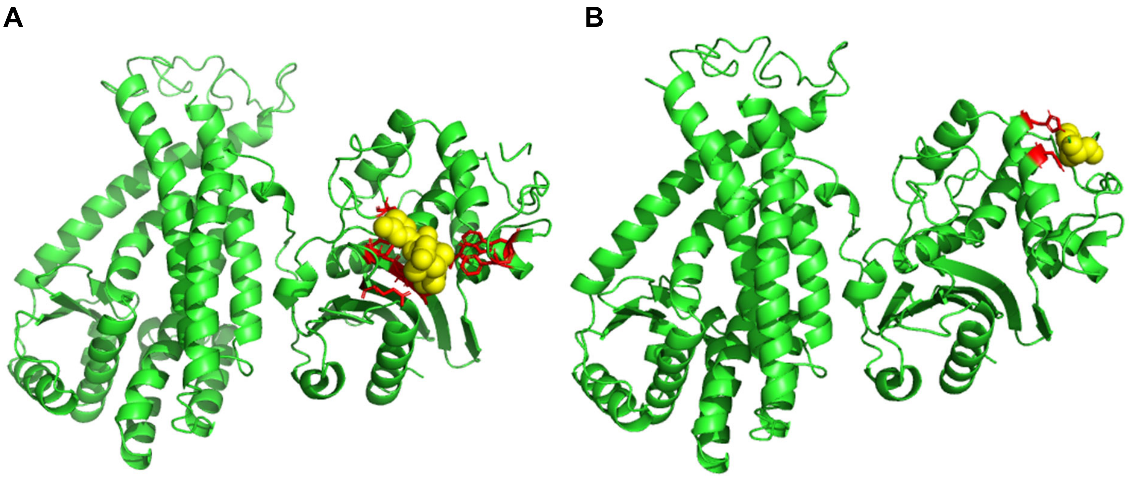


Figure S.53. FunFOLD3 ligand-binding site predictions for CASP13 target T0994
(A) Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted acylated ceftazidime (CAZ) ligand shown as sphere and coloured *yellow* **(B)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted histidine (HIS) ligand is shown as sphere and coloured *yellow*. Two different binding sites for the same ligand were identified for this target. Structure cancelled by CASP organisers

The twenty-first predicted CASP13 target is Q79ER8_STAAU and no PDB ID is associated with this target and the CASP13 organisers cancelled the structure, therefore no analysis can be made against the observed structures.

The UniProtKB entry identifies Q79ER8_STAAU as a beta-lactam sensor/signal transducer MecR1. In terms of annotation, the lowest score with one out of five is associated with the protein and the protein is inferred by homology and indicates that the existence of a protein is probable because clear orthologs exist in closely related species.(UniProt Consortium, 2019)

In relation to GO terms penicillin binding (GO:0008658) and integral component of membrane (GO:0016021). In comparison the GO terms predicted by FunFOLD3 are given below in Table S.16.

Table S.16. Predicted GO terms for CASP13 target T0994 as predicted by FunFOLD3
The GO terms for CASP13 target T0994 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0000166	Molecular function	nucleotide binding
GO:0015408	Molecular function	ATPase-coupled ferric iron transmembrane transporter activity
GO:0016787	Molecular function	hydrolase activity
GO:0005524	Molecular function	ATP binding
GO:0016887	Molecular function	ATPase activity
GO:0017111	Molecular function	nucleoside-triphosphatase activity
GO:0005215	Molecular function	transporter activity
GO:0015426	Molecular function	ATPase-coupled polar amino acid-transporter activity
GO:0015423	Molecular function	ATPase-coupled maltose transmembrane transporter activity
GO:0043865	Molecular function	methionine transmembrane transporter activity
GO:0048474	Molecular function	D-methionine transmembrane transporter activity
GO:0046872	Molecular function	metal ion binding
GO:0006200	Biological process	obsolete ATP catabolic process
GO:0006810	Biological process	transport
GO:0008643	Biological process	carbohydrate transport
GO:0015768	Biological process	maltose transport
GO:0042956	Biological process	maltodextrin transport
GO:0006865	Biological process	amino acid transport
GO:0015821	Biological process	methionine transport
GO:0048473	Biological process	D-methionine transport
GO:0005886	Cellular Component	plasma membrane
GO:0043190	Cellular Component	ATP-binding cassette (ABC) transporter complex
GO:0016020	Cellular Component	membrane
GO:0009276	Cellular Component	Gram-negative-bacterium-type cell wall

FunFOLD3 did not predict any GO terms which were associated with the protein as per UniProtKB. The closest GO terms could be GO:0009276 and the protein is a beta-lactam and as antibiotics have activity on both gram-negative and gram-positive bacteria.(Fisher and Mobashery, 2016) In terms of the predicted ligand by FunFOLD3, ceftazidime is a novel β-lactamase inhibitor with activity against multi-drug resistant gram-negative bacteria.(Zasowski, Rybak and Rybak, 2015) The role of this predicted ligand ties in with the predicted GO term 0009276 and also the protein as a beta-lactam.

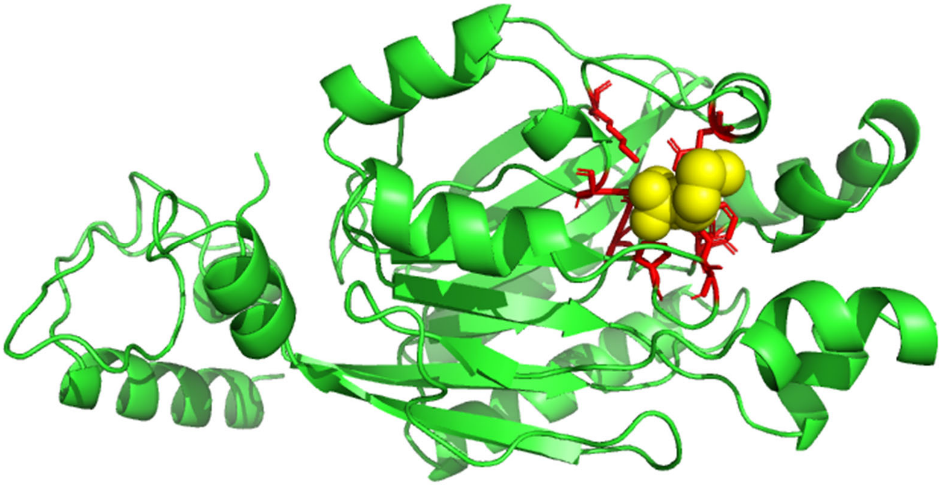


Figure S.54. FunFOLD3 ligand-binding site predictions for CASP13 target T0995
Predicted ligand binding site residues shown as sticks with predictions in red, the predicted 4-methylsulfanyl-2-ureido-butyric acid (CDT) ligand shown as sphere and coloured yellow. No observed structure was released by CASP organisers

The twenty-second predicted CASP13 target is B3GNT7 (B3GNTY_BACPU) and no PDB ID is associated with this entry. The UniProtKB entry identifies the protein as cyanide dehydratase(UniProt Consortium, 2019) and as expected of a protein target with no PDB ID, has a low annotation score with one out of five and has the status of ‘protein predicted’.

Information about function states an active site and this subsection of function relates specifically to enzymes and indicates the residues directly involved in catalysis. The active site position is 48 as per UniProtKB(UniProt Consortium, 2019) and the predicted ligand-binding site residues were 48,54,130,137,164,189 and 192. The description of the site is proton acceptor.

GO terms associated with the protein are nitrilase activity (molecular function, GO:0000257) and nitrogen compound metabolic process (biological process, GO:0006807). In comparison the GO terms predicted by FunFOLD3 are given below in Table S.17 with the matched GO term for biological process. Furthermore, the family and domains section of the entry identifies position of 8-270 as a CN hydrolase of which FunFOLD3 has predicted terms related to this (GO:0047417 and GO:0016787).(UniProt Consortium, 2019)

Table S.17. Predicted GO terms for CASP13 target T0995 as predicted by FunFOLD3
The GO terms for CASP13 target T0995 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0047417	Molecular function	N-carbamoyl-D-amino acid hydrolase activity
GO:0016787	Molecular function	hydrolase activity
GO:0016810	Molecular function	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0003824	Molecular function	catalytic activity
GO:0047710	Molecular Function	bis(5'-adenosyl)-triphosphatase activity
GO:0006807	Biological process	nitrogen compound metabolic process
GO:0006139	Biological process	nucleobase-containing compound metabolic process
GO:0008152	Biological process	metabolic process
GO:0005575	Cellular Component	cellular component

In terms of templates, 1ems is classified as an antitumor protein and is NITFHIT protein and N-carbamyl-D-amino acid amidohydrolase and classified as hydrolase and has CDT as a ligand. No other further information can be found about the protein in literature.

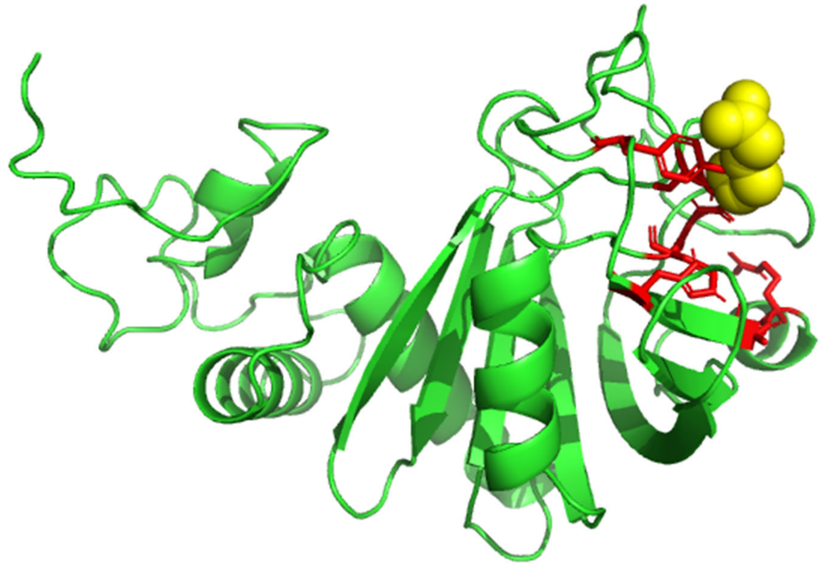


Figure S.55. FunFOLD3 ligand-binding site predictions for CASP13 target T0997
Predicted ligand binding site residues shown as sticks with predictions in red, the predicted d-glutamic acid (DGL) ligand shown as sphere and coloured yellow**.** No observed structure was released by CASP organisers

The twenty-third CASP13 target is Q6MN59, as uncharacterised protein from gene Bd1402 as per the UniProtKB entry for the protein.(UniProt Consortium, 2019) The status of the protein is unreviewed and has the lowest annotation score with one out of five and has the status of protein predicted for the level of existence. As with all proteins with limited information, there is no PDB ID associated with the target.

Post-translation modifications/processing of the protein occurs and there is the presence of an N-terminal signal peptides.(UniProt Consortium, 2019) Signal peptides are found in proteins that are targeted to the endoplasmic reticulum and eventually destined to be either secreted/extracellular/periplasmic, retained in the lumen of the endoplasmic reticulum, of the lysosome or of any other organelle along the secretory pathway or to be I single-pass membrane proteins.(UniProt Consortium, 2019)

FunFOLD3 did not predict any GO terms for this target. However, the following templates and their roles were predicted; 3tur (M.tuberculosis LD-transpeptidase and classified as peptidoglycan binding protein), 3vyp (mycobacterium tuberculosis L,D-transpeptidase and classified as transferase), 4k73 (L,D-transpeptidase and classified as transferase), 4x9l (heat shock protein and classified as chaperone), 5k69 (mycobacterium tuberculosis L,D-transpeptidase 2 classified as transferase/transferase inhibitor), 5kis (YenB/RHS2 complex, classified as a toxin), 5mps (spliceosome and classified as splicing) and 6br8 (A6 and classified as a viral protein). Based on the predicted templates there appears to be come consensus, with mycobacterium tuberculosis L,D-transpeptidase being the most popular. Additionally, 3tur had the same ligand as was predicted by FunFOLD3, DGL.

D-glutamic acid, as glutamic acid is the most common excitatory neurotransmitter in the neurotransmitter and the role of the ligand, potentially fits in with the post-translational modifications of the protein target as a signal peptides. No information in literature was available for the role of DGL and mycobacterium tuberculosis L,D-transpeptidase.

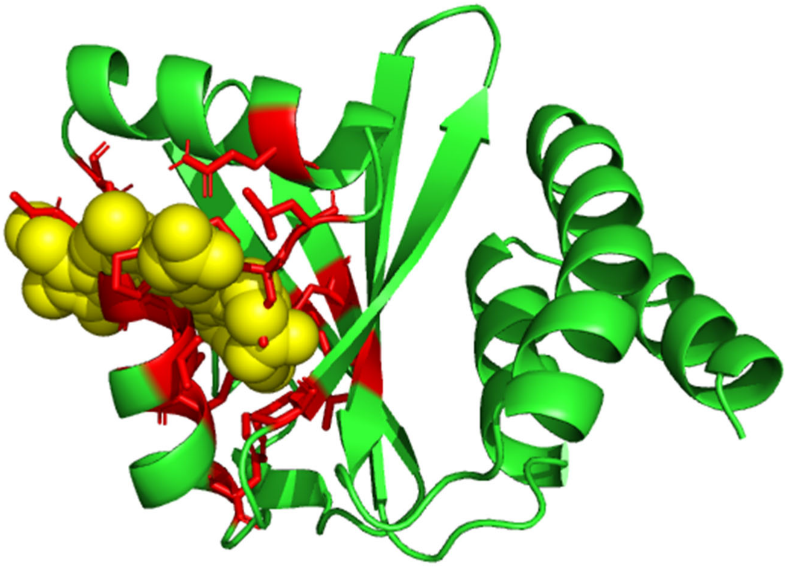


Figure S.56. FunFOLD3 ligand-binding site predictions for CASP13 target T1001
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted biliverdine IX alpha (BLA) ligand shown as sphere and coloured *yellow***.** No observed structure was released by CASP organisers

The twenty-fourth predicted CASP13 target is Q6MIM9, a sensor histidine kinase from gene Bd3125.(UniProt Consortium, 2019) There are limited information available on the protein as there is no PDB ID associated with the target. The UniProtKB entry states the protein is predicted and has the lowest annotation score.(UniProt Consortium, 2019) As a result of no PDB ID associated with the target or CASP13 organisers not releasing an observed structure, no comparisons of the predicted target can be made against a known (observed) protein.

In the family and domains section of the UniProtKB entry, a feature of this protein is coiled coil which occurs at position 136-156.(UniProt Consortium, 2019) The total number of residues provided for this target is 140. Therefore, it would be difficult to visualise based on the current target. Coiled coils are built by two or more alpha-helices that wind around each other to form a super coil there can be two, three or four helices is in the bundle and they might either running the same (parallel) or in the opposite (antiparallel) directions.(UniProt Consortium, 2019) The role of sensor histidine kinases (SHKs) constitute the main means by which bacteria gather information about their surroundings and are found in plants and certain other eukaryotes.(Berntsson *et al.*, 2017) SHKs detect external signals, such as chemicals, light or pH changes.(Berntsson *et al.*, 2017) This in turn triggers structural changes and the activity of the histidine kinase output domains is modulated.(Berntsson *et al.*, 2017) The structure of SHKs is well understood and contains a coiled coil linker, which is supported in the UniProtKB entry. Information from a study by Berntsson et al.,(Berntsson *et al.*, 2017) it was proposed that left-handed supercoiling is the structural mechanism by which signals are relayed from the sensor to the effector module of the SHK and by which activity is switched from kinase to phosphatase.(Berntsson *et al.*, 2017) Furthermore, this is supported by the function of the target as per UniProtKB which lists kinase activity (GO:0016301) as part of molecular function.(UniProt Consortium, 2019) No GO terms were predicted by FunFOLD3, however in terms of templates signalling proteins were identified (4fof, 4s21 and 5akp – which has the ligand BLA).

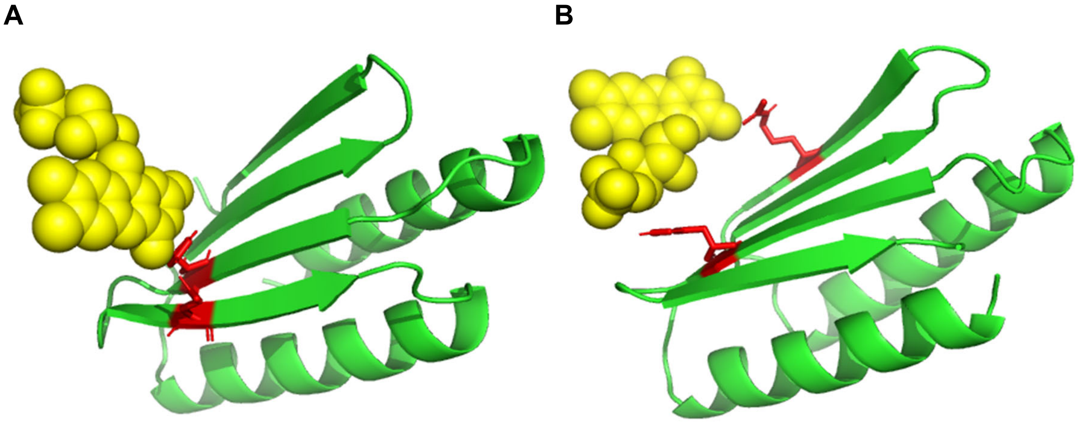


Figure S.57. FunFOLD3 ligand-binding site predictions for CASP13 target T1008 (PDB ID 6msp)
(A) Predicted ligand binding site residues shown as sticks with predictions in red, the predicted flavin monoucleotide (FMN) ligand shown as sphere and coloured yellow **(B)** Predicted ligand binding site residues shown as sticks with predictions in red, the predicted FMN ligand is shown as sphere and coloured yellow. Two different binding sites for the same ligand were identified for this target. No observed structure was released by CASP organisers and no biologically relevant ligands were found in the PDB file

The twenty-sixth predicted CASP13 target is UW_engnr, a de novo designed protein folfit3 and as per the PDB entry is classified as a de novo protein and the organism is a synthetic construct. There is an article associated with the PDB entry.

The authors wanted to investigate how crowd-based creativity could contribute to solving the de novo protein design problem.(Koepnick *et al.*, 2019) The authors incorporated de novo design tools into the protein folding game Foldit.(Koepnick *et al.*, 2019) Foldit is a free online computer game developed to crowdsource problems in protein modelling and provide full control over the three-dimensional structure of a protein model.(Koepnick *et al.*, 2019) Players compete to build a model with the lowest free energy, as calculated by the Rosetta energy function.(Koepnick *et al.*, 2019) Players were repeatedly challenged to design stably folded proteins from scratch and iteratively improved the game based on their results.(Koepnick *et al.*, 2019) Four of the player-designed proteins had high-resolution strutures solved, of which Foldit3, T1008, was one of them and was nominated as a target for the CASP COMMONS Community Outreach program.(Koepnick *et al.*, 2019) As the protein is de novo, there are no information available about the target in literature.

Figure S.58 below, shows the TMalign superposition for the predicted T1008 and the observed structure as per the PDB entry. A TM-score of 0.63629 was achieved demonstrating similar folds.

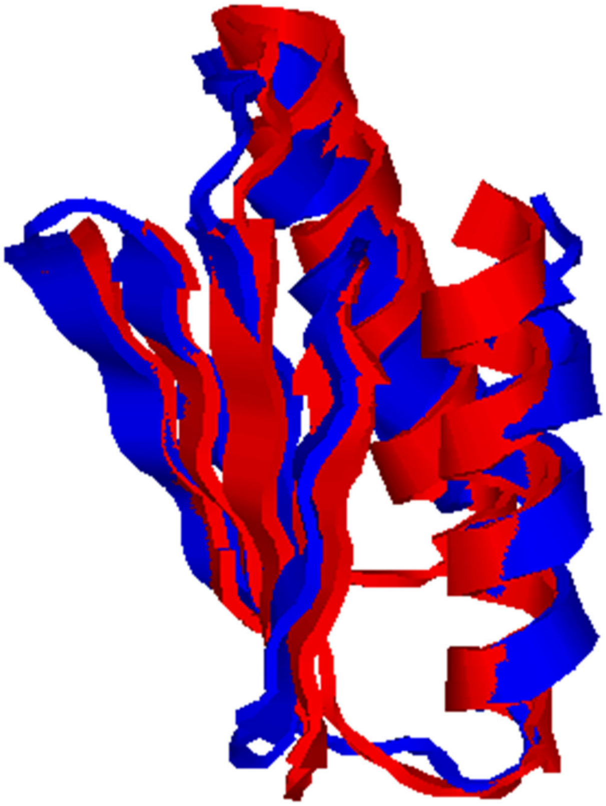


Figure S.58. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted and observed structure for T1008 (PDB ID 6msp)
The structure in blue is the observed structure from 6msp and the predicted structure is in red. A TM-score of 0.63629 was achieved for protein structures. The score was normalised for the observed structure T1003 as it is the reference molecule

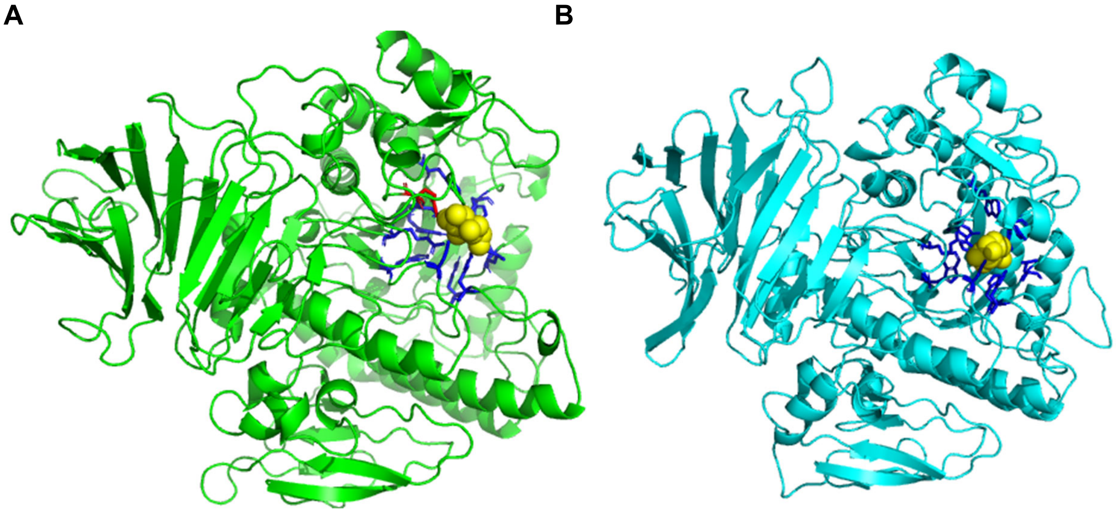


Figure S.59. FunFOLD3 ligand-binding site predictions for CASP13 target T1009 (PDB ID 6dru)
(A) Predicted ligand binding site residues shown as sticks with correct predictions in blue and incorrect predictions in red the predicted alpha-D-glucopyranose (GLC) ligand shown as sphere and coloured yellow**.** An MCC and BDT score of 0.91 and 0.94 was achieved, respectively **(B)** The observed ligand binding site residues shown as sticks and coloured blue, the alpha-D-xylopyranose (XYS) ligand is shown as sphere and coloured yellow.

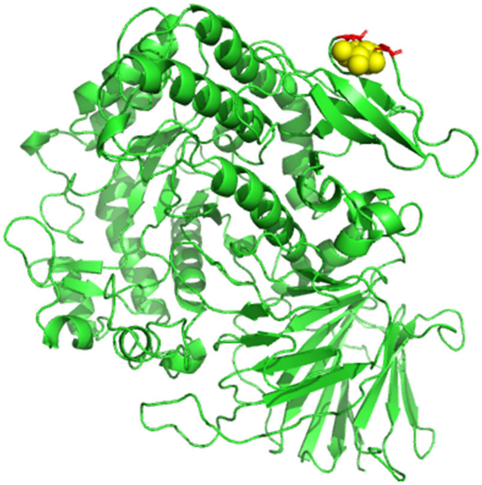


Figure S.60. FunFOLD3 ligand-binding site predictions for CASP13 target T1009 (PDB ID 6dru)
Predicted ligand binding site residues shown as sticks with predictions in red, the predicted GLC ligand shown as sphere and coloured yellow

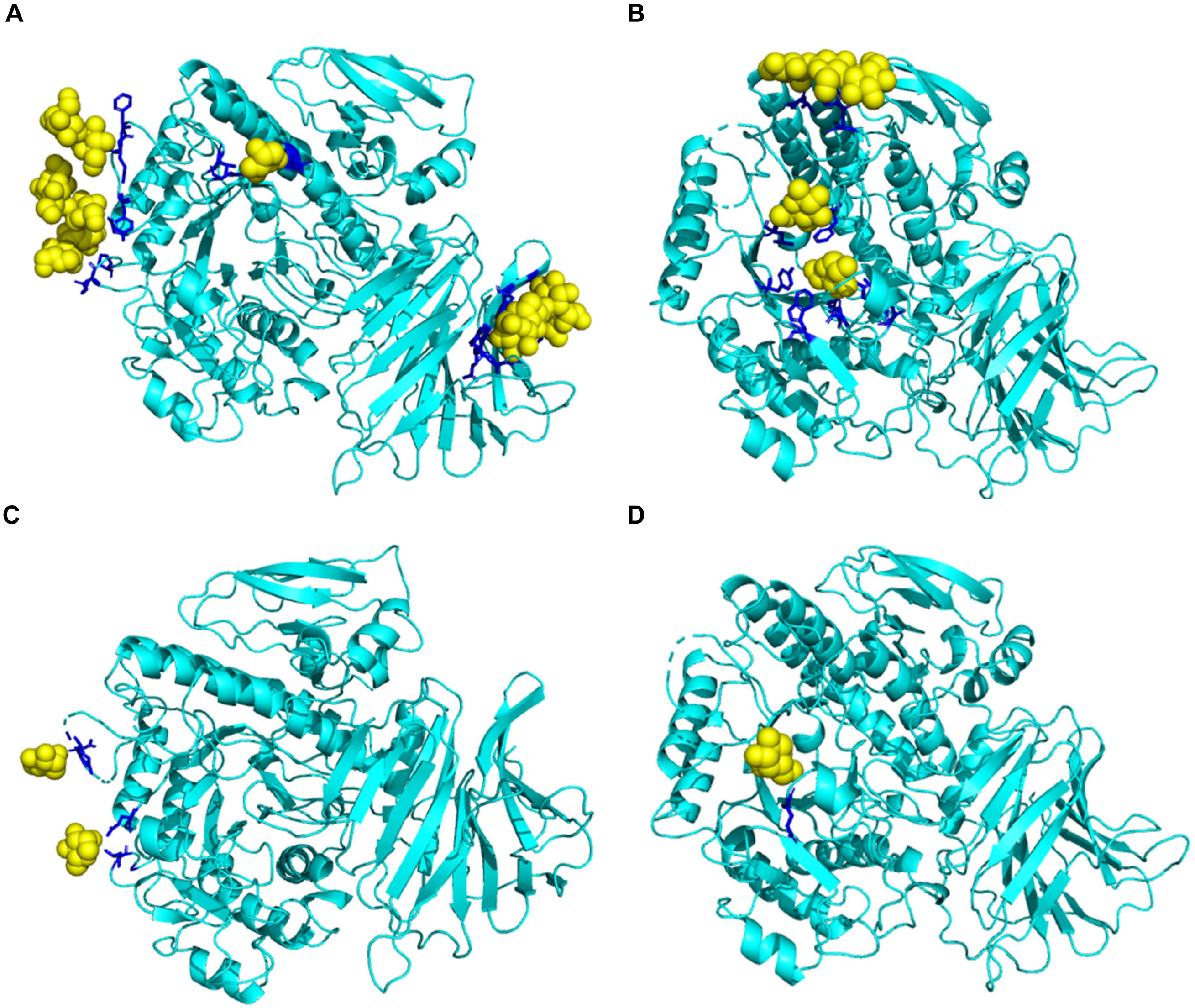


Figure S.61. FunFOLD3 ligand-binding site predictions for CASP13 target T1009 (PDB ID 6dru)
(A) Observed ligand binding site residues shown as sticks and coloured blue the observed ligand alpha-D-mannopyranose (MAN) ligand is shown as sphere and coloured yellow. The MAN ligand was predicted in four different locations on the target. **(B)** Observed ligand binding site residues shown as sticks and coloured blue the observed ligand beta-D-glucopyranose (BGC) ligand is shown as sphere and coloured yellow. The BGC ligand was predicted in three different locations on the target. **(C)** Observed ligand binding site residues shown as sticks and coloured blue the observed ligand beta-D-mannopyranose (BMA) ligand is shown as sphere and coloured yellow. The BMA ligand was predicted in two different locations on the target **(D)** Observed ligand binding site residues shown as sticks and coloured blue the observed ligand beta-D-galactopyranose (GAL) ligand is shown as sphere and coloured yellow.

The twenty-seventh predicted CASP13 target is A2QTU5.1 or xylosidase as per the PDB entry and is classified as a hydrolase. FunFOLD3 predicted the GLC ligand at two different locations. In comparison, the biologically relevant ligands identified in the observed structure as released by the CASP13 organisers were XYS, MAN, BGC, BMA and GAL. The MAN, BGC and BMA ligand was identified in multiple locations within the protein structure. The PDB entry identifies NAG, XYS and GOL as ligands. Despite FunFOLD3 not predicting the same ligands as in the observed structure, the predicted GLC ligand has similar residues to the XYS ligand. Therefore, a comparison was made between these two ligands and an MCC and BDT score was calculated and was 0.91 and 0.94,respectively. Due to the very good MCC and BDT score it is fair to compare them as the same ligand and Figure S.61 below compares the two ligand structures to one another and there is clear homology between the two ligands.

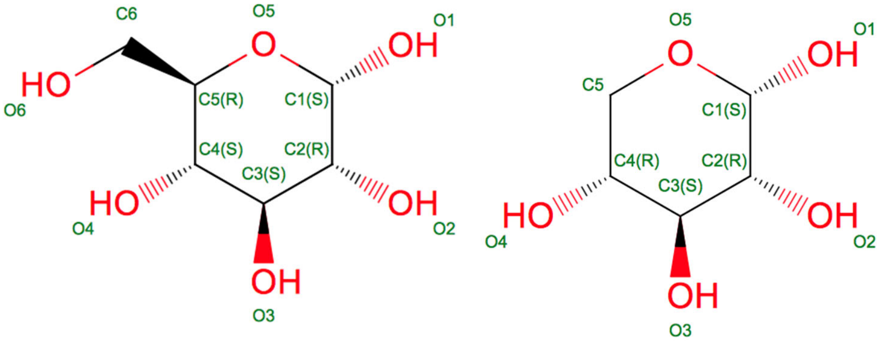


Figure S.62. Comparison of GLC and XYS ligand
The difference between GLC and XYS ligand. The GLC ligand was predicted by FunFOLD3 and the XYS ligand was a biologically relevant ligand identified in the observed structure

In literature information available on the protein, a key molecular determinant of substrate specificity corresponding to Tyr286, this residue was also predicted in the CASP13 target and this is despite the poor sequence conservation at this position among GH31 family enzymes.(Cao *et al.*, 2020) Almost all the structurally available GH31 α-xylosidases possess a bulky aromatic residue at the spatially equivalent position to Tyr286(Cao *et al.*, 2020) expect for the E.Coli α-xylosidases with Cys307 (PDB ID 2f2h)(Cao *et al.*, 2020), of which this template was used by FunFOLD3 for prediction of the ligand-binding sites.

Figure S.63 below, shows the TMalign superposition for the predicted T1009 and the observed structure as released by CASP organisers. A TM-score of 0.86740 was achieved demonstrating the higher end of the same fold.

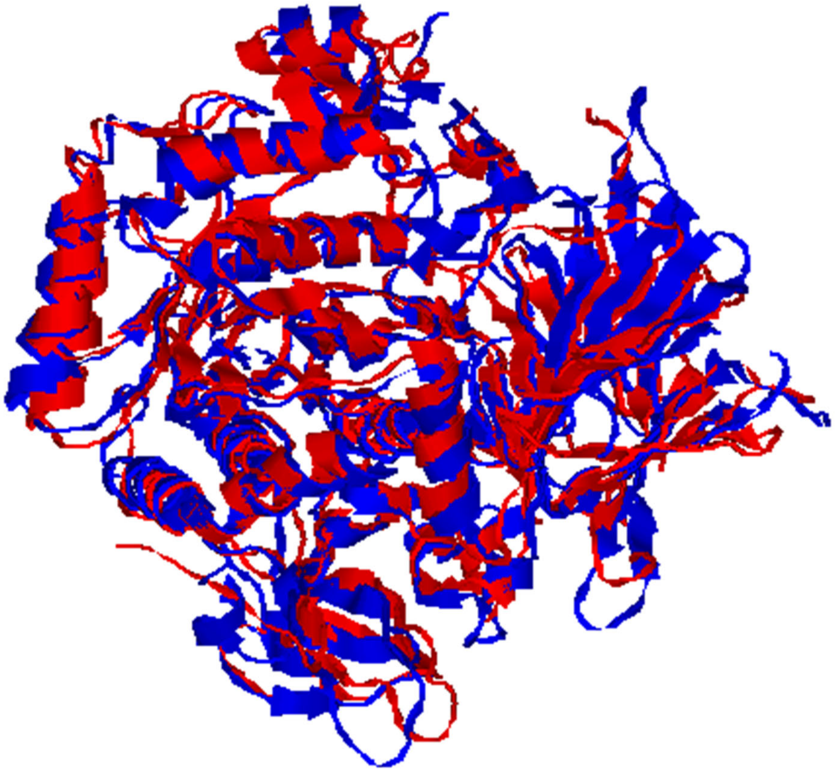


Figure S.63. Comparison of TMalign structures for predicted and observed structure for T1009 (PDB ID 6dru)
The structure in blue is the observed structure as released by the CASP13 organisers and the predicted structure is in red. A TM-score of 0.86740 was achieved for protein structures. The score was normalised for the observed structure T1009 as it is the reference molecule

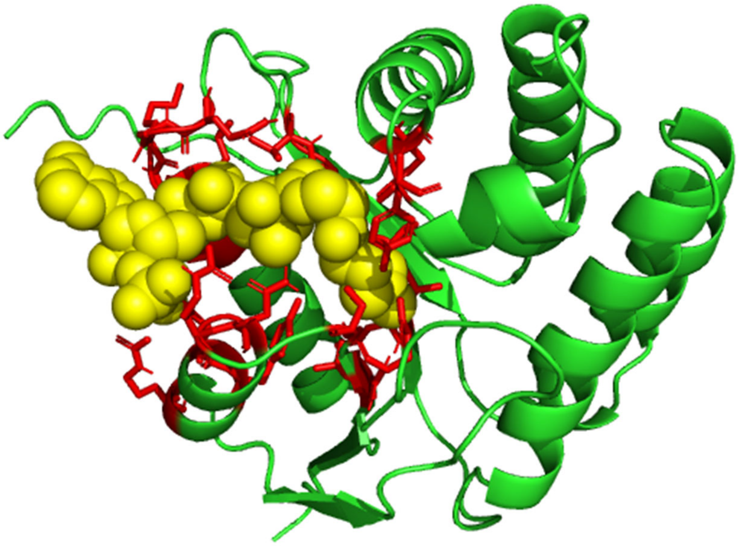


Figure S.64. FunFOLD3 ligand-binding site predictions for CASP13 target T1012
Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted acetyl coenzyme-A (ACO) ligand shown as sphere and coloured *yellow*. Observed structure cancelled by CASP organisers

The twenty-eighth predicted CASP13 target is puromycin N-acetyltransferase and the structure was cancelled by CASP13 organisers. Additionally, no PDB ID is associated with this target. The status of the protein is predicted and has an annotation score of two out of five. The function of the protein is detoxification of puromycin and the UniProtKB entry identifies N-acetyltransferase activity (GO:008080) as the molecular function and response to antibiotic (GO:0046677) as a biological process. FunFOLD3, also predicted N-acetyltransferase activity and the full list of GO terms predicted by FunFOLD3 are depicted in Table S.18 below.

Table S.18. Predicted GO terms for CASP13 target T1012 as predicted by FunFOLD3
The GO terms for CASP13 target T1012 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0004059	Molecular function	aralkylamine N-acetyltransferase activity
GO:0005515	Molecular function	protein binding
GO:0008080	Molecular function	N-acetyltransferase activity
GO:0016740	Molecular function	transferase activity
GO:0016746	Molecular Function	transferase activity, transferring acyl groups
GO:0008483	Molecular Function	Transaminase activity
GO:0006474	Biological process	N-terminal protein amino acid acetylation
GO:0007623	Biological process	circadian rhythm
GO:0030187	Biological process	melatonin biosynthetic process
GO:0048511	Biological process	rhythmic process
GO:0071320	Biological process	cellular response to cAMP

GO:0005737	Cellular Component	cytoplasm
GO:0048471	Cellular Component	perinuclear region of cytoplasm

There are no recent literature publications on the role of puromycin acetyltransferase. Puromycin is an aminonucleoside antibiotic with structural similarity to aminoacyl tRNA.(Cary *et al.*, 2014) This allows the drug to bind to the ribosomal A site and incorporate into nascent polypeptides, causing chain termination, ribosomal subunit dissociation and widespread translational arrest at high concentrations.(Cary *et al.*, 2014) Puromycin N-acetyltransferase is a bacterial enzyme which inactivates puromycin by acetylating the amino position of its tyrosinyl moiety.(Lahoz, de Haro and Esponda, 1991) This could be facilitated by the predicted ACO ligand. Acetyl-CoA represents a key node in metabolism due to its intersection with many other metabolic pathways and transformations.(Shi and Tu, 2015)

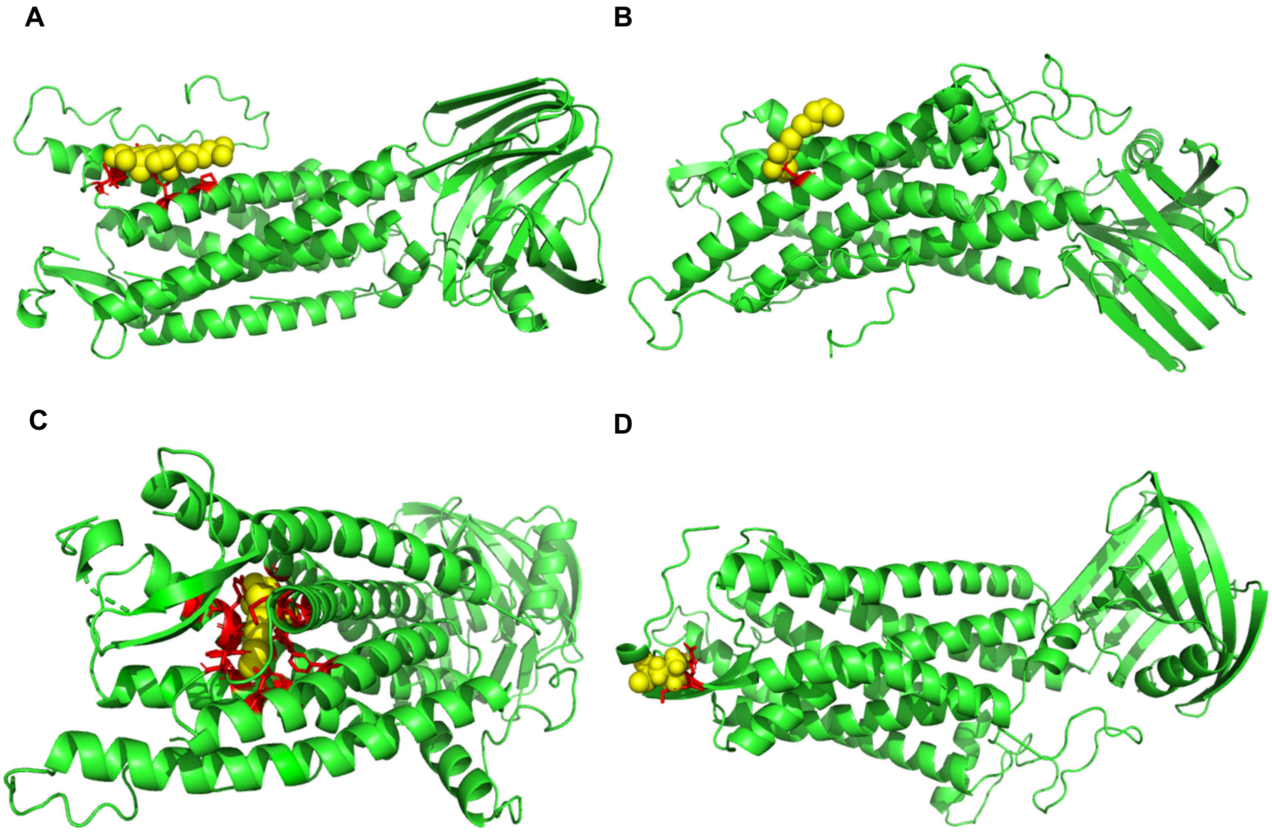


Figure S.65. FunFOLD3 ligand-binding site predictions for CASP13 target T1013
(A) Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted cholesterol (CLR) ligand shown as sphere and coloured *yellow* **(B)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted [Z-octadec-9-enyl] (2R)-2,3-bis(oxidanyl)propanoate (MPG) ligand is shown as sphere and coloured *yellow*. **(C)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted retinal (RET) ligand shown as sphere and coloured *yellow* **(D)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted leucine (LEU) ligand is shown as sphere and coloured *yellow*. No observed structure released by CASP organisers

The twenty-ninth predicted CASP13 target is UNK2, as no observed structure was released by CASP13 organisers and a search on UniProtKB isn't conclusive, information from the FunFOLD3 prediction will be presented. Table S.19 below shows the GO terms predicted by FunFOLD3. The GO terms are quite varied across the three groups, however there appears to be some consistency with the classification of the protein potentially a hydrolase, due to two different GO term predictions around hydrolase activity. Additionally, the function could potentially involve light as there are a number of GO terms related to this, in particular the cellular component where this could occur is the membrane. Additional support for this comes from the prediction of a retinal ligand.

Table S.19. Predicted GO terms for CASP13 target T1013 as predicted by FunFOLD3
The GO terms for CASP13 target T1013 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0003796	Molecular function	lysozyme activity
GO:0003824	Molecular function	catalytic activity
GO:0004871	Molecular function	obsolete signal transducer activity
GO:0004930	Molecular function	G protein-coupled receptor activity
GO:0004969	Molecular Function	histamine receptor activity
GO:0016787	Molecular Function	hydrolase activity
GO:0016798	Molecular Function	hydrolase activity
GO:0005515	Molecular Function	protein binding
GO:0009881	Molecular Function	photoreceptor activity
GO:0046872	Molecular Function	metal ion binding
GO:0004995	Molecular Function	tachykinin receptor activity
GO:0006954	Biological Process	inflammatory response
GO:0007165	Biological Process	signal transduction
GO:0007186	Biological Process	G protein-coupled receptor signaling pathway
GO: 0007200	Biological Process	phospholipase C-activating G protein-coupled receptor signaling
GO:0007268	Biological Process	chemical synaptic transmission
GO:0008152	Biological Process	metabolic process
GO:0009253	Biological Process	peptidoglycan catabolic process
GO:0009629	Biological Process	response to gravity
GO:0010894	Biological Process	negative regulation of steroid biosynthetic process
GO:0016998	Biological Process	cell wall macromolecule catabolic process
GO:0019835	Biological Process	cytolysis
GO:0032962	Biological Process	positive regulation of inositol triphosphate biosynthetic process
GO:0042742	Biological Process	defense response to bacterium
GO:0045429	Biological Process	positive regulation of nitric oxide biosynthetic process
GO:0045907	Biological Process	Positive regulation of vasoconstriction
GO:0048016	Biological Process	inositol phosphate-mediated signaling
GO:0071420	Biological Process	cellular response to histamine
GO:0006468	Biological Process	protein phosphorylation
GO:0007601	Biological Process	visual perception
GO:0007602	Biological Process	phototransduction
GO:0009416	Biological Process	response to light stimulus
GO:0009583	Biological Process	detection of light stimulus
GO:0018298	Biological Process	protein-chromophore linkage
GO:0050896	Biological Process	response to stimulus
GO:0050953	Biological Process	sensory perception of light stimulus
GO:0060041	Biological Process	retina development in camera-type eye
GO:0071482	Biological Process	cellular response to light stimulus
GO:0005634	Cellular Component	nucleus
GO:0005737	Cellular Component	cytoplasm
GO:0005886	Cellular Component	plasma membrane
GO:0005887	Cellular Component	integral component of plasma membrane
GO:0016020	Cellular Component	membrane

GO:0016021	Cellular Component	integral component of membrane
GO:0005730	Cellular Component	nucleolus
GO:0001750	Cellular Component	photoreceptor outer segment
GO:0001917	Cellular Component	photoreceptor inner segment
GO:0005794	Cellular Component	Golgi apparatus
GO:0042622	Cellular Component	photoreceptor outer segment membrane
GO:0060342	Cellular Component	photoreceptor inner segment membrane

In terms of templates, out of the seventeen templates that were associated with the protein target, six of those were signalling protein and five were hydrolase. Thereby, suggesting that the protein target is potentially a signalling protein or a hydrolase.

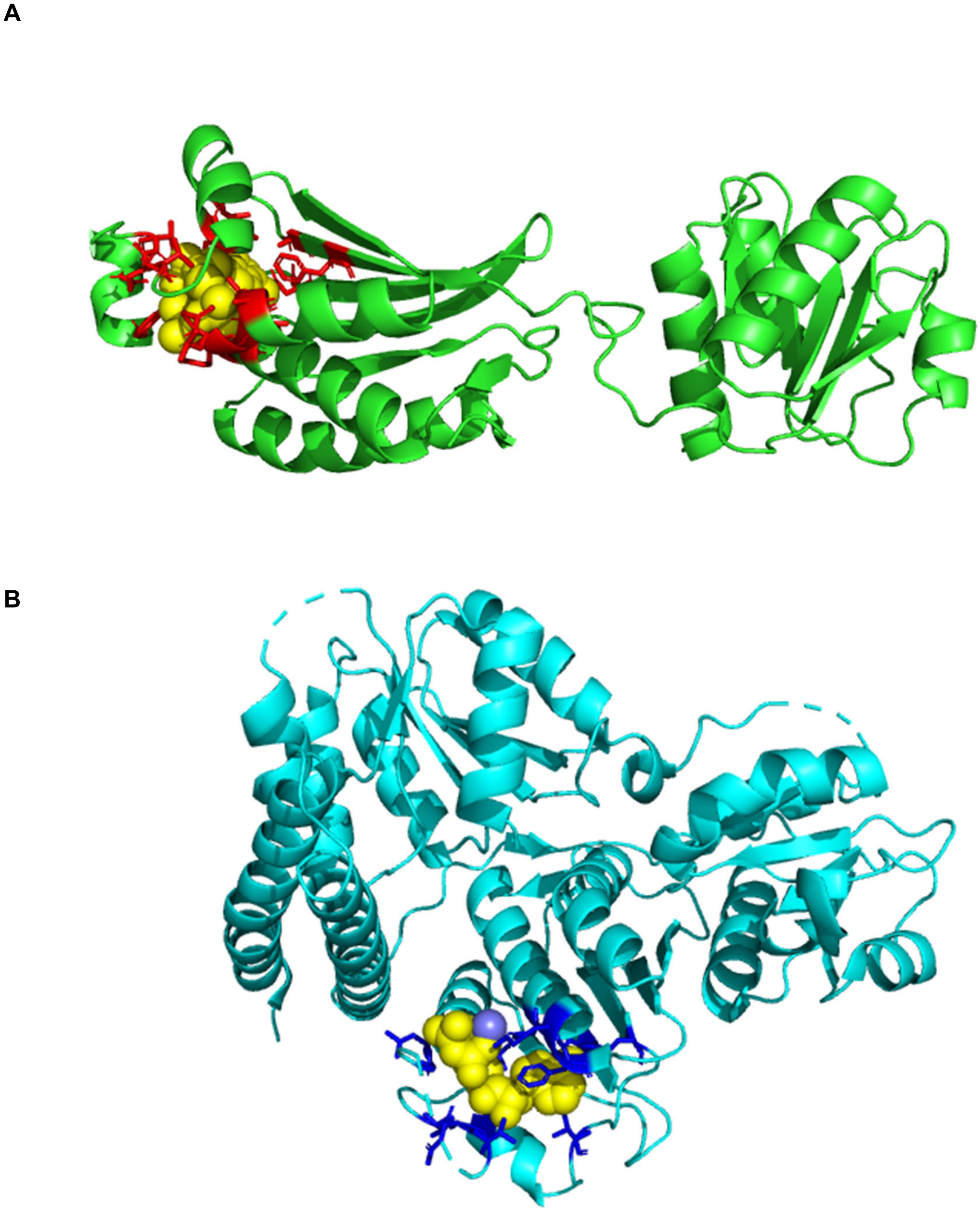


Figure S.66. FunFOLD3 ligand-binding site predictions for CASP13 target T1014 (PDB ID 6qrj)
Predicted ligand binding site residues shown as sticks with incorrect predictions in red, the predicted adenosine-5'-phosphate (ADP) ligand shown as sphere and coloured yellow. MCC and BDT score was -0.05 and 0.05, respectively **(B)** The observed ligand binding site residues shown as sticks for T0953s2 with binding site residues coloured in blue and the correctly predicted ligands MG shown as sphere and coloured blue and the phosphoaminophosphonic acid-adenylate ester (ANP) ligand shown as sphere and coloured yellow. As no observed structure was released by CASP organisers the predictions have been made against the PDB structure

The thirtieth predicted CASP13 target is WP_010918027.1 or ShkA, as per the PDB entry for the target and is classified as a signalling protein. FunFOLD3 predicted ADP as the biologically relevant ligand, whereas the observed ligands are magnesium and ANP and this is also as per the PDB entry. The observed ligands ANP and MG share two ligand-binding site residues; 139 and 142. ShkA, is a noncanonical hybrid histidine kinase lacking any N-terminal input domain.(Dubey *et al.*, 2020) Information in literature, suggest that ShkA is an ATPase, converting ATP to ADP via phosho-enzyme intermediates and in the presence of c-di-GMP, ShkA efficiently catalyses ATP turnover by enabling autophosphorylation and subsequent phosphotransfer and dephosphorylation.(Dubey *et al.*, 2020)

Phosphoaminophosphonic acid-adenylate ester is also known as AMPPNP and the structure of ShkA has been resolved in the presence of AMPPNP and magnesium, as shown by the PDB structure.(Dubey *et al.*, 2020) AMPPNP, is a nonhydrolyzable anlog of ATP in which the bridging O atom between the two terminal phosphate groups is substituted by the imido function.(Dauter and Dauter, 2011)

Figure S.66 below, shows the TMalign superposition for the predicted T1014 and PDB structure for 6qrj. A TM-score of 0.31010 was achieved demonstrating almost random structural homology.

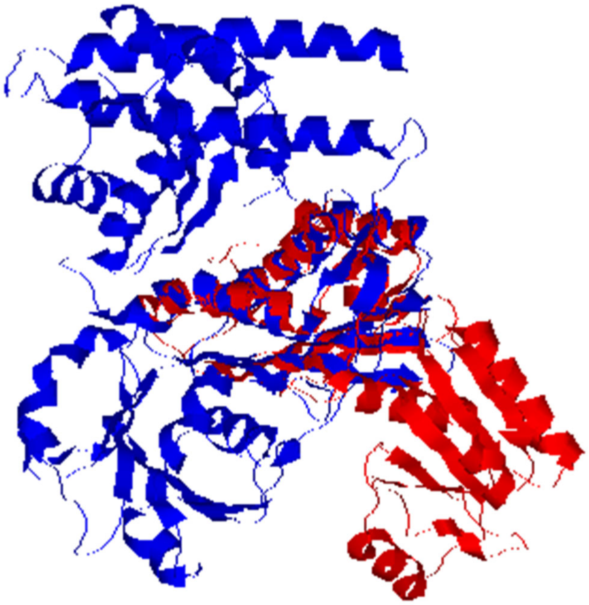


Figure S.67. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted and observed structure for T1014 (PDB ID 6qrj)
The structure in blue is the observed structure from the PDB entry and the predicted structure is in red. A TM-score of 0.31010 was achieved for protein structures. The score was normalised for the observed structure 6qrj as it is the reference molecule

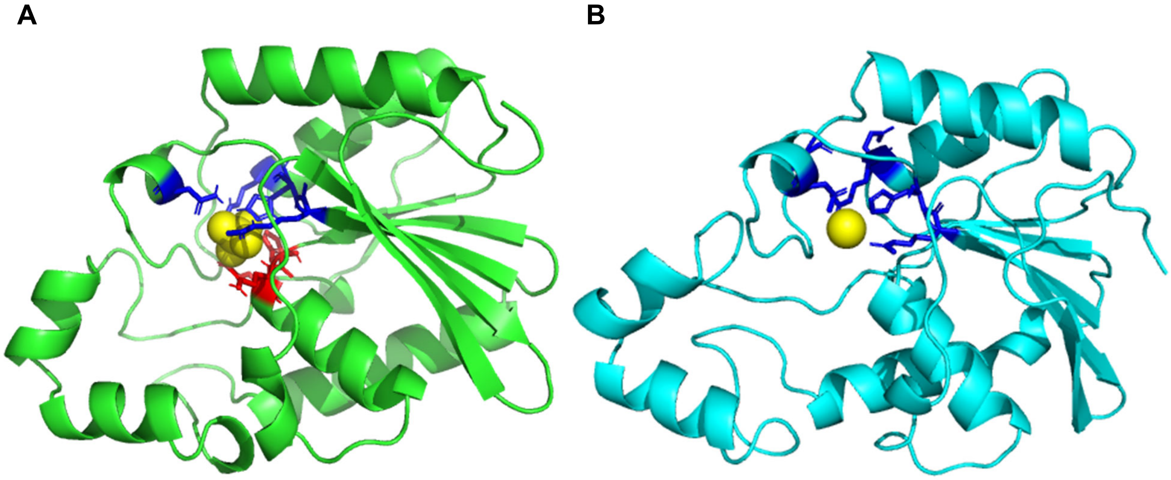


Figure S.68. FunFOLD3 ligand-binding site predictions for CASP13 target T1016 (PDB ID 6e4b)
(A) Predicted ligand binding site residues shown as sticks with correct predictions in *blue* and over predictions in *red* the predicted phosphate (PO4) ligand shown as sphere and coloured *yellow.* An MCC and BDT score of 0.556 and 0.646, respectively was achieved **(B)** The observed ligand binding site residues shown as sticks and coloured *blue*, the CL ligand is shown as sphere and coloured *yellow*.

The thirty-first predicted CASP13 target is IDP96117 or alpha-ribazole-5'-P-phosphatase as per the PDB entry for the target and is classified as hydrolase. FunFOLD3 predicted phosphate as the ligand, in comparison chloride was the observed ligand and this was also supported the PDB entry for the protein. Despite the difference between the two ligands an MCC and BDT score of 0.556 and 0.646, respectively.

Information in UniProtKB about the function of alpha-ribazole-5'-P phosphatase, identifies the molecular function as alpha-ribazole phosphatase activity (GO:0043755) and the biological process as cobalamin biosynthetic process (GO:0009236).(UniProt Consortium, 2019) Further entries regarding alpha-ribazole phosphatase relate the protein to adenosylcobalamin. For one particular entry, the function is catalysing the conversion of adenosylcobalamin 5'-phosphate to adenosylcobalamin (vitamin B12); involved in the assembly of the nucleotide loop of cobalamin.(UniProt Consortium, 2019) Additionally, two active sites at position eight and 81 are denoted with the description of tele-phosphohistidine intermediate and proton donor/acceptor, respectively(UniProt Consortium, 2019) (UniProt KB – P39701) both these residues were predicted by FunFOLD3 and are also part of the observed structure ligand-binding site predictions. In addition to the molecular function mentioned above, intramolecular transferase activity, phosphotransferase (GO:0016868).(UniProt Consortium, 2019) In terms of catalytic activity adenosylcob(III)alamin 5'-phosphate and water results in adenosylcob(III)alamin and phosphate.(UniProt Consortium, 2019)

Figure S.69 below, shows the TMalign superposition for the predicted T1016 and the observed structure. A TM-score of 0.89374 was achieved demonstrating very good structural homology.

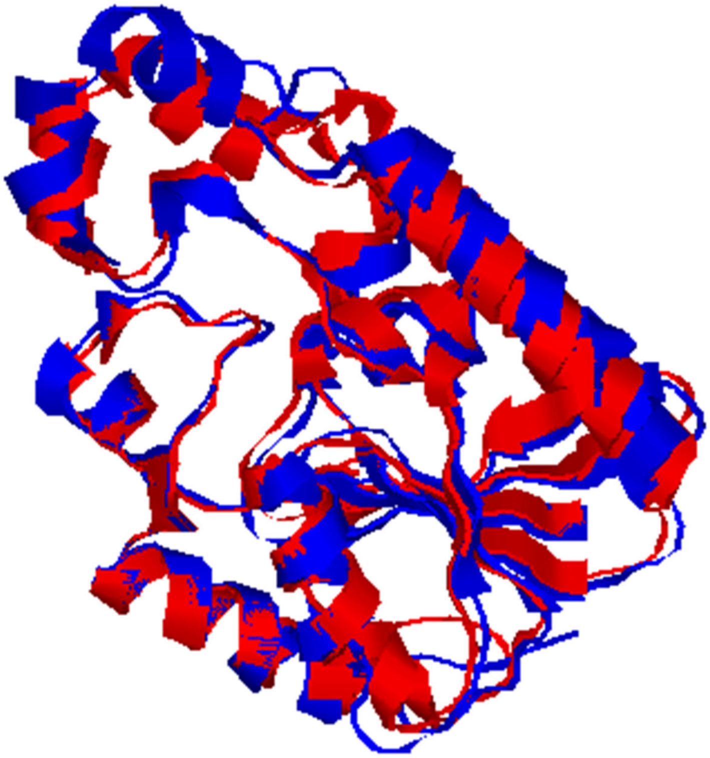


Figure S.69. Comparison of TMalign(Zhang and Skolnick, 2005) structures for predicted and observed structure for T1016 (PDB ID 634b)
The structure in blue is the observed structure from the PDB entry and the predicted structure is in red. A TM-score of 0.89374 was achieved for protein structures. The score was normalised for the observed structure as it is the reference molecule

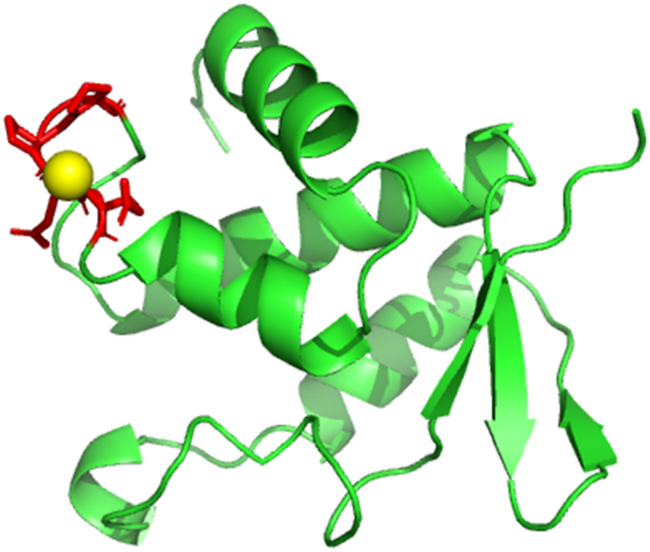


Figure S.70. FunFOLD3 ligand-binding site predictions for CASP13 target T1017s1
Predicted ligand binding site residues shown as sticks with predictions in *red* the predicted ZN ligand shown as sphere and coloured *yellow***.** No observed structure released by CASP organisers

The thirty-second predicted CASP13 target is 201_INDD4, there is no information available on the protein target on UniProtKB and additionally, there is no PDB ID associated with the target. Therefore, no comparisons can be made to an observed structure. Table S.20 below shows the GO terms predicted by FunFOLD3 which can be used to provide insight into the role and function of the protein target.

Table S.20. Predicted GO terms for CASP13 target T1017s1 as predicted by FunFOLD3
The GO terms for CASP13 target T1017s1 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function
GO:0008270	Molecular function	zinc ion binding
GO:0008892	Molecular function	guanine deaminase activity
GO:0016787	Molecular function	hydrolase activity
GO:0046872	Molecular function	metal ion binding
GO:0006147	Biological process	guanine catabolic process
GO:0006195	Biological process	purine nucleotide catabolic process
GO:0006139	Biological process	nucleobase-containing compound metabolic process
GO:0006144	Biological process	purine nucleobase metabolic process
GO:0007399	Biological process	nervous system development
GO:0044281	Biological process	small molecule metabolic process
GO:0055086	Biological process	nucleobase-containing small molecule metabolic process
GO:0005829	Cellular Component	cytosol

GO:0005622	Cellular Component	intracellular anatomical structure

In terms of modelled templates the following were identified 2heo (potassium channel blocker and classified as immune system/DNA), 2v39 (N-wasp WH2 domain), 2xvc (ESCRT-III and classified as cell cycle), 4cc9 (SAMHDI and classified as protein binding) and 4ooj (legionella pneumophilia and classified as unknown function). The templates don't show a consensus among each other so trying to infer the role and function of the target is difficult.

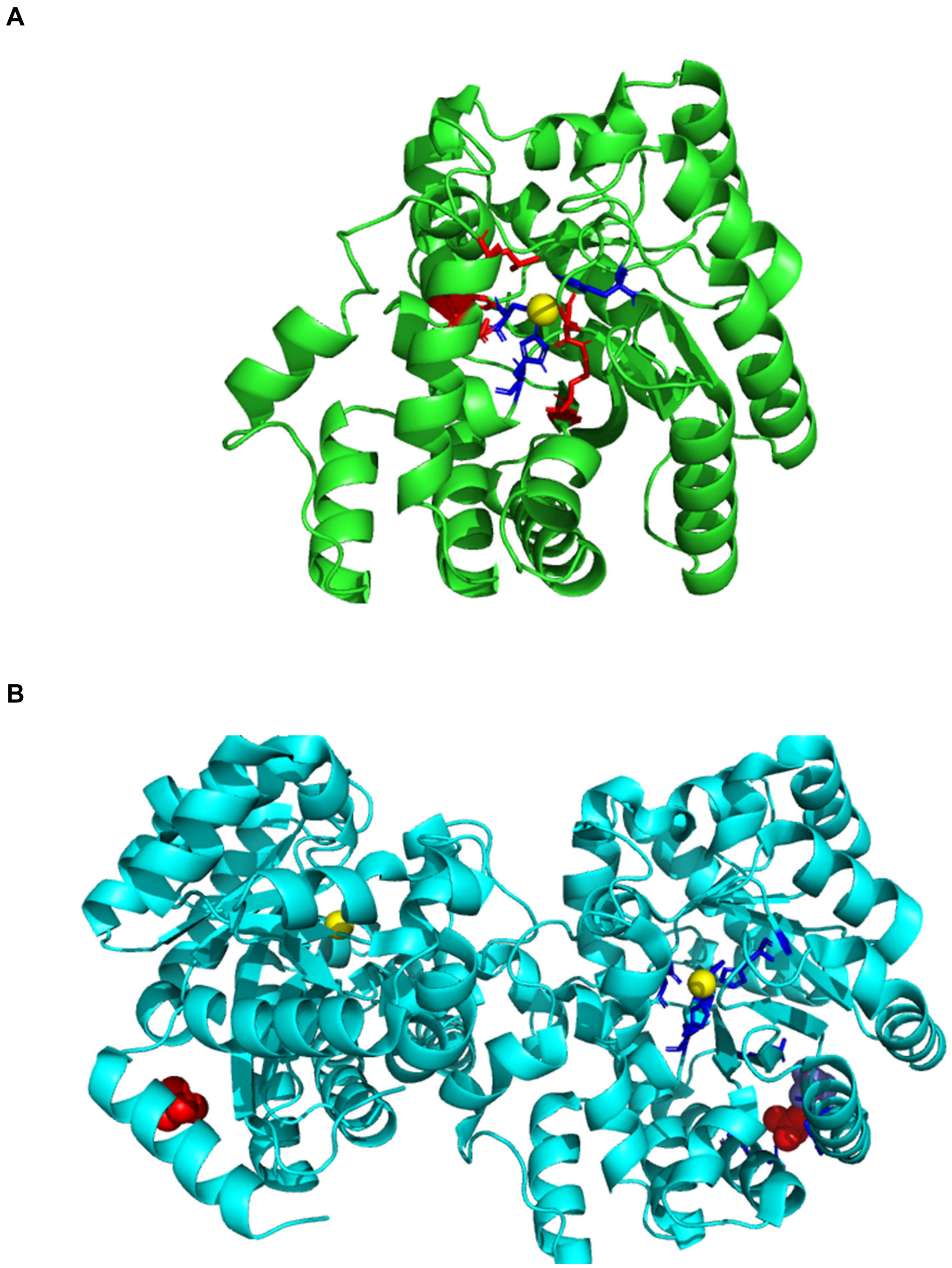


Figure S.71. FunFOLD3 ligand-binding site predictions for CASP13 target T1018 (PDB ID 6n91)
(A) Predicted ligand binding site residues shown as sticks with correct predictions in blue and incorrect predictions in red the predicted zinc (ZN) ligand shown as sphere and coloured yellow. An MCC and BDT score of 0.522 and 0.48, respectively was achieved **(B)** The observed ligand binding site residues shown as sticks and coloured blue, the ZN ligand is shown as sphere and coloured yellow, the phosphate (PO4) ligand is shown as sphere and coloured red and the sulphate (SO4) ligand is shown as sphere and coloured blue. The gull dimer structure has been shown for demonstrative purposes

The thirty-third predicted CASP13 targets is IDP04388 or adenosine deaminase as per the PDB entry and is classified as a hydrolase. FunFOLD3, correctly predicted one of the biologically relevant ligands, ZN. There were three biologically relevant ligands identified in the PDB file obtained from PDB, as the CAP13 organisers did not realise an observed structure for the target. The PDB entry lists nine molecules, under the ligands section; 2'-deoxycorformycin (DCF), 3-cyclohexyl-1-propoylsulpfonic acid (CXS), sulphate ion (SO4), phosphate ion (PO4), glyercol (GOL), 1,2-ethanediol (EDO), formic acid (FMT), sodium ion (NA) and ZN. In term of biological relevance; ZN, SO4 and PO4 were predicted. Interestingly, the ligands were predicted in only the A chain of the dimer of the observed protein with the ZN and SO4 ligand being present in both chains A and B but biological relevance being confined to one chain as demonstrated in Figure S.70A.

Adenosine deaminase (ADA) is a key enzyme in purine metabolism and crucial for normal immune competence and contains a tightly bound zinc, which is required for activity.(Niu *et al.*, 2010) Ada catalyses the irreversible deamination of adenosine or 2'-deoxyadenosine to inosine or 2'-deoxyinosine and ammonia.(Niu *et al.*, 2010) Removing zinc or mutating amino acids involved in metal co-ordination leads to loss of the enzyme activity, confirming the role of Zinc in the catalytic function of ADA.(Niu *et al.*, 2010) The role of zinc is therefore clearly understood in literature. As a result of both the predicted and observed structure containing zinc the MCC and BDT scores were calculated based on just the zinc ligand.

Figure S.72 below, shows the TMalign structures for the predicted T1018 and the observed structure as obtained from PDB. A TM-score of 0.92543 was achieved demonstrating very good structural homology.

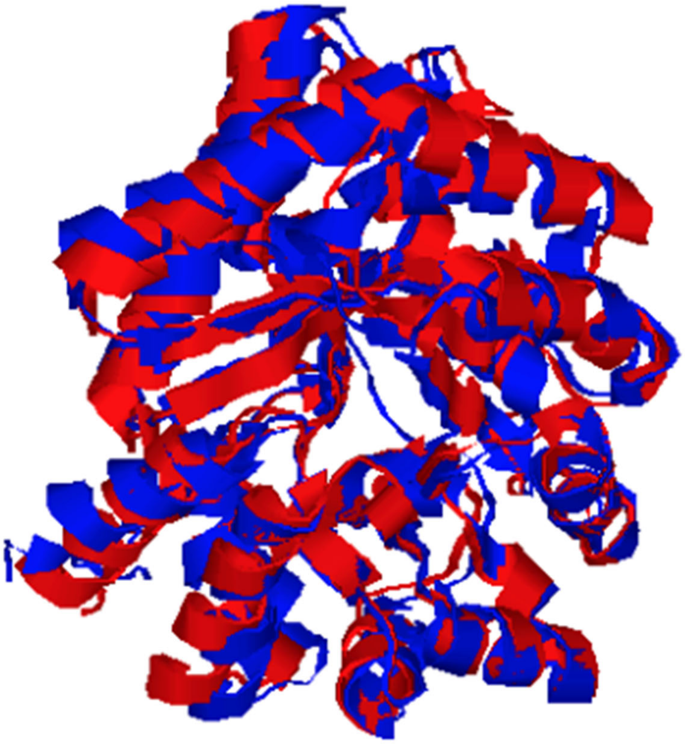


Figure S.72. Comparison of TMalign(Zhang and Skolnick, 2005) **structures for predicted and observed structure for T1018 (PDB ID 6n91)**
The structure in blue is the observed structure from the PDB entry and the predicted structure is in red. A TM-score of 0.92543 was achieved for protein structures. The score was normalised for the observed structure as it is the reference molecule

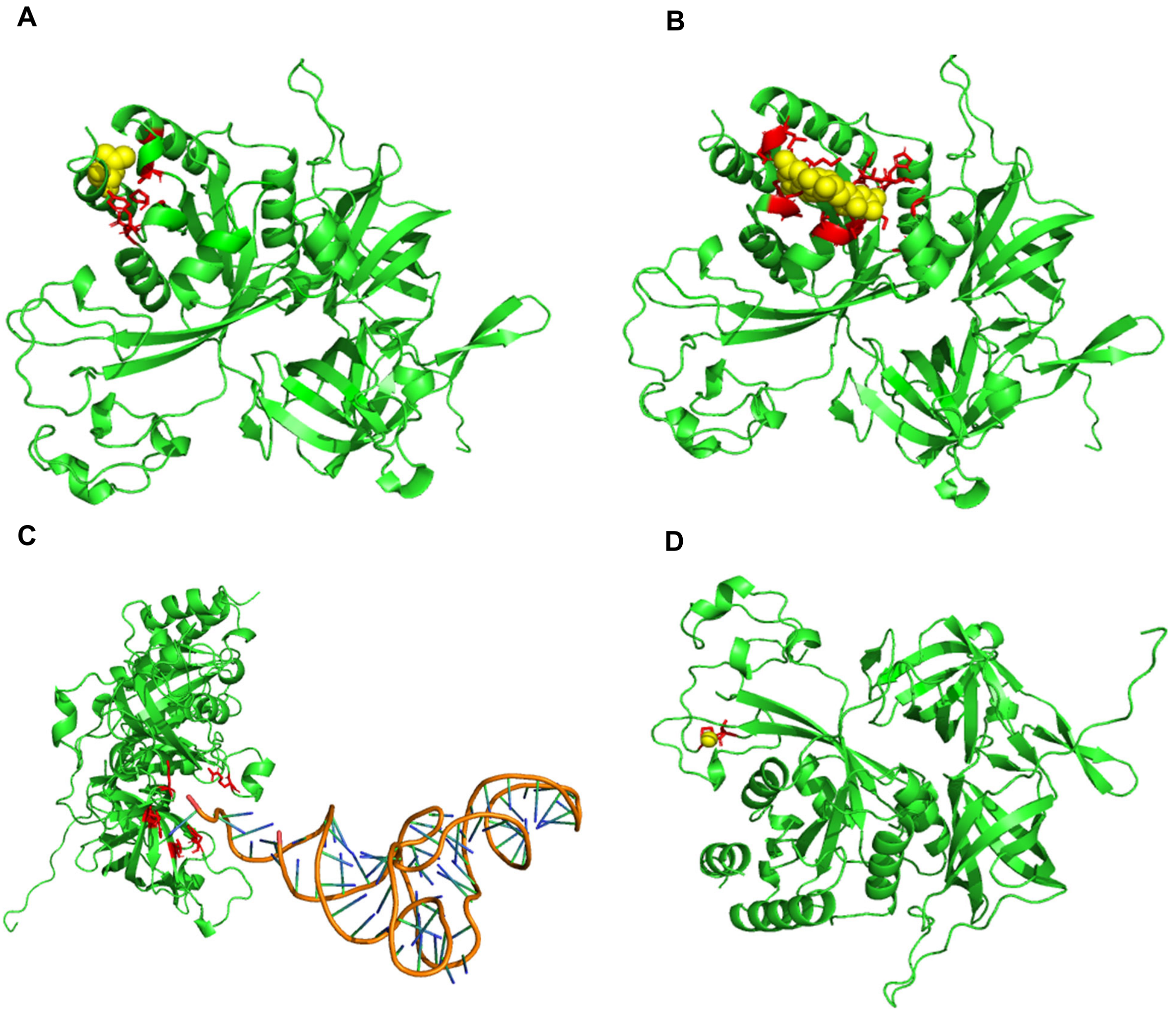


Figure S.73. FunFOLD3 ligand-binding site predictions for CASP13 target T1023s3
(**A**) Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted leucine (LEU) ligand shown as sphere and coloured *yellow* **(B)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted guanosine-5'-diphosphate (GDP) ligand is shown as sphere and coloured *yellow*. **(C)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted RC ligand shown as sphere and coloured *yellow* **(D)** Predicted ligand binding site residues shown as sticks with predictions in *red*, the predicted ZN ligand is shown as sphere and coloured *yellow*. Structure cancelled by CASP organisers

The thirty-fourth predicted CASP13 target and the final predicted target for CASP13 is eIF2 or eukaryotic translation initiation factor 2A, as per the UniProtKB entry.(UniProt Consortium, 2019) The protein has the highest annotation score with five out of five and has experimental evidence at protein level indicating that there is clear experimental evidence for the existence of the protein.

In the functions section of the entry, the protein functions in the early steps of protein synthesis of a small number of specific mRNAs and acts by directing the binding of methionyl-tRNAi to 40S ribosomal subunits.(Zoll *et al.*, 2002)

Additionally, the entry identifies the GO terms related to molecular function and biological process. Table S.21 below shows the GO terms predicted by FunFOLD3

Table S.21. Predicted GO terms for CASP13 target T1023 as predicted by FunFOLD3
The GO terms for CASP13 target T1017s1 and their associated term domains and function are shown below. Molecular function coloured green, biological process coloured red and cellular component is coloured purple

GO term	GO term domain	Function	GO term as per UniProt
GO:0000166	Molecular function	nucleotide binding	
GO:0003746	Molecular function	translation elongation factor activity	
GO:0003924	Molecular function	GTPase activity	
GO:0005525	Molecular function	GTP binding	
GO:0003743	Molecular function	translation initiation factor activity	✓
GO:0003723	Molecular function	RNA binding	GO:0000049 (GO:0003723 ancestor term)
GO:0006412	Biological process	translation	
GO:0006414	Biological process	translational elongation	
GO:0006413	Biological process	translational initiation	
GO:0001514	Biological process	selenocysteine incorporation	
GO:0005622	Cellular Component	intracellular anatomical structure	
GO:0005737	Cellular Component	cytoplasm	

In comparison, the following molecular function GO terms are associated with the entry as per UniProtKB entry(UniProt Consortium, 2019) cadherin binding (GO:0045296) and ribosome binding (GO:0043022). For biological process the following are associated positive regulation of signal transduction (GO:0009967), protein phosphorylation (GO:0006468), regulation of translation (GO:0006417), response to amino acid starvation (GO:1990928), ribosome assembly (GO:0042255), SREBP signalling pathway (GO:0032933). FunFOLD3 correctly identified one GO term and predicted a GO term which is an ancestor term of another GO term. The function description as per the UniProtKB entry, in terms of mRNA aligns with the predicted RNA ligand as per FunFOLD3. Literature information about the protein identifies the role of GTP, in one of the pathways, methionylated initiator tRNA is loaded onto the ribsome in a guanosine-triphosphate (GTP)-dependent and mRNA-independent manner.(Kashiwagi, Ito and Yokoyama, 2014) The predicted ligand, GDP is a hydrolysed version of GTP. Despite no observed structure being released and no PDB entry associated with the target, the ligands predicted by FunFOLD3 and the available information in literature provide insight into the role and ultimately function of the protein.

Appendix 3

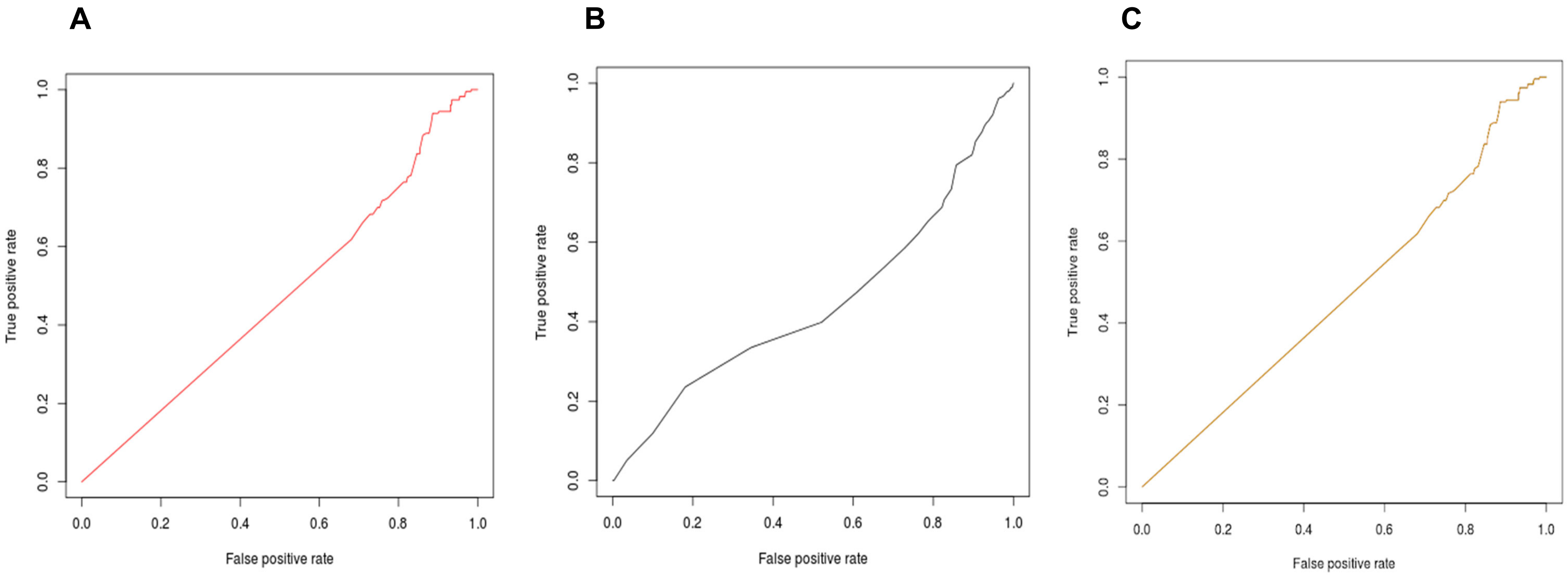


Figure S.74. Receiver operator characteristic curves for the three methods used in the CAFA3 challenge
(A) ROC plot for FunFOLDQ, area under curve for this method is 0.47 (B) ROC plot for HHsearch, area under curve for this method is 0.46 (C) ROC plot for Combined, area under curve for this method is 0.47

1. Target files for CAFA3 GO prediction

Target files moonlighting proteins

```
>M96060000001 IPPK_HUMAN
MEEGKMDENEWGYHGEGNKSLVVAHAQRCVVLRFLKFPPNRKKTSEEIFQHLQNIVDFGK
NVMKEFLGENYVHYGEVVQLPLEFVKQLCLKIQSERPESRCDKDLDTLSGYAMCLPNLTR
LQTYRFAEHRPILCVEIKPKCGFIPFSSDVTHEMKHKVCRYCMHQHLKVATGKWKQISKY
CPLDLYSGNKQRMHFALKSLLQEAQNNLKIFKNGELIYGCKDARSPVADWSELAHHLKPF
FFPSNGLASGPHCTRAVIRELVHVITRVLLSGSDKGRAGTLSPGLGPQGPRVCEASPFSR
SLRCQGKNTPERSGLPKGCLLYKTLQVQMLDLLDIEGLYPLYNRVERYLEEFPEERKTLQ
IDGPYDEAFYQKLLDLSTEDDGTVAFALTKVQQYRVAMTAKDCSIMIALSPCLQDASSDQ
RPVVPSSRSRFAFSVSVLDLDLKPYESIPHQYKLDGKIVNYYSKTVRAKDNAVMSTRFKE
SEDCTLVLHKV
>M100900000002 HXK1_MOUSE
MGWGAPLLSRMLHGPGQAGETSPVPERQSGSENPASEDRRPLEKQCSHHLYTMGQNCQ
RG
QAVDVEPKIRPPLTEEKIDKYLYAMRLSDEILIDILTRFKKEMKNGLSRDYNPTASVKML
PTFVRSIPDGSEKGDFIALDLGGSSFRILRVQVNHEKSQNVSMESEVYDTPENIVHGSGS
QLFDHVAECLGDFMEKRKIKDKKLPVGFTFSFPCRQSKIDEAVLITWTKRFKASGVEGAD
VVKLLNKAIKKRGDYDANIVAVVNDTVGTMMTCGYDDQQCEVGLIIGTGTNACYMEELRH
IDLVEGDEGRMCINTEWGAFGDDGSLEDIRTEFDRELDRGSLNPGKQLFEKMVSGMYMGE
LVRLILVKMAKESLLFEGRITPELLTRGKFTTSDVAAIETDKEGVQNAKEILTRLGVEPS
HDDCVSVQHVCTIVSFRSANLVAATLGAILNRLRDNKGTPRLRTTVGVDGSLYKMHPQYS
RRFHKTLRRLVPDSDVRFLLSESGSGKGAAMVTAVAYRLAEQHRQIEETLSHFRLSKQAL
MEVKKKLRSEMEMGLRKETNSRATVKMLPSYVRSIPDGTEHGDFLALDLGGTNFRVLLVK
IRSGKKRTVEMHNKIYSIPLEIMQGTGDELFDHIVSCISDFLDYMGIKGPRMPLGFTFSF
PCKQTSLDCGILITWTKGFKATDCVGHDVATLLRDAVKRREEFDLDVVAVVNDTVGTMMT
CAYEEPSCEIGLIVGTGSNACYMEEMKNVEMVEGNQGQMCINMEWGAFGDNGCLDDIRT
D
FDKVVDEYSLNSGKQRFEKMISGMYLGEIVRNILIDFTKKGFLFRGQISEPLKTRGIFET
KFLSQIESDRLALLQVRAILQQLGLNSTCDDSILVKTVCGVVSKRAAQLCGAGMAAVVEK
IRENRGLDHLNVTVGVDGTLYKLHPHFSRIMHQTVKELSPKCTVSFLLSEDGSGKGAALI
TAVGVRLRGDPTNA
>M37020000003 HXK1_ARATH
MGKVAVGATVVCTAAVCAVAVLVVRRRMQSSGKWGRVLAILKAFEEDCATPISKLRQVAD
AMTVEMHAGLASDGGSKLKMLISYVDNLPSGDEKGLFYALDLGGTNFRVMRVLLGGKQER
VVKQEFEEVSIPPHLMTGGSDELFNFIAEALAKFVATECEDFHLPEGRQRELGFTFSFPV
KQTSLSSGSLIKWTKGFSIEEAVGQDVVGALNKALERVGLDMRIAALVNDTVGTLAGGRY
YNPDVVAAVILGTGTNAAYVERATAIPKWHGLLPKSGEMVINMEWGNFRSSHLPLTEFDH
TLDFESLNPGEQILEKIISGMYLGEILRRVLLKMAEDAAFFGDTVPSKLRIPFIIRTPHM
SAMHNDTSPDLKIVGSKIKDILEVPTTSLKMRKVVISLCNIIATRGARLSAAGIYGILKK
LGRDTTKDEEVQKSVIAMDGGLFEHYTQFSECMESSLKELLGDEASGSVEVTHSNDGSGI
GAALLAASHSLYLEDS
>M2627230000004 ENO_MYCS5
MSAIKKIHAREVLDSRGNPTVQVEVYTELKGYGSAMVPSGASTGSREALELRDKGSKFES
NWFGGKGVMQAVENVNKLIAPALIGFEVTDQRQVDLAMKALDGTKNKEKLGANAILGVSL
AVARAAANELDLPLYKYLGGFNAHKLPLPMLNVINGGEHASNTLDFQEFMVMPVGAKSFR
EALQMANFVFHNLAKLLKKHGHGVQVGDEGGFAPNFKSHEEALDFLVEAIKLSGYKPATS
GEKAVAIAMDCASSELYKDGKYTFGKLKKAIEEKQPGFENLGKTKLVYTTDELIDYLDHL
VSKYPIVSIEDGLAESDWAGFEKLTKRLGHKLQVVGDDLTVTNTELLAKAIERKAMNSIL
IKVNQIGSLTETFEAIQMAQMANMTAVVSHRSGETEDTTIADVAVAMNTGQIKTGSMSRT
DRIAKYNRLLAIEEELSKASTFPKDVFYNLKK
>M96060000005 PFKAM_HUMAN
MTHEEHHAAKTLGIGKAIAVLTSGGDAQGMNAAVRAVVRVGIFTGARVFFVHEGYQGLVD
```

```
GGDHIKEATWESVSMMLQLGGTVIGSARCKDFREREGRLRAAYNLVKRGITNLCVIGGDG
SLTGADTFRSEWSDLLSDLQKAGKITDEEATKSSYLNIVGLVGSIDNDFCGTDMTIGTDS
ALHRIMEIVDAITTTAQSHQRTFVLEVMGRHCGYLALVTSLSCGADWVFIPECPPDDDWE
EHLCRRLSETRTRGSRLNIIIVAEGAIDKNGKPITSEDIKNLVVKRLGYDTRVTVLGHVQ
RGGTPSAFDRILGSRMGVEAVMALLEGTPDTPACVVSLSGNQAVRLPLMECVQVTKDVTK
AMDEKKFDEALKLRGRSFMNNWEVYKLLAHVRPPVSKSGSHTVAVMNVGAPAAGMNAAV
R
STVRIGLIQGNRVLVVHDGFEGLAKGQIEEAGWSYVGGWTGQGGSKLGTKRTLPKKSFEQ
ISANITKFNIQGLVIIGGFEAYTGGLELMEGRKQFDELCIPFVVIPATVSNNVPGSDFSV
GADTALNTICTTCDRIKQSAAGTKRRVFIIETMGGYCGYLATMAGLAAGADAAYIFEEPF
TIRDLQANVEHLVQKMKTTVKRGLVLRNEKCNENYTTDFIFNLYSEEGKGIFDSRKNVLG
HMQQGGSPTPFDRNFATKMGAKAMNWMSGKIKESYRNGRIFANTPDSGCVLGMRKRALV
F
QPVAELKDQTDFEHRIPKEQWWLKLRPILKILAKYEIDLDTSDHAHLEHITRKRSGEAAV
>M96060000006 F16P1_HUMAN
MADQAPFDTDVNTLTRFVMEEGRKARGTGELTQLLNSLCTAVKAISSAVRKAGIAHLYGI
AGSTNVTGDQVKKLDVLSNDLVMNMLKSSFATCVLVSEEDKHAIIVEPEKRGKYVVCFDP
LDGSSNIDCLVSVGTIFGIYRKKSTDEPSEKDALQPGRNLVAAGYALYGSATMLVLAMDC
GVNCFMLDPAIGEFILVDKDVKIKKKGKIYSLNEGYARDFDPAVTEYIQRKKFPPDNSAP
YGARYVGSMVADVHRTLVYGGIFLYPANKKSPNGKLRLLYECNPMAYVMEKAGGMATTGK
EAVLDVIPTDIHQRAPVILGSPDDVLEFLKVYEKHSAQ
>M96060000007 KPYM_HUMAN
MSKPHSEAGTAFIQTQQLHAAMADTFLEHMCRLDIDSPPITARNTGIICTIGPASRSVET
LKEMIKSGMNVARLNFSHGTHEYHAETIKNVRTATESFASDPILYRPVAVALDTKGPEIR
TGLIKGSGTAEVELKKGATLKITLDNAYMEKCDENILWLDYKNICKVVEVGSKIYVDDGL
ISLQVKQKGADFLVTEVENGGSLGSKKGVNLPGAAVDLPAVSEKDIQDLKFGVEQDVDMV
FASFIRKASDVHEVRKVLGEKGKNIKIISKIENHEGVRRFDEILEASDGIMVARGDLGIE
IPAEKVFLAQKMMIGRCNRAGKPVICATQMLESMIKKPRPTRAEGSDVANAVLDGADCIM
LSGETAKGDYPLEAVRMQHLIAREAEAAIYHLQLFEELRRLAPITSDPTEATAVGAVEAS
FKCCSGAIIVLTKSGRSAHQVARYRPRAPIIAVTRNPQTARQAHLYRGIFPVLCKDPVQE
AWAEDVDLRVNFAMNVGKARGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP
>M96060000008 ECHB_HUMAN
MTILTYPFKNLPTASKWALRFSIRPLSCSSQLRAAPAVQTKTKKTLAKPNIRNVVVVDGV
RTPFLLSGTSYKDLMPHDLARAALTGLLHRTSVPKEVVDYIIFGTVIQEVKTSNVAREAA
LGAGFSDKTPAHTVTMACISANQAMTTGVGLIASGQCDVIVAGGVELMSDVPIRHSRKMR
KLMLDLNKAKSMGQRLSLISKFRFNFLAPELPAVSEFSTSETMGHSADRLAAAFAVSRLE
QDEYALRSHSLAKKAQDEGLLSDVVPFKVPGKDTVTKDNGIRPSSLEQMAKLKPAFIKPY
GTVTAANSSFLTDGASAMLIMAEEKALAMGYKPKAYLRDFMYVSQDPKDQLLLGPTYATP
KVLEKAGLTMNDIDAFEFHEAFSGQILANFKAMDSDWFAENYMGRKTKVGLPPLEKFNNW
GGSLSLGHPFGATGCRLVMAAANRLRKEGGQYGLVAACAAGGQGHAMIVEAYPK
>M99130000009 LDHA_BOVIN
MATLKDQLIQNLLKEEHVPQNKITIVGVGAVGMACAISILMKDLADEVALVDVMEDKLKG
EMMDLQHGSLFLRTPKIVSGKDYNVTANSRLVIITAGARQQEGESRLNLVQRNVNIFKFI
IPNIVKYSPNCKLLVVSNPVDILTYVAWKISGFPKNRVIGSGCNLDSARFRYLMGERLGV
HPLSCHGWILGEHGDSSVPVWSGVNVAGVSLKNLHPELGTDADKEQWKAVHKQVVDSAY
E
VIKLKGYTSWAIGLSVADLAESIMKNLRRVHPISTMIKGLYGIKEDVFLSVPCILGQNGI
SDVVKVTLTHEEEACLKKSADTLWGIQKELQF
>M96060000010 MDHC_HUMAN
MSEPIRVLVTGAAGQIAYSLLYSIGNGSVFGKDQPIILVLLDITPMMGVLDGVLMELQDC
ALPLLKDVIATDKEDVAFKDLDVAILVGSMPRREGMERKDLLKANVKIFKSQGAALDKYA
KKSVKVIVVGNPANTNCLTASKSAPSIPKENFSCLTRLDHNRAKAQIALKLGVTANDVKN
VIIWGNHSSTQYPDVNHAKVKLQGKEVGVYEALKDDSWLKGEFVTTVQQRGAAVIKARKL
```

```
SSAMSAAKAICDHVRDIWFGTPEGEFVSMGVISDGNSYGVPDDLLYSFPVVIKNKTWKFV
EGLPINDFSREKMDLTAKELTEEKESAFEFLSSA
>M96060000011 AUHM_HUMAN
MAAAVAAAPGALGSLHAGGARLVAACSAWLCPGLRLPGSLAGRRAGPAIWAQGWVPAAG
G
PAPKRGYSSEMKTEDELRVRHLEEENRGIVVLGINRAYGKNSLSKNLIKMLSKAVDALKS
DKKVRTIIIRSEVPGIFCAGADLKERAKMSSSEVGPFVSKIRAVINDIANLPVPTIAAID
GLALGGGLELALACDIRVAASSAKMGLVETKLAIIPGGGGTQRLPRAIGMSLAKELIFSA
RVLDGKEAKAVGLISHVLEQNQEGDAAYRKALDLAREFLPQGPVAMRVAKLAINQGMEVD
LVTGLAIEEACYAQTIPTKDRLEGLLAFKEKRPPRYKGE
>M72270000012 Q9VMB4_DROME
MPEQDNYDDELVSSNPNQRNWRGILIALLVIIIVLALIVTSVVLLTPPDEGPRVKGQRIK
LQDIVDGLFVPQHSNGSWIDGEEFLYQDHLGRICLLNAANRSERVLMSNVTFKTLSPFTF
TISADKRYLLLAQNVVKLFRHSYLAQYTLYDIQTSESIKLRHSPHQDEWPYLHYARFTPA
GNALVWVQSYDIYYREEVRSASVHRITHDAVPGVVYNGIPDWLYEEEILHANNAIWMSDN
GQLMLYATFNDTHVQEQHFAWYGTTGPSAGGAAAAAAVGAGGAGTGSPGAGGSNPHAS
LY
PEIRSLRYPKPGTQNPTVTLRVADLKDPLKVHITDLHPPQIIANEDHYFSSASWVSHSKI
AVVWLNRPQNISVVSVCKAPLFQCIETHRVSGDGRGWVDTVAVPLFAANASIYVAISPLR
DGLFGYFRHIVHVDIDKNRVLPLTHGPYEVNRLLHWDQLDNWIYFLGTPERLPSQQHLYR
VSALPARQGQALRSPDCLTCPAVSQWSEGYDEGHTKSPPKLVTAWDDDWEDSEEAEAQP
P
QPALPVEQQPPGRGQSAPLPPPPSDCLYHEAKFPISRQAKYVLIDCLGPVVPTSILYGLK
SAAADSAKTKRHSTQQEEPPTEGGDEKPGEEPKSQFLELLVIVQNNTRLKEKMAKTAMPQ
IKTFPVMISGGYHAQVRLYLPPVLREDEITRYPTILHVYSGPGSQLVTDHWHVDWNTYLS
GSKDYIVVEIDGRGSAGQGYQLLHEVYKRLGSVEVSDQLEVSEYLRDNLHFIDSRRMGVW
GWSYGGYTAALALAGQQSIFQCGISVSPVTNWKLYDSTYAERYLSFPNVTDNYKGYEESD
LSKYVDNLRDRQFLLVHGTADDNVHVQQSMVLARSLTSKGVLYKQQIYPDEGHSLSGVKR
HLYRSMTAFFEDCFKKLVPPESKAGLGNGGDMQQQ
>M96060000013 EPS15_HUMAN
MAAAAQLSLTQLSSGNPVYEKYYRQVDTGNTGRVLASDAAAFLKKSGLPDLILGKIWDLA
DTDGKGILNKQEFFVALRLVACAQNGLEVSLSSLNLAVPPPRFHDTSSPLLISGTSAAEL
PWAVKPEDKAKYDAIFDSLSPVNGFLSGDKVKPVLLNSKLPVDILGRVWELSDIDHDGML
DRDEFAVAMFLVYCALEKEPVPMSLPPALVPPSKRKTWVVSPAEKAKYDEIFLKTDKDMD
GFVSGLEVREIFLKTGLPSTLLAHIWSLCDTKDCGKLSKDQFALAFHLISQKLIKGIDPP
HVLTPEMIPPSDRASLQKNIIGSSPVADFSAIKELDTLNNEIVDLQREKNNVEQDLKEKE
DTIKQRTSEVQDLQDEVQRENTNLQKLQAQKQQVQELLDELDEQKAQLEEQLKEVRKKCA
EEAQLISSLKAELTSQESQISTYEEELAKAREELSRLQQETAELEESVESGKAQLEPLQQ
HLQDSQQEISSMQMKLMEMKDLENHNSQLNWCSSPHSILVNGATDYCSLSTSSSETANLN
EHVEGQSNLESEPIHQESPARSSPELLPSGVTDENEVTTAVTEKVCSELDNNRHSKEEDP
FNVDSSSLTGPVADTNLDFFQSDPFVGSDPFKDDPFGKIDPFGGDPFKGSDPFASDCFFR
QSTDPFATSSTDPFSAANNSSITSVETLKHNDPFAPGGTVVAASDSATDPFASVFGNESF
GGGFADFSTLSKVNNEDPFRSATSSSVSNVVITKNVFEETSVKSEDEPPALPPKIGTPTR
PCPLPPGKRSINKLDSPDPFKLNDPFQPFPGNDSPKEKDPEIFCDPFTSATTTTNKEADP
SNFANFSAYPSEEDMIEWAKRESEREEEQRLARLNQQEQEDLELAIALSKSEISEA
>M96060000014 ARRB2_HUMAN
MGEKPGTRVFKKSSPNCKLTVYLGKRDFVDHLDKVDPVDGVVLVDPDYLKDRKVFVTLTC
AFRYGREDLDVLGLSFRKDLFIATYQAFPPVPNPPRPPTRLQDRLLRKLGQHAHPFFFTI
PQNLPCSVTLQPGPEDTGKACGVDFEIRAFCAKSLEEKSHKRNSVRLVIRKVQFAPEKPG
PQPSAETTRHFLMSDRSLHLEASLDKELYYHGEPLNVNVHVTNNSTKTVKKIKVSVRQYA
DICLFSTAQYKCPVAQLEQDDQVSPSSTFCKVYTITPLLSDNREKRGLALDGKLKHEDTN
LASSTIVKEGANKEVLGILVSYRVKVKLVVSRGGDVSVELPFVLMHPKPHDHIPLPRPQS
AAPETDVPVDTNLIEFDTNYATDDDIVFEDFARLRLKGMKDDDYDDQLC
```

```
>M6560610000015 D5GCF2_TUBMM
MSRPPLSLTAELEKLEQSITLTLQEIDHNFSKAHRIVTTSIIPIVERYAKESEAVWEGSK
FWKQFFEASANVALSNYQEEEETYEETGVATNAETYITASSPGNYGEQSTRITQEDDPRH
RETQWENIESPFDALRKDDDADMTFEPTLLPATPQTSRRTRPAKGSFETPKSSPFYPSAA
GAIGKKTPGGANEDQLLHRVLDKNWRLQATPLGKPPPSRYRTIGAATATTPKAQILPPPG
SESPMSSPPKPHFYSADIFSSPIPGFGGFDGGKKPKPSTTSNTTVLAGDPKTPISRRYGT
AKVTTTHHHELSQQGDEGRFAYGYDDDDDSDDLDLPPGLSPPVTIQFSLPPSKLLATPAR
EASRRIVHDILQTAGAADESGATGGSSPPVVRDVGPLDDTF
>M99250000016 Q9MZB4_CAPHI
RLLLEYTDSNYEEKKYTMGDAPDYDRSQWLNEKSKLGLDFPNLPYLIDGTHKLTQSNAIL
RHIARKYNMCGETEEEKIRVDLLENQVMDVRLHMARICYSPDFEKLKPGYLKEIPGRMKL
FSVFLGKRCWFAGNKLTYVDFLAYDILDLQRIFEPRCLDEFRNLKDFLTRFEGLKKISGY
MKSSRFLP
>M99250000017 GSTP1_CAPHI
MASYTIVYFPVQGRCEAMRMLLADQDQSWKEEVVAMQSWLQGPLKASCLYGQLPKFQDG
D
LTLYQSNAILRHLGRTLGLYGKDQREAALVDMVNDGVEDLRCKYVSLIYTNYQAGKEDYV
KALPQHLKPFETLLSQNKGGQAFIVGDQISFADYNLLDLLRIHQVLAPSCLDSFPLLSAY
VARLNSRPKLKAFLASPEHVNRPINGNGKQ
>M2813090000018 Q6HEJ4_BACHK
MAFEFKLPDIGEGIHEGEIVKWFIKPGDEVNEDDVLLEVQNDKAVVEIPSPVKGKVLEVL
VEEGTVAVVGDTLIKFDAPGYENLKFKGDDHDEAPKAEEAKEEAPKAEATPAATAEVVNE
RVIAMPSVRKYARENGVDIHKVAGSGKNGRIVKADIDAFANGGQAVAATEAPAAVEATPA
AAKEEAPKAQPIPAGEYPETREKMSGIRKAIAKAMVNSKHTAPHVTLMDEVDVTELVAHR
KKFKAVAADKGIKLTYLPYVVKALTSALREYPMLNTSLDDASQEVVHKHYFNIGIAADTD
KGLLVPVVKDTDRKSIFTISNEINDLAGKAREGRLAPAEMKGASCTITNIGSAGGQWFTP
VINHPEVAILGIGRIAEKPVVKNGEIVAAPVLALSLSFDHRLIDGATAQKALNQIKRLLN
DPQLLVMEA
>M56610000019 Q95VF2_LEIDO
MGKDKVHMNLVVVGHVDAGKSTATGHLIYKCGGIDKRTIEKFEKEAAEIGKASFKYAWVL
DKLKAERERGITIDIALWKFESPKSVFTIIDAPGHRDFIKNMITGTSQADAAILMIDSTH
GGFEAGISKDGQTREHALLAFTLGVKQMVVCCNKMDDKTVTYAQSRYDEISKEVGAYLKR
VGYNPEKVRFIPISGWQGDNMIERSDNMPWYKGPTLLDALDMLEPPVRPVDKPLRLPLQD
VYKIGGIGTVPVGRVETGIMKPGDVVTFAPANVTTEVKSIEMHHEQLAEAQPGDNVGFNV
KNVSVKDIRRGNVCGNSKNDPPKEAADFTAQVIVLNHPGQISNGYAPVLDCHTSHIACRF
AEIESKIDRRSGKELEKNPKAIKSGDAAIVKMVPQKPMCVEVFNDYAPLGRFAVRDMRQT
VAVGIIKGVNKKEGSGGKVTKAAAKAAKK
>M9810870000020 E9BTJ1_LEIDB
MSRVTIFQSQLPACNRLKTPYESELIATVKKLTTPGKGLLAADESIGSCTKRFEPIGLSN
TEEHRRQYRALMLEAEGFEQYISGVILHEETVGQKASNGQTFPEYLTARGVVPGIKTDMG
LCPLLEGAEGEQMTEGLDGYVKRASAYYKKGCRFCKWRNVYKIQNGTVSESAVRFNAETL
ARYAILSQISGLVPIVEPEVMIDGKHDIDTCQRVSEHVWREVVAALQRHGVIWEGCLLKP
NMVVPGAESGQTAAPAQVAHYTVMTLARTMPAMLPGVMFLSGGLSEVQASEYLNAINNSP
LPRPYFLSFSYARALQSSALKAWGGKDSGVAAGRRAFLHRARMNSMAQLGKYKRADDDA
S
SSSLYVKGNTY
>M58330000021 ENO_PLAFA
MAHVITRINAREILDSRGNPTVEVDLETNLGIFRAAVPSGASTGIYEALELRDNDKSRYL
GKGVQKAIKNINEIIAPKLIGMNCTEQKKIDNLMVEELDGSKNEWGWSKSKLGANAILAI
SMAVCRAGAAPNKVSLYKYLAQLAGKKSDQMVLPVPCLNVINGGSHAGNKLSFQEFMIVP
VGAPSFKEALRYGAEVYHTLKSEIKKKYGIDATNVGDEGGFAPNILNANEALDLLVTAIK
SAGYEGKVKIAMDVAASEFYNSENKTYDLDFKTPNNDKSLVKTGAQLVDLYIDLVKKYPI
VSIEDPFDQDDWENYAKLTAAIGKDVQIVGDDLLVTNPTRITKALEKNACNALLLKVNQI
```

```
GSITEAIEACLLSQKNNWGVMVSHRSGETEDVFIADLVVALRTGQIKTGAPCRSERNAKY
NQLLRIEESLGNNAVFAGEKFRLQLN
>M96060000022 GPX4_HUMAN
MSLGRLCRLLKPALLCGALAAPGLAGTMCASRDDWRCARSMHEFSAKDIDGHMVNLDKY
R
GFVCIVTNVASQUGKTEVNYTQLVDLHARYAECGLRILAFPCNQFGKQEPGSNEEIKEFA
AGYNVKFDMFSKICVNGDDAHPLWKWMKIQPKGKGILGNAIKWNFTKFLIDKNGCVVKRY
GPMEEPLVIEKDLPHYF
>M96060000023 TPIS_HUMAN
MAEDGEEAEFHFAALYISGQWPRLRADTDLQRLGSSAMAPSRKFFVGGNWKMNGRKQSL
G
ELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGM
IKDCGATWVVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKV
VFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQ
STRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ
>M96060000024 VDAC2_HUMAN
MATHGQTCARPMCIPPSYADLGKAARDIFNKGFGFGLVKLDVKTKSCSGVEFSTSGSSNT
DTGKVTGTLETKYKWCEYGLTFTEKWNTDNTLGTEIAIEDQICQGLKLTFDTTFSPNTGK
KSGKIKSSYKRECINLGCDVDFDFAGPAIHGSAVFGYEGWLAGYQMTFDSAKSKLTRNNF
AVGYRTGDFQLHTNVNDGTEFGGSIYQKVCEDLDTSVNLAWTSGTNCTRFGIAAKYQLDP
TASISAKVNNSSLIGVGYTQTLRPGVKLTLSALVDGKSINAGGHKVGLALELEA
>M1711010000025 Q8DNW9_STRR6
MTRYQDDFYDAINGEWQQTAEIPADKSQTGGFVDLDQEIEDLMLATTDKWLAGEEVPEDA
ILENFVKYHRLVRDFDKREADGITPVLPLLKEFQELETFADFTAKLAEFELAGKPNFLPF
GVSPDFMDARINVLWASAPSTILPDTTYYAEEHPQREELLTLWKESSANLLKAYDFSDEE
IEDLLEKRLELDRRVAAVVLSNEESSEYAKLYHPYSYEDFKKFAPALPLDDFFKAVIGQL
PDKVIVDEERFWQAAEQFYSEESWSLLKATLILSVVNLSTSYLTEDIRVLSGAYSRALSG
VPEAKDKVKAAYHLAQEPFKQALGLWYAREKFSPEAKADVEKKVATMIDVYKERLLKNDW
LTPETCKQAIVKLNVIKPYIGYPEELPARYKDKVVNETASLFENALAFARVEIKHSWSKW
NQPVDYKEWGMPAHMVNAYYNPQKNLIVFPAAILQAPFYDLHQSSSANYGGIGAVIAHEI
SHAFDTNGASFDENGSLKDWWTESDYAAFKEKTQKVIDQFDGQDSYGATINGKLTVSENV
ADLGGIAAALEAAKREADFSAEEFFYNFGRIWRMKGRPEFMKLLASVDVHAPAKLRVNVQ
VPNFDDFFTTYDVKEGDGMWRSPEERVIIW
>M8887450000026 E7S2A7_STRA8
MSLVGKEIIEFSAQAYHDGKFITVTNEDVKGKWAVFCFYPADFSFVCPTELGDLQEQYET
LKSLDVEVYSVSTDTHFVHKAWHDDSDVVGTITYPMIGDPSHLISQGFDVLGQDGLAQRG
TFIIDPDGVIQMMEINADGIGRDASTLIDKVRAAQYIRQHPGEVCPAKWKEGAETLTPSL
DLVGKI
>M1711010000027 ALF_STRR6
MAIVSAEKFVQAARDNGYAVGGFNTNNLEWTQAILRAAEAKKAPVLIQTSMGAAKYMGGY
KVARNLIANLVESMGITVPVAIHLDHGHYEDALECIEVGYTSIMFDGSHLPVEENLKLAK
EVVEKAHAKGISVEAEVGTIGGEEDGIIGKGELAPIEDAKAMVETGIDFLAAGIGNIHGP
YPVNWEGLDLDHLQKLTEALPGFPIVLHGGSGIPDEQIQAAIKLGVAKVNVNTECQIAFA
NATRKFARDYEANEAEYDKKKLFDPRKFLADGVKAIQASVEERIDVFGSEGKA
>M1711010000028 ALF_STRR6
MAIVSAEKFVQAARDNGYAVGGFNTNNLEWTQAILRAAEAKKAPVLIQTSMGAAKYMGGY
KVARNLIANLVESMGITVPVAIHLDHGHYEDALECIEVGYTSIMFDGSHLPVEENLKLAK
EVVEKAHAKGISVEAEVGTIGGEEDGIIGKGELAPIEDAKAMVETGIDFLAAGIGNIHGP
YPVNWEGLDLDHLQKLTEALPGFPIVLHGGSGIPDEQIQAAIKLGVAKVNVNTECQIAFA
NATRKFARDYEANEAEYDKKKLFDPRKFLADGVKAIQASVEERIDVFGSEGKA
>M72270000029 CH60_DROME
MFRLPVSLARSSISRQLAMRGYAKDVRFGPEVRAMMLQGVDVLADAVAVTMGPKGRNVII
EQSWGSPKITKDGVTVAKSIELKDKFQNIGAKLVQDVANNTNEEAGDGTTTATVLARAIA
```

```
KEGFEKISKGANPVEIRRGVMLAVETVKDNLKTMSRPVSTPEEIAQVATISANGDQAIGN
LISEAMKKVGRDGVITVKDGKTLTDELEVIEGMKFDRGYISPYFINSSKGAKVEFQDALL
LLSEKKISSVQSIIPALELANAQRKPLVIIAEDIDGEALSTLVVNRLKIGLQVAAVKAPG
FGDNRKSTLTDMAIASGGIVFGDDADLVKLEDVKVSDLGQVGEVVITKDDTLLLKGKGKK
DDVLRRANQIKDQIEDTTSEYEKEKLQERLARLASGVALLRVGGSSEVEVNEKKDRVHDA
LNATRAAVEEGIVPGGGTALLRCIEKLEGVETTNEDQKLGVEIVRRALRMPCMTIAKNAG
VDGAMVVAKVENQAGDYGYDALKGEYGNLIEKGIIDPTKVVRTAITDASGVASLLTTAEA
VVTEIPKEDGAPAMPGMGGMGGMGGMGGMGGMM
>M72270000030 CSN7_DROME
MTQDMLLGNEEPSKSKETFLEKFCVLAKSSTGAALLDVIRQALEAPNVFVFGELLAEPSV
LQLKDGPDSKHFETLNLFAYGTYKEYRAQPEKFIELTPAMQKKLQHLTIVSLAIKAKSIP
YALLLSELEIDNVRHLEDIIIEAIYADIIHGKLFQNTRILEVDYAQGRDIPPGYTGQIVE
TLQAWVNSCDSVSNCIEMQIKYANAEKSKRLINKERVEQDLINLKKVLKSQTSDSDESMQ
IDTHGPGTSGGLGQSELRKKPSKLRNPRSAAVGLKFSK
>M99860000031 ALDOA_RABIT
MPHSHPALTPEQKKELSDIAHRIVAPGKGILAADESTGSIAKRLQSIGTENTEENRRFYR
QLLLTADDRVNPCIGGVILFHETLYQKADDGRPFPQVIKSKGGVVGIKVDKGVVPLAGTN
GETTTQGLDGLSERCAQYKKDGADFAKWRCVLKIGEHTPSALAIMENANVLARYASICQQ
NGIVPIVEPEILPDGDHDLKRCQYVTEKVLAAVYKALSDHHIYLEGTLLKPNMVTPGHAC
TQKYSHEEIAMATVTALRRTVPPAVTGVTFLSGGQSEEEASINLNAINKCPLLKPWALTF
SYGRALQASALKAWGGKKENLKAAQEEYVKRALANSLACQGKYTPSGQAGAAASESLFIS
NHAY
>M30550000032 A8J7F6_CHLRE
MQATTRVPAKSGVSSSAKRVAASGRRVLVVPNAVKDVFMPALSSTMTEGKIVSWLKNVGD
KVKKGEALVVVESDKADMDVESFADGILGAIVVQEGERAVVGAPIAFVAENANEAPAAAP
APAPAPVAAPAPPAPTPVPAAPVGRADGRIVATPYAKQLAKDLKVDLATVAGTGPNGRIT
AADATTVSELRGTTKPFSTLQAAVARNMNESLKVPEFRVSYAITTDKLDALYQQLKPKGV
TMTALLAKACGVALAKHPLLYAACTPDGNGITYSSQINVALAVAMPDGGLITPVLKNADS
TDLYQMSRNWADLVKRARSKQLQPDEYNSGNFTISNLGMYGVETFDAILPPGTAAIMAVG
GSKPTVVASPDGMIGVKKVMNVNLTADHRIVYGADAAEFLQTLKAVIENPDQLLF
>M56710000033 Q95U89_LEIIN
MLRRLPTSCFLKRSQFRGFAATSPLLNLDYQMYRTATVREAAPQFSGQAVVNGAIKDINM
NDYKGKYIVLFFYPMDFTFVCPTEIIAFSDRHADFEKLNTQVVAVSCDSVYSHLAWVNTP
RKKGGLGEMHIPVLADKSMEIARDYGVLIEESGIALRGLFIIDKKGILRHSTINDLPVGR
NVDEALRVLEAFQYADENGDAIPCGWKPGQPTLDTTKAGEFFEKNM
>M21330000034 PGK_SPICI
MTNKKELKDVQVKGKKVLVRVDFNVPMKDGQVTDDNRIIAALPTIKYLIAQEAKVILFSH
LGKVKTADDLEKRDMAPVAKVLEQKLGQPVKFINAFEGKQLEEAINEMHNKEVILFQNTR
FADIINSNGQISVDSEGKAAAKRESKNDSALGKYWASLGDVFVNDAFGTAHRAHASNVGI
AENITESCLGFLVEKEVKMLSQCVDNPVKPFVAIIGGAKVSDKIGVIEHLLTKADKILIG
GGMAYTFFAAQGHKIGNSLLEVDKVEIAKTFLAKGQGKIILPIDALEAPEFADVPAKVTT
GFDIDDGYMGLDIGPKTIELFKKELADAKTVTWNGPMGVFEFKNYSIGTKAVCEAIAELK
GAFTLIGGGDSAAAAIQLGYKDKFTHISTGGGASLEYMEGKPLPGIEAVQSK
>M1854310000035 Q383B2_TRYB2
MLRRLATHGLQATCLTSEKLAYRYCLSICVPTIAESISSGKVVGWTKKVGDAVAEDEIIC
QIESDKLNVDVRAPAAGVITKINFEEGTVVDVGAELSTMKEGEAPAAKAETADKPKQNAP
AAAAPPKASPTEAAPKPAPAAAPVTSRGADPRVRSVRISSMRQRIADRLKASQNTCAMLT
TFNEIDMTPLIELRNRYKDDFFKKNGVKLGFMSPFVKACAIALQDVPIVNASFGTDCIEY
HDYVDISVAVSTPKGLVVPVLRDVQNSNFAQIEKQIADFGERARSNKLTMAEMTGGTFTI
SNGGVFGSWMGTPIVNPPQSAILGMHATKKKPWVVGNSVVPRDIMAVALTYDHRLIDGSD
AVTFLVKVKNLIEDPARIVLDLA
>M54760000036 HGT1_CANAX
MSSKIERIFSGPALKINTYLDKLPKIYNVFFIASISTIAGMMFGFDISSMSAFIGAEHYM
```

```
RYFNSPGSDIQGFITSSMALGSFFGSIASSFVSEPFGRRLSLLTCAFFWMVGAAIQSSVQ
NRAQLIIGRIISGIGVGFGSAVAPVYGAELAPRKIRGLIGGMFQFFVTLGIMIMFYLSFG
LGHINGVASFRIAWGLQIVPGLCLFLGCFFIPESPRWLAKQGQWEAAEEIVAKIQAHGDR
ENPDVLIEISEIKDQLLLEESSKQIGYATLFTKKYIQRTFTAIFAQIWQQLTGMNVMMYY
IVYIFQMAGYSGNSNLVASSIQYVINTCVTVPALYFIDKVGRRPLLIGGATMMMAFQFGL
AGILGQYSIPWPDSGNDSVNIRIPEDNKSASKGAIACCYLFVASFAFTWGVGIWVYCAEI
WGDNRVAQRGNAISTSANWILNFAIAMYTPTGFKNISWKTYIIYGVFCFAMATHVYFGFP
ETKGKRLEEIGQMWEERVPAWRSRSWQPTVPIASDAELARKMEVEHEEDKLMNEDSNSE
S
RENQA
>M5592920000037 GPD2_YEAST
MLAVRRLTRYTFLKRTHPVLYTRRAYKILPSRSTFLRRSLLQTQLHSKMTAHTNIKQHKH
CHEDHPIRRSDSAVSIVHLKRAPFKVTVIGSGNWGTTIAKVIAENTELHSHIFEPEVRMW
VFDEKIGDENLTDIINTRHQNVKYLPNIDLPHNLVADPDLLHSIKGADILVFNIPHQFLP
NIVKQLQGHVAPHVRAISCLKGFELGSKGVQLLSSYVTDELGIQCGALSGANLAPEVAKE
HWSETTVAYQLPKDYQGDGKDVDHKILKLLFHRPYFHVNVIDDVAGISIAGALKNVVALA
CGFVEGMGWGNNASAAIQRLGLGEIIKFGRMFFPESKVETYYQESAGVADLITTCSGGRN
VKVATYMAKTGKSALEAEKELLNGQSAQGIITCREVHEWLQTCELTQEFPLFEAVYQIVY
NNVRMEDLPEMIEELDIDDE
>M5592920000038 SODC_YEAST
MVQAVAVLKGDAGVSGVVKFEQASESEPTTVSYEIAGNSPNAERGFHIHEFGDATNGCVS
AGPHFNPFKKTHGAPTDEVRHVGDMGNVKTDENGVAKGSFKDSLIKLIGPTSVVGRSVVI
HAGQDDLGKGDTEESLKTGNAGPRPACGVIGLTN
>M833330000039 ADHE_ECOLI
MAVTNVAELNALVERVKKAQREYASFTQEQVDKIFRAAALAAADARIPLAKMAVAESGMG
IVEDKVIKNHFASEYIYNAYKDEKTCGVLSEDDTFGTITIAEPIGIICGIVPTTNPTSTA
IFKSLISLKTRNAIIFSPHPRAKDATNKAADIVLQAAIAAGAPKDLIGWIDQPSVELSNA
LMHHPDINLILATGGPGMVKAAYSSGKPAIGVGAGNTPVVIDETADIKRAVASVLMSKTF
DNGVICASEQSVVVVDSVYDAVRERFATHGGYLLQGKELKAVQDVILKNGALNAAIVGQP
AYKIAELAGFSVPENTKILIGEVTVVDESEPFAHEKLSPTLAMYRAKDFEDAVEKAEKLV
AMGGIGHTSCLYTDQDNQPARVSYFGQKMKTARILINTPASQGGIGDLYNFKLAPSLTLG
CGSWGGNSISENVGPKHLINKKTVAKRAENMLWHKLPKSIYFRRGSLPIALDEVITDGHK
RALIVTDRFLFNNGYADQITSVLKAAGVETEVFFEVEADPTLSIVRKGAELANSFKPDVI
IALGGGSPMDAAKIMWVMYEHPETHFEELALRFMDIRKRIYKFPKMGVKAKMIAVTTTSG
TGSEVTPFAVVTDDATGQKYPLADYALTPDMAIVDANLVMDMPKSLCAFGGLDAVTHAME
AYVSVLASEFSDGQALQALKLLKEYLPASYHEGSKNPVARERVHSAATIAGIAFANAFLG
VCHSMAHKLGSQFHIPHGLANALLICNVIRYNANDNPTKQTAFSQYDRPQARRRYAEIAD
HLGLSAPGDRTAAKIEKLLAWLETLKAELGIPKSIREAGVQEADFLANVDKLSEDAFDDQ
CTGANPRYPLISELKQILLDTYYGRDYVEGETAAKKEAAPAKAEKKAKKSA
>M96060000040 PGK1_HUMAN
MSLSNKLTLDKLDVKGKRVVMRVDFNVPMKNNQITNNQRIKAAVPSIKFCLDNGAKSVVL
MSHLGRPDGVPMPDKYSLEPVAVELKSLLGKDVLFLKDCVGPEVEKACANPAAGSVILLE
NLRFHVEEEGKGKDASGNKVKAEPAKIEAFRASLSKLGDVYVNDAFGTAHRAHSSMVGVN
LPQKAGGFLMKKELNYFAKALESPERPFLAILGGAKVADKIQLINNMLDKVNEMIIGGGM
AFTFLKVLNNMEIGTSLFDEEGAKIVKDLMSKAEKNGVKITLPVDFVTADKFDENAKTGQ
ATVASGIPAGWMGLDCGPESSKKYAEAVTRAKQIVWNGPVGVFEWEAFARGTKALMDEV
V

KATSRGCITIIGGGDTATCCAKWNTEDKVSHVSTGGGASLELLEGKVLPGVDALSNI
```

As previously mentioned Chapter 4, the results for combined, FunFOLDQ and HHblits were variable and two examples were given; one of which was a good prediction example (ACE_RAT) and one poor prediction (ACH1_CANAL). This Appendix has further examples five of which have predicted at least one correct GO term and five of which have predicted no correct predictions. As the combined method was examined in CAFA3, further results for this method are given below in the Tables S.22-S.31. Tables S.22-S.26 are considered good predictions due to having correct predictions, whereas Tables S.27-S.31 are considered poor predictions due to no correct predictions.

Table S.22. GO terms predicted by combined method for 1433E_MOUSE

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	**Predicted GO term**	**Comparison to CAFA3 annotation**
Biological Process	GO:0000086	-
Biological Process	GO:0000278	-
Biological Process	**GO:0001764**	Exact match in CAFA3
Molecular Function	**GO:0005515**	Ancestor of GO:0019904
Cellular Component	**GO:0005737**	Ancestor of GO:0005739 and GO:0005829
Cellular Component	**GO:0005739**	Exact match in CAFA3
Cellular Component	**GO:0005829**	Exact match in CAFA3
Cellular Component	GO:0005871	-
Biological Process	**GO:0006605**	Exact match in CAFA3
Molecular Function	**GO:0019904**	Exact match in CAFA3

Table S.23. GO terms predicted by combined method for ACEA_HYMPE

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Comparison to CAFA3 annotation
Biological Process	**GO:0001101**	Exact match in CAFA3
Molecular Function	**GO:0003824**	Ancestor of GO:0004451
Molecular Function	**GO:0004451**	Exact match in CAFA3
Cellular Component	GO:0005737	Ancestor of GO:0005829
Cellular Component	GO:0005829	-
Cellular Component	GO:0005886	-
Biological Process	**GO:0006097**	Exact match in CAFA3
Biological Process	**GO:0006099**	Exact match in CAFA3
Biological Process	GO:0006102	-
Biological Process	**GO:0008152**	Ancestor of GO:0006102

Table S.24. GO terms predicted by combined method for 1433Z_MOUSE

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Comparison to CAFA3 annotation
Biological Process	GO:0000086	-
Biological Process	GO:0000278	-
Biological Process	GO:0001764	-
Molecular Function	GO:0005515	Ancestor of GO:0019904
Cellular Component	**GO:0005737**	Ancestor of GO:0005739
Cellular Component	**GO:0005739**	Exact match in CAFA3
Cellular Component	**GO:0005829**	Exact match in CAFA3
Cellular Component	GO:0005871	-
Biological Process	**GO:0006605**	Exact match in CAFA3
Molecular Function	**GO:0019904**	Exact match in CAFA3

Table S.25. GO terms predicted by combined method for 6P21_YEAST

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold, ancestor terms are given in bold and green and UniProtKB matches are in blue and bold

GO domain	Predicted GO term	Comparison to CAFA3 annotation
Molecular Function	GO:0000166	-
Molecular Function	**GO:0003824**	Ancestor of GO:0003873
Molecular Function	**GO:0003873**	Exact match in CAFA3
Molecular Function	GO:0004331	-
Molecular Function	**GO:0005524**	Match on UniProtKB
Biological Process	**GO:0006000**	Match on UniProtKB
Biological Process	**GO:0006003**	Exact match in CAFA3
Biological Process	GO:0008152	Ancestor of GO:0006003
Molecular Function	GO:0016301	-
Biological Process	GO:0016310	-

Table S.26. GO terms predicted by combined method for ACE_MOUSE

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Comparison to CAFA3 annotation
Biological Process	GO:0006508	-
Molecular Function	**GO:0008237**	Exact match in CAFA3
Molecular Function	**GO:0008241**	Exact match in CAFA3
Cellular Component	GO:0016020	-

Table S.27. GO terms predicted by combined method for 2AAA_DROME

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Relevant info
No entry	GO:0005087	-
Molecular Function	GO:0005515	-
Cellular Component	GO:0005634	Ancestor of UniProtKB related term GO:0005654
Cellular Component	GO:0005635	-
Cellular Component	GO:0005643	-
Cellular Component	GO:0005737	Ancestor of UniProtKB related term GO:0005829
Biological Process	GO:0006606	-
Biological Process	GO:0006612	-
Biological Process	GO:0006656	-
Biological Process	GO:0006810	-

Table S.28. GO terms predicted by combined method for 6PGD_SHEEP

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Relevant info
Molecular Function	GO:0000166	Ancestor of UniProtKB related term GO:0050661
Molecular Function	GO:0003979	-
Molecular Function	GO:0004735	-
Biological Process	GO:0006561	-
Molecular Function	GO:0016491	Ancestor of UniProtKB related term GO:0004616
Molecular Function	GO:0016616	Ancestor of UniProtKB related term GO:0004616
Biological Process	GO:0042121	-
Molecular Function	GO:0047919	-
Biological Process	GO:0051287	-

Biological Process	GO:0055114	Obsolete entry

Table S.29. GO terms predicted by combined method for 3HAO_HUMAN

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Relevant info
Cellular Component	GO:0005737	-
Molecular Function	GO:0008127	-
Molecular Function	GO:0016491	-
Molecular Function	GO:0016702	-
Molecular Function	GO:0016829	-
Molecular Function	GO:0016831	-
Biological Process	GO:0017000	-
Molecular Function	GO:0045735	-
Molecular Function	GO:0046872	Ancestor of UniProtKB related term GO:0008198
Molecular Function	GO:0055114	Obsolete entry

Table S.30. GO terms predicted by combined method for ACL6A_HUMAN

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold, ancestor terms are given in bold and green and UniProtKB matches are in blue and bold

GO domain	Predicted GO term	Relevant info
Molecular Function	GO:0000166	-
No entry	GO:0001725	-
Molecular Function	GO:0004844	-
Molecular Function	GO:0005524	-
Cellular Component	GO:0005694	-
Cellular Component	GO:0005737	-
Cellular Component	GO:0005856	-
Cellular Component	GO:0005865	-
Cellular Component	GO:0005884	Actin filament (GO term) and protein is an actin-protein 6A
Biological Process	**GO:0006281**	Match on UniprotKB
Biological Process	GO:0006284	Base-exicision repair (GO term) is a DNA repair so is related to GO:0006281
Cellular Component	GO:0015629	Actin cytoskeleton
Molecular Function	GO:0016799	-
Cellular Component	GO:0030017	-
Biological Process	GO:0030240	-
Biological Process	GO:0048741	-
Biological Process	GO:0051276	-

Table S.31. GO terms predicted by combined method for ACL7B_MOUSE

Predicted terms with their associated GO terms are given below. Correct matches are given in black and bold and ancestor terms are given in bold and green

GO domain	Predicted GO term	Relevant info
Molecular Function	GO:0000166	-
Biological Process	GO:0000902	-
Molecular Function	GO:0003674	-
Molecular Function	GO:0005524	-
Biological process	GO:0007049	-
Cellular Component	GO:0016020	-
No entry	GO:0043241	-
Molecular Function	GO:0051082	-
Biological Process	GO:0051085	-
Biological Process	GO:0051301	-

The chart below explains the difference between the colours on hierarchical mapping of GO ontology categories and denote the organism in which the activity was identified in

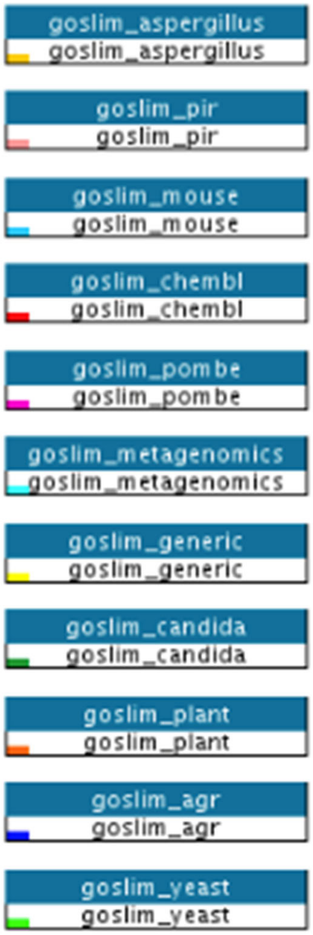


Figure S.75. GO slim colours

Appendix 4

Source code to run the different grid box calculations for AutoDock Vina

```
#!/usr/bin/env python3

# Program: Centre_v2.py
# Function: Calculate the "bounding box" for the ligand.
# Determine the "centre of geometry" of the ligand (atoms are equal mass) to get the centre
of the cube and then the maximum
# distance between ligand atoms will be the size of the cube needed to contain the ligand.


# Call functions and create empty lists/dicts to store Residue and Ligand Atom co-ordinates.
import math
import numpy
import pandas as pd
from scipy.spatial import distance
Lig_coords = []
Lig_x = {}
Lig_y = {}
Lig_z = {}
Rg_x = {}
Rg_y = {}
Rg_z = {}
dist_x = {}
dist_y = {}
dist_z = {}
New_list=[]

# Main routine to calculate the centre and maximum distances
with open('T0849.txt',"r") as pdb2, open('CentreV2_out.txt', "w") as Lig:
    # Read each line of the pdb file
    for line in pdb2:
        # Slice the lines and Limit the lines to chosen Ligand atoms
        if line[0:6]=='HETATM' and line[17:20]=='GSH':
            # Assign the split line parts to varibles to use below
            sp0=line[0:6]
            sp1=line[7:11]
            sp2=line[13:16]
            sp3=line[17:20]
            sp4=line[22:26]
            sp5=line[31:38]
            sp6=line[39:46]
            sp7=line[47:54]
            # Append selected parts of the lines to the Lig_coords list and create x, y and z
dictionaries
            Lig_coords.append([sp1,sp2,sp3,sp4,float(sp5),float(sp6),float(sp7)])
            Lig_x[sp1] = float(sp5)
            Lig_y[sp1] = float(sp6)
            Lig_z[sp1] = float(sp7)
# Calculation of distances from centre to furthest atom in x,y,z planes - REMOVED in place
of Rg calculation
```

```
#            # Compare the coordinates to find the maximum distance between any two atoms
#            # List through the Lig_coords list once
#            for i in range(len(Lig_coords)):
#                s=Lig_coords[i]
#                # Create a nested for loop to list through the Lig_coords list again
#                for j in range(len(Lig_coords)):
#                    p=Lig_coords[j]
#                # Calculate the distance between x coordinates, y coordinates and z coordinates,
use square to get positive values
#                dist_x[i,j] = float(round(math.sqrt((p[4]-s[4])**2),4))
#                dist_y[i,j] = float(round(math.sqrt((p[5]-s[5])**2),4))
#                dist_z[i,j] = float(round(math.sqrt((p[6]-s[6])**2),4))
    # Format the Lig_coords list into a dataframe table for output to check that it looks
sensible
    pd.set_option('display.max_rows', len(Lig_coords))
    df = pd.DataFrame(Lig_coords)
    new_header = df.iloc[0] #grab the first row for the header
    df = df[1:] #take the data less the header row
    df.columns = new_header #set the first row as the header
    Lig.write(str(df))
    Lig.write('\n')
    pd.reset_option('display.max_rows')
# Calculation of distances from centre to furthest atom in x,y,z planes - REMOVED in place
of Rg calculation
## Calculate the maximum distance between any two atoms (including VdW radii) in the x, y
and z planes
#Mltpr = raw_input('Please choose a % to add to your maximum ligand distance: ')
#Pct = 1 + (float(Mltpr)/100)
#Maxx_dist = round(max(dist_x.values()) + (1.6 * 2),1)
#Maxy_dist = round(max(dist_y.values()) + (1.6 * 2),1)
#Maxz_dist = round(max(dist_z.values()) + (1.6 * 2),1)
# Calculate the xyz centroids from xyz dictionaries and assign the values to variables
Centroid_x = round(sum(Lig_x.values())/len(Lig_x),2)
Centroid_y = round(sum(Lig_y.values())/len(Lig_y),2)
Centroid_z = round(sum(Lig_z.values())/len(Lig_z),2)
# Calulate the radius of gyration (Rg) of the ligand
# In the x plane
for x in range(len(Lig_coords)):
    s=Lig_coords[x]
    R = (s[4]-Centroid_x)**2
    Rg_x[x] = round(float(R),4)
Rgx = round(math.sqrt(sum(Rg_x.values())),2)
# In the y plane
for y in range(len(Lig_coords)):
    s=Lig_coords[y]
    R = (s[5]-Centroid_y)**2
    Rg_y[y] = round(float(R),4)
Rgy = round(math.sqrt(sum(Rg_y.values())),2)
# In the z plane
for z in range(len(Lig_coords)):
    s=Lig_coords[z]
    R = (s[6]-Centroid_z)**2
    Rg_z[z] = round(float(R),4)
```

```
Rgz = round(math.sqrt(sum(Rg_z.values())),2)
#Add in the 2.9 times Rg multiplier to get the box size
Multp = 2.9
# Print out the centroid coordinates and the distances
print ('Radius of gyration in the x plane is ' + str(Rgx))
print ('Radius of gyration in the y plane is ' + str(Rgy))
print ('Radius of gyration in the z plane is ' + str(Rgz))
print ('Box size is now calculated by Rg x 2.9 in each plane:')
print ('Centroid x ' + str(Centroid_x) + ' box size in x plane = ' + str(round((Rgx * Multp),1)))
print ('Centroid y ' + str(Centroid_y) + ' box size in y plane = ' + str(round((Rgy * Multp),1)))
print ('Centroid z ' + str(Centroid_z) + ' box size in z plane = ' + str(round((Rgz * Multp),1)))
Lig.close()
# End
```

The remaining targets following docking with AutoDock Vina is shown below

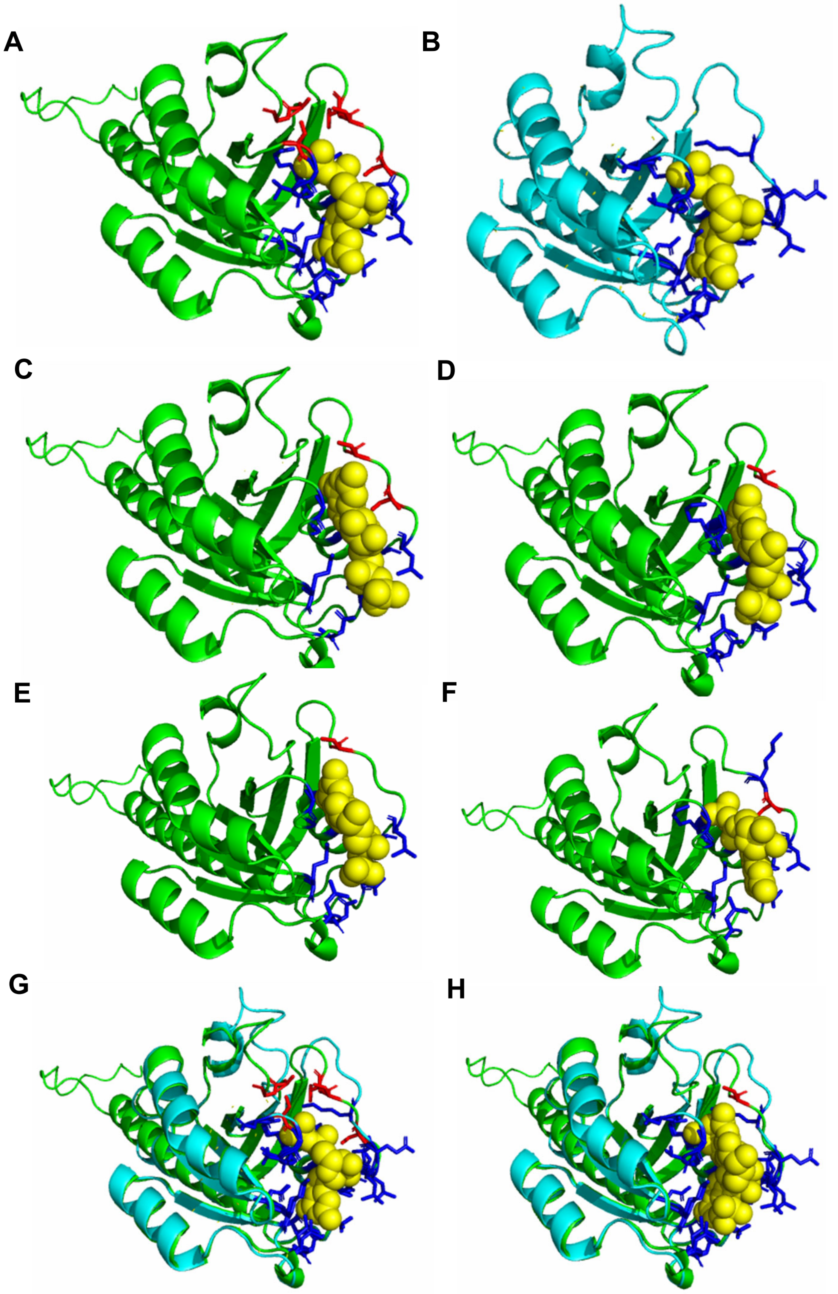


Figure S.76. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0798 (PDB ID 4ojk)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the GDP ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0798 (PDB ID 4ojk) shown as sticks and coloured blue, the GDP ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model from a 10% box calculation with the protein coloured green and the observed structure coloured cyan. BDT and MCC score of 0.65 and 0.74, respectively

Table S.32. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5 Å for T0798 (PDB ID 4ojk)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	14,16,19,29,31,33,35,118,121	0.51	-0.24	0.41	-0.387
2	35,36,38,39,61,63,66,67,69,73	-0.075	-0.83	0.066	-0.731
3	38,39,58,60,61,63,67,69,70,72,73	-0.079	-0.83	0.048	-0.749
4	**14,16,19,29,31,33,35,118,120,149**	**0.55**	**-0.20**	**0.46**	**-0.337**
5	**14,16,19,29,31,33,35,118,121,149**	**0.55**	**-0.20**	**0.46**	**-0.337**
6	14,18,29,33,34,35,36,118,121,149	0.47	-0.28	0.42	-0.377
7	35,36,37,38,39,61,63,67,68,69	-0.075	-0.83	0.070	-0.727
8	35,38,60,61,63,67,69	-0.062	-0.82	0.053	-0.744
9	38,39,60,61,63,66,67,69	-0.067	-0.82	0.040	-0.757

Table S.33. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0798 (PDB ID 4ojk)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	14,16,17,19,29,31,33,35,118,149	0.55	-0.203	0.46	-0.337
2	**14,16,17,18,19,29,30,31,35,118,120,121,149**	**0.74**	**-0.013**	**0.65**	**-0.147**
3	14,16,17,18,19,29,30,31,33,35,118	0.59	-0.163	0.51	-0.287
4	14,18,29,31,33,34,35,36,118,121,149	0.52	-0.233	0.47	-0.327
5	13,14,16,18,19,29,31,33,34,35,36,118,149	0.54	-0.213	0.55	-0.247
6	14,16,18,19,29,31,33,34,35,118	0.55	-0.203	0.46	-0.337
7	14,16,18,19,29,31,33,35,118	0.51	-0.243	0.41	-0.387
8	13,14,16,17,18,19,29,33,34,35,36,118,121	0.54	-0.213	0.55	-0.247
9	14,16,19,29,31,33,34,118,120	0.59	-0.163	0.44	-0.357

Table S.34. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculaton for T0798 (PDB ID 4ojk)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	14,16,19,29,31,33,35,118,121	0.51	-0.243	0.41	-0.387
2	**14,16,18,19,29,31,35,118,120,121,149**	**0.67**	**-0.083**	**0.55**	**-0.247**
3	14,16,18,19,29,30,31,33,35,118,121	0.59	-0.163	0.51	-0.287
4	14,18,29,31,33,34,35,36,118,121,149	0.52	-0.233	0.47	-0.327
5	13,14,16,17,19,29,31,33,34,35,36,118,121	0.54	-0.213	0.55	-0.247
6	16,19,29,30,31,33,34,35,118,121,149	0.59	-0.163	0.51	-0.287
7	14,29,31,33,34,35,118,120,121,149	0.55	-0.203	0.46	-0.337
8	13,14,16,19,29,31,33,34,35,36,118	0.44	-0.313	0.44	-0.357
9	Same as complex 2	-	-	-	-

Table S.35. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0798 (PDB ID 4ojk)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	14,16,19,29,31,35,118,149	0.54	-0.213	0.39	-0.407
2	13,14,16,17,18,19,29,30,31,33,35,61,118	0.54	-0.213	0.54	-0.257
3	14,16,19,29,31,35,118,120,121,149	0.63	-0.123	0.49	-0.307
4	14,16,18,19,29,30,31,33,35,118,121	0.59	-0.163	0.51	-0.287
5	14,18,29,31,33,34,35,36,118,121,149	0.52	-0.233	0.47	-0.327
6	13,14,16,17,18,19,31,33,34,35,36,118,120,121	0.58	-0.173	0.60	-0.197
7	13,14,17,18,19,31,33,34,35,118	0.47	-0.283	0.43	-0.367
8	14,16,18,19,30,31,33,35,85,88,118	0.45	-0.303	0.43	-0.367
9	**14,16,17,18,19,29,31,33,34,118,121,149**	**0.70**	**-0.053**	**0.60**	**-0.197**

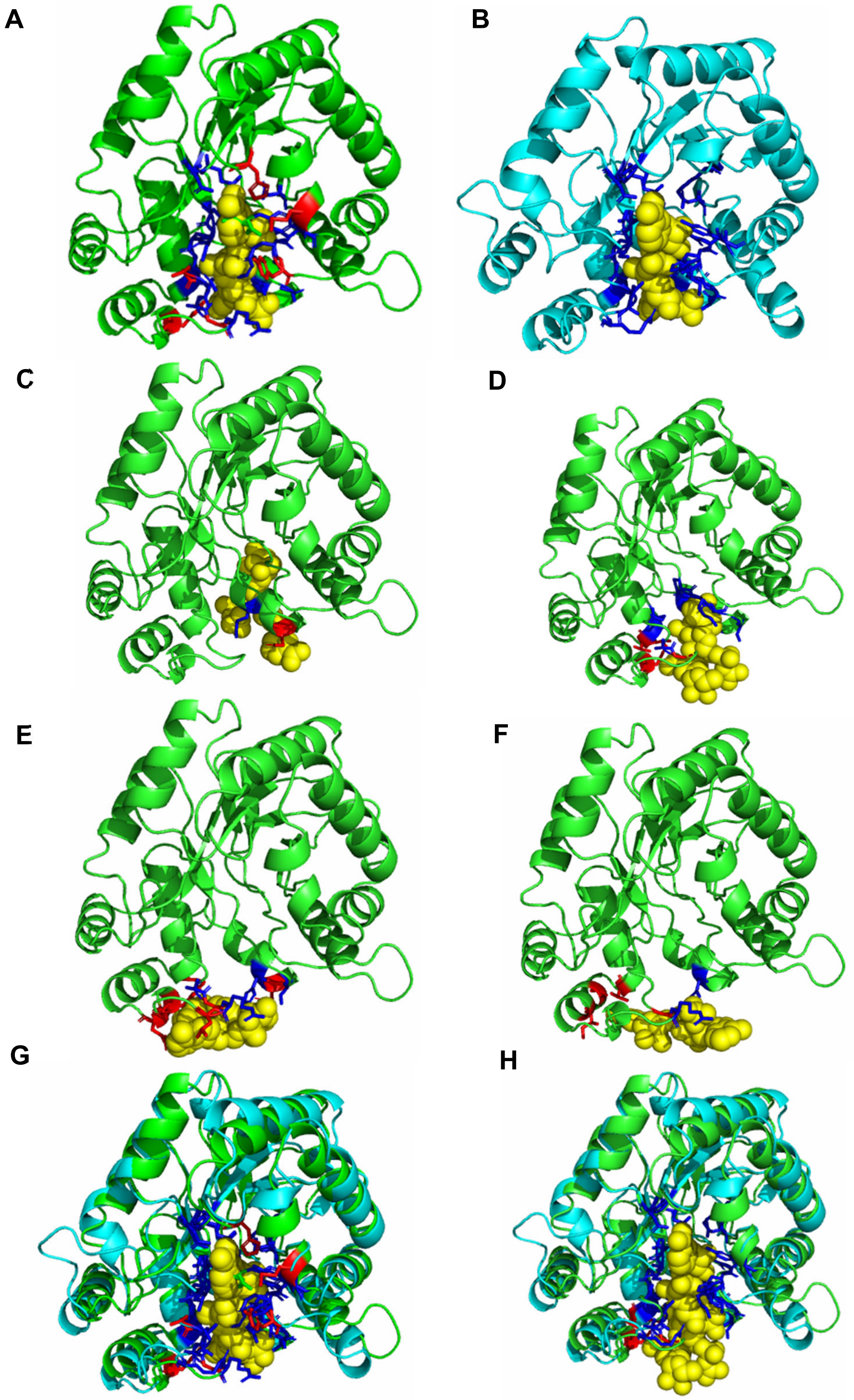


Figure S.77. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0807 (PDB ID 4wgh)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAP ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0807 (PDB ID 4wgh) shown as sticks and coloured blue, the NAP ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model from 10% grid box calculation with the protein coloured green and the observed structure coloured cyan. BDT and MCC score of 0.43 and 0.49, respectively

Table S.36. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0807 (PDB ID 4wgh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	10,128,131,132,137,159	-0.047	-0.818	0.0089	-0.8401
	2	132,137	-0.027	-0.798	0.0029	-0.8461
	3	10,131,132,134,137,157,158,159	-0.054	-0.825	0.012	-0.837
	4	10,128,131,132,134,137,157,158,159	-0.058	-0.829	0.013	-0.836
	5	10,128,131,132,134,137,159	-0.051	-0.822	0.010	-0.839
2	1	225,247,250,251,254	-0.050	-0.821	0.093	-0.756
	2	**247,251**	**0.119**	**-0.652**	0.053	-0.796
	3	1,2,225,247,250,251,254,256,257	0.020	-0.751	**0.10**	**-0.749**
	4	1,225,247,250,251,254,256	0.038	-0.733	0.098	-0.751
	5	247,250,254,256	-0.038	-0.809	0.045	-0.804
	6	1,2,225,247,250,251,254	0.038	-0.733	0.094	-0.755
3	1	88,92,125,128,129,132,159	-0.051	-0.822	0.010	-0.839
	2	88,92,125,129,157,158	-0.047	-0.818	0.009	-0.84
	3	88,125,128,129,159	-0.043	-0.814	0.01	-0.839
	4	88,92,125,126,128,129,159	-0.051	-0.822	0.011	-0.838
	5	88,92,125,128,129,132,159	-0.051	-0.822	0.010	-0.839
	6	88,92,125,126,128,129,132,159	-0.054	-0.825	0.012	-0.837
4	1	270,271,272,274	-0.038	-0.809	0.005	-0.844
	2	272,274	-0.027	-0.798	0.003	-0.846
	3	204,205,209,270,272,274	-0.047	-0.818	0.019	-0.83
	4	171,204,209,270,271,272,274	-0.051	-0.822	0.019	-0.83
	5	204,270,272,273,274	-0.043	-0.814	0.013	-0.836
	6	171,204,270,271,272,273,274	-0.051	-0.822	0.018	-0.831
5	1	10,11,13,137	-0.038	-0.809	0.006	-0.843
	2	10,13	-0.027	-0.798	0.003	-0.846
	3	10,11,73,74,137	-0.043	-0.814	0.008	-0.841
	4	10,11,12,13,137	-0.047	-0.818	0.009	-0.84
6	1	175,177,184,234,263,266,269	-0.051	-0.822	0.011	-0.838
	2	177,184,263,266	-0.038	-0.809	0.005	-0.844
	3	175,177,180,184,233,234,235,266,269	-0.058	-0.829	0.016	-0.833
	4	175,177,184,233,234,263,266,269	-0.054	-0.825	0.012	-0.837
	5	175,177,184,263,266	-0.043	-0.814	0.008	-0.841
	6	175,177,233,234,263,266,269	-0.051	-0.822	0.011	-0.838
7	1	184,188,189,233,234,235,265	-0.051	-0.822	0.016	-0.833
	2	6,7,184,188,189,235	-0.047	-0.818	0.014	-0.835
	3	184,189,233,234,235,265	-0.047	-0.818	0.013	-0.836
	4	184,188,189,233,234,235,265	-0.051	-0.822	0.016	-0.833
	5	183,184,188,189,233,234,235,265	-0.054	-0.825	0.018	-0.831

Table S.37. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0807 (PDB ID 4wgh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	196,200,207,222,224,241,242,243,245,248,251,252	0.48	-0.291	0.37	-0.479
	2	196,199,200,201,207,222,241,242,243,247,248,252	0.42	-0.351	0.35	-0.499
	3	196,207,222,224,241,242,243,245,248,249,251,252	0.48	-0.291	0.37	-0.479
	4	196,200,207,222,223,224,241,242,243,245, 248,249,251,252	0.44	-0.331	0.40	-0.449
	5	**196,199,200,207,222,223,224,241,242,243, 245,248,251,252**	**0.49**	**-0.281**	**0.43**	**-0.419**

Table S.38. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0807 (PDB ID 4wgh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	200,202,207,222,245,248	0.12	-0.651	0.12	-0.729
	2	199,200,201,207,222,248	0.21	-0.561	0.14	-0.709
	3	200,201,202,203,205,207,208,222,245,247,248	0.13	-0.641	0.19	-0.659
	4	**199,200,201,202,205,207,208,222,245,247,248**	**0.19**	**-0.581**	**0.22**	**-0.629**
	5	200,202,207,208,222,245,248	0.11	-0.661	0.12	-0.729
	6	200,207,208,222,245,248	0.12	-0.651	0.11	-0.739

Table S.39. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0807 (PDB ID 4wgh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	**200,201,213,216,223,248**	**0.12**	**-0.651**	**0.11**	**-0.739**
	2	200,213,216,222,223	-0.042	-0.813	0.045	-0.804
	3	213,223,245,248	0.17	-0.601	0.095	-0.754
	4	213,216,222,223,248	0.05	-0.721	0.069	-0.780
	5	200,201,207,213,216,221,222,223, 248	0.082	-0.689	0.13	-0.719
	6	200,201,207,208,213,216,223,247,248	0.082	-0.689	0.14	-0.709
2	1	200,213,216,221,247,248,251	0.11	-0.661	0.12	-0.729
	2	213,216,221,222	-0.038	-0.809	0.02	-0.829
	3	213,216,245,247,248	0.14	-0.631	0.099	-0.75
	4	200,213,216,221,222,247,248,251	0.093	-0.678	0.13	-0.719
	5	200,213,216,221,222,223,247,248	0.020	-0.751	0.10	-0.749
	6	200,213,216,221,223,247,248	0.028	-0.743	0.094	-0.755

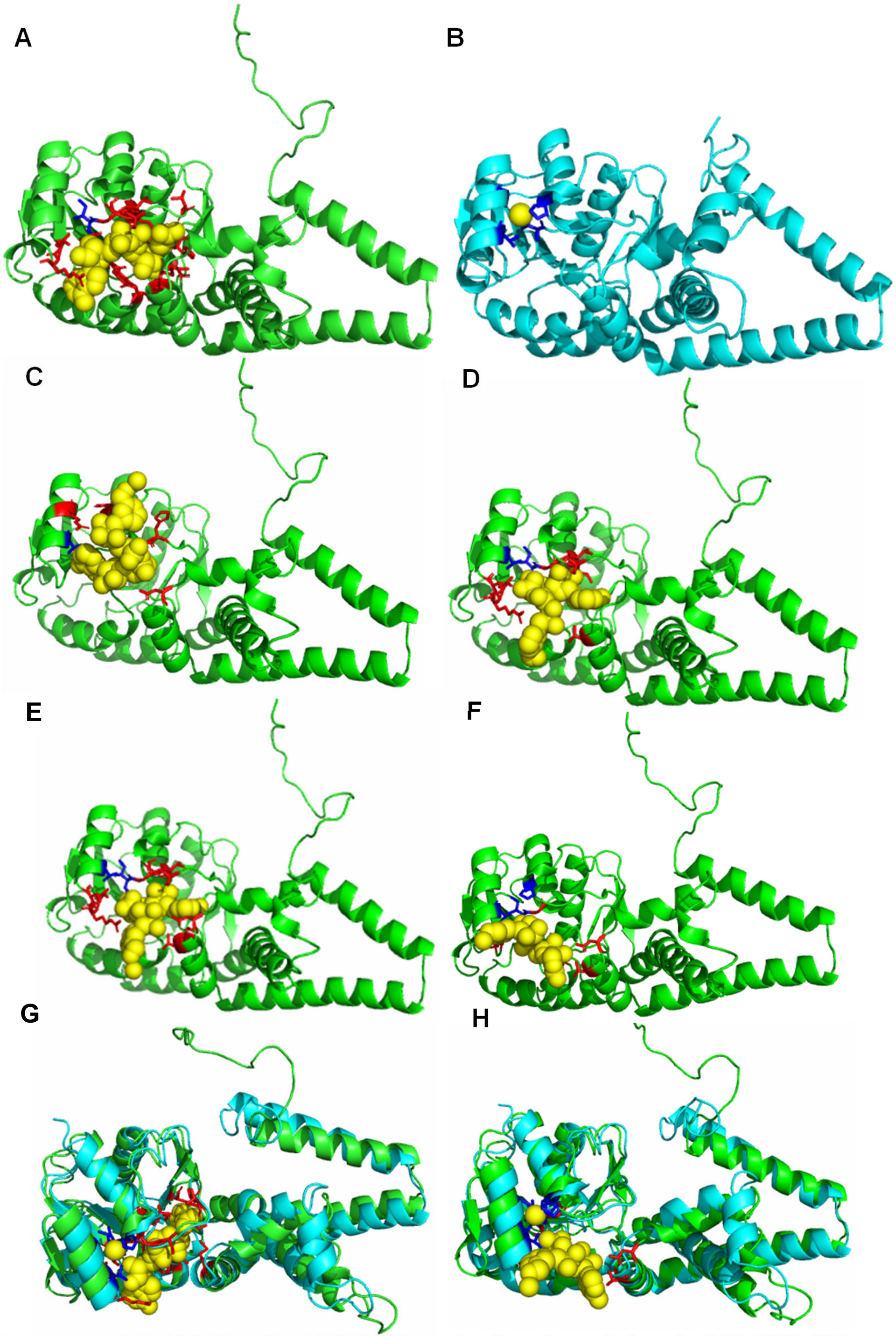


Figure S.78. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for NAD T0813 (PDB ID 4wji) **(A)** Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAD ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0813 (PDB ID 4wji) shown as sticks and coloured blue, the MG ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model from 50% grid box calculation with the predicted structure coloured green and the observed structure coloured cyan. BDT and MCC score of 0.43 and 0.46 respectively

Table S.40. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for NAD T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å with the predicted ligand NAD

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	128,139,144	-0.011	-0.097	0.029	-0.161
	2	128,129,144,227,287,288,289	-0.018	-0.104	0.026	-0.164
	3	128,137,139,140,144,227,287,288,289	-0.020	-0.106	0.035	-0.155
	4	128,137,139,140,144,227,287	-0.018	-0.104	0.043	-0.147
2	1	290,291,293,298,299,300,301	-0.018	-0.104	0.012	-0.178
	2	290,295,297,299,301,303	-0.015	-0.101	0.012	-0.178
	3	290,291,293,299,301,302,305	-0.015	-0.101	0.013	-0.177
	4	290,293,299,301,302	-0.015	-0.101	0.013	-0.177
3	1	40,43,55,57,62,65	-0.017	-0.103	0.14	-0.05
	2	43,54,57,61,62	-0.015	-0.101	0.12	-0.07
	3	34,57,61,62,65	-0.015	-0.101	0.05	-0.14
4	1	39,45,131,132,134,135,232	-0.018	-0.104	0.26	0.07
	2	37,39,42,45,130,131,134,232	0.16	0.074	0.31	0.12
	3	**42,45,130,131,134,135,232**	**0.17**	**0.084**	**0.33**	**0.14**
	4	45,130,131,134,135,232	-0.01	-0.096	0.22	0.03

Table S.41. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for NAD T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site with the predicted ligand NAD

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	**12,13,14,15,16,37,38,39,42,72,131,234**	**0.27**	**0.184**	**0.33**	**0.14**
	2	12,14,15,37,38,72,73,74,232,234,235	-0.02	-0.106	0.16	-0.03
	3	12,13,14,15,37,38,39,42,131,132,232,234,235	0.26	0.174	0.32	0.13
2	1	16,38,72,73,74,77,80,81,234,238	-0.021	-0.107	0.08	-0.11
	2	12,14,15,37,38,72,73,74,77,81,234,238	-0.024	-0.11	0.15	-0.04
	3	14,15,16,38,72,73,74,77,81,234,238	-0.023	-0.109	0.13	-0.06
	4	14,15,16,38,72,74,77,81,234,238	-0.02	-0.106	0.13	-0.06

Table S.42. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for NAD T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site with the predicted ligand NAD

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	38,57,73,74,77,132,232,234,238	-0.02	-0.106	0.15	-0.04
	2	13,14,38,57,73,74,77,234	0.16	0.074	0.23	0.04
	3	38,57,74,77,232,234,238	-0.018	-0.104	0.064	-0.126
2	1	12,13,14,15,16,37,38,42,72,74,131,132,232,234,235	0.24	0.154	0.29	0.1
	2	14,15,37,38,73,74,132,232,234,235	-0.021	-0.107	0.16	-0.03
	3	**12,13,14,15,37,38,39,42,131,132,232,234,235**	**0.26**	**0.174**	**0.32**	**0.13**

Table S.43. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for NAD T0813 (PDB ID 4wji)

Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site with the predicted ligand NAD

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	12,13,14,15,37,72,74,131,132,232,234,235	0.12	0.034	0.26	0.07
	2	14,15,38,73,74,232,234,235	-0.019	-0.105	0.13	-0.06
	3	12,13,14,15,37,39,132,232,234,235	0.14	0.054	0.29	0.1
	4	12,13,14,15,37,131,132,232,234,235	0.14	0.054	0.29	0.1
2	1	12,13,14,37,38,39,42,74,133,232,234	0.44	0.354	0.41	0.22
	2	37,38,39,42,74,234	0.19	0.104	0.28	0.09
	3	**12,13,14,37,39,42,74,133,232,234**	**0.46**	**0.374**	**0.43**	**0.24**
3	1	38,59,73,74,77,84,132,234	-0.019	-0.105	0.11	-0.08
	2	13,14,38,73,74,77,234	0.17	0.084	0.25	0.06
	3	38,59,74,77,234,238	-0.016	-0.102	0.062	-0.128
	4	38,59,74,77,84,234,238	-0.018	-0.104	0.057	-0.133

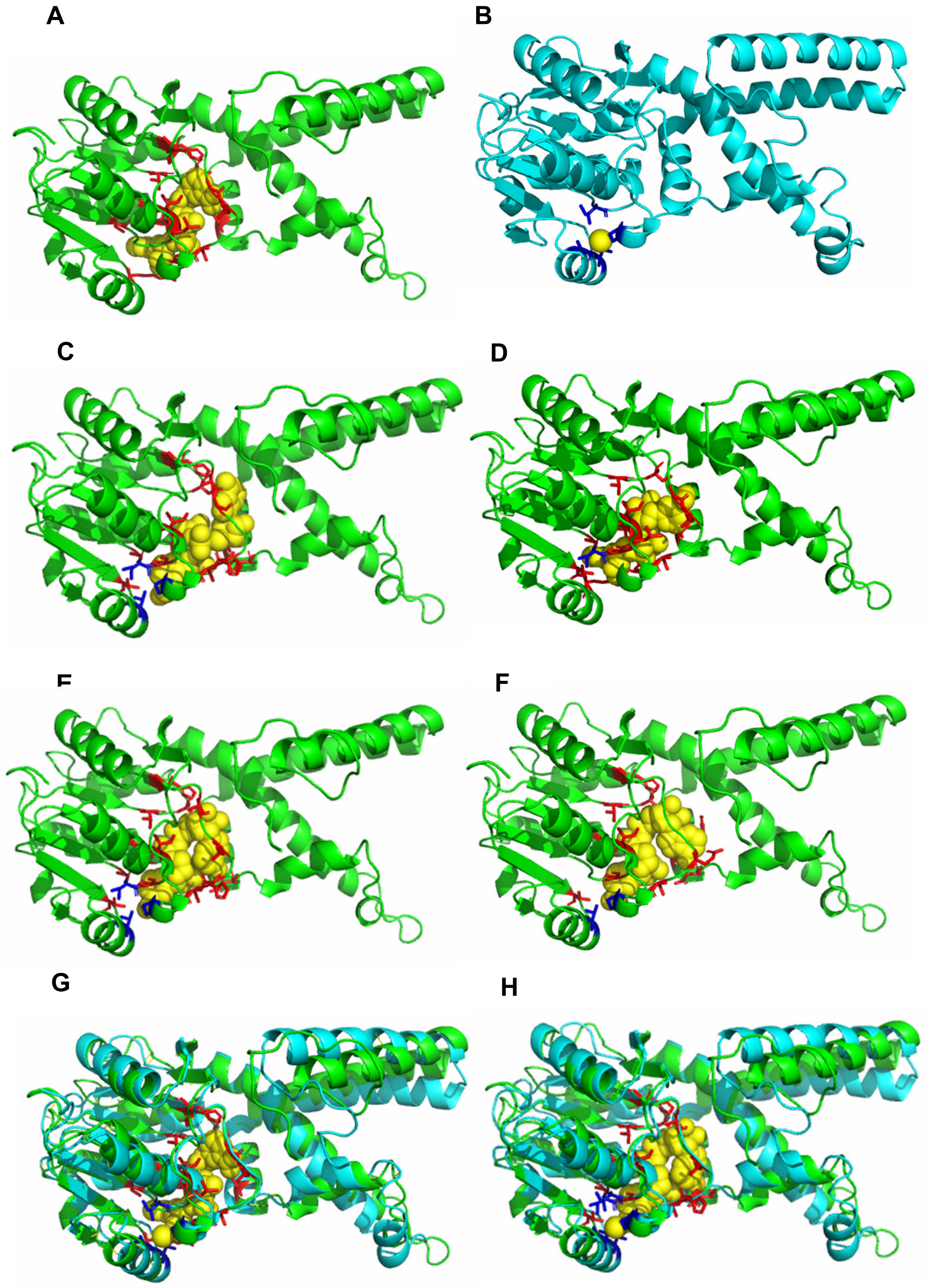


Figure S.79. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for NAI T0813 (PDB ID 4wji) **(A)** Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAI ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0813 (PDB ID 4w66) shown as sticks and coloured blue, the MG ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model complex 2 with 22.5Å grid box calculation with the predicted structure coloured green and the observed structure coloured cyan. BDT and MCC score of 0.28 and 0.34, respectively

Table S.44. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for NAI T0813 (PDB ID 4wji)

Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å with the predicted ligand NAI

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	12,13,14,15,16,37,38,42,72,73,74,98,100,131,235	0.24	0.269	0.27	0.16
2	**12,13,14,16,37,42,74,100,123,126,127,130,131,133,232,235,236**	**0.35**	**0.379**	**0.28**	**0.17**
3	12,14,15,16,72,100,123,128,129,130,132,189,193,232,235,236	-0.027	0.002	0.13	0.02
4	16,100,123,125,126,127,128,129,189,193,232,235,236	-0.025	0.004	0.047	-0.063
5	14,15,16,37,38,74,98,99,100,123,124,128,130,131,132,232,235,236	-0.029	0	0.12	0.01
6	189,192,193,196,224,225,227,228,235,236	-0.021	0.008	0.022	-0.088
7	15,16,72,128,129,130,131,192,193,196,224,228,232,235,236,284	-0.027	0.002	0.069	-0.041
8	12,14,15,16,37,38,72,73,74,81,126,127,128,129,131,132,232,235,236	-0.03	-0.001	0.14	0.03
9	12,13,14,16,37,38,74,100,123,126,127,128,129,131,232,235,236	0.097	0.126	0.17	0.06

Table S.45. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for NAI T0813 (PDB ID 4wji)

Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site with the predicted ligand NAI

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	13,14,15,16,37,38,42,72,73,74,75,98,99,100,128,129,130,131,132,235	0.20	0.229	0.22	0.11
2	11,12,14,15,16,37,38,72,73,74,98,99,100,129,130,131,232,235	-0.029	0	0.14	0.03
3	12,13,14,16,37,38,72,73,74,81,98,100,127,128,130,235,236	0.097	0.126	0.18	0.07
4	11,12,13,14,16,37,38,72,73,74,75,98,99,100,128,129,130,131,232,235	0.086	0.115	0.17	0.06
5	11,12,14,15,16,37,38,72,73,74,128,130,131,232,235	-0.026	0.003	0.16	0.05
6	**12,13,14,15,16,37,38,42,73,74,98,100,127,128,131,132,235**	**0.22**	**0.249**	**0.26**	**0.15**
7	11,13,14,16,37,38,42,73,74,98,99,100,130,131,232,234,235	0.22	0.249	0.22	0.11
8	14,15,16,37,38,42,72,73,74,75,131,232,235	0.12	0.149	0.21	0.1
9	12,14,15,16,37,38,72,73,74,100,123,126,127,128,130,131,132,236	-0.029	0	0.15	0.04

Table S.46. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for NAI T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site with the predicted ligand NAI

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	11,12,13,14,15,16,37,38,72,73,74,123,126,127,128,132, 235,236	0.093	-0.017	0.20	0.09
2	**12,13,14,16,37,42,72,73,98,100,123,126,128,130,131, 133,232,235**	**0.34**	**0.23**	**0.28**	**0.17**
3	12,13,14,15,16,37,38,72,73,74,100,129,130,131,132, 235,236	0.097	-0.013	0.21	0.1
4	12,13,14,16,37,72,74,100,123,124,126,128,132,232, 235,236	0.10	-0.01	0.19	0.08
5	12,13,16,37,38,42,72,73,74,75,98,99,100,128,129, 130,131,132,235	0.21	0.1	0.22	0.11
6	12,13,14,15,16,37,38,42,72,73,74,98,131,232,235	0.24	0.13	0.27	0.16
7	12,14,16,37,38,42,73,74,81,98,127,128,235	0.12	0.01	0.21	0.1
8	12,14,37,38,72,73,74,77,123,126,128,131,232,235	-0.026	-0.136	0.13	0.02
9	12,37,38,72,73,74,100,127,128,130,235,236	-0.024	-0.134	0.11	0

Table S.47. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for NAI T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site with the predicted ligand NAI

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	11,12,13,14,15,16,37,38,72,73,74,100,123,126,127,128, 130,235,236	0.089	0.118	0.18	0.07
2	12,13,14,16,72,73,74,100,123,128,129,131,132, 232,234,235,236	0.097	0.126	0.19	0.08
3	12,14,15,16,37,72,98,100,123,126,127,128,129, 130,131,232,235,236	-0.029	0	0.12	0.01
4	12,13,14,16,42,72,73,74,75,98,100,126,130, 232,235,236	0.23	0.259	0.23	0.12
5	**14,16,37,42,72,73,98,100,123,129,131,133,232, 235,236**	**0.24**	**0.269**	**0.23**	**0.12**
6	12,14,16,37,38,72,73,74,75,98,99,100,128,130, 131,235	-0.027	0.002	0.13	0.02
7	12,13,14,15,16,37,38,72,73,74,127,128,129,130,236	0.11	0.139	0.21	0.1
8	12,14,16,37,38,72,73,74,75,98,99,100,130,131, 232,235	-0.027	0.002	0.13	0.02
9	14,15,16,72,73,74,75,98,99,123,125,126,128, 129,130,131,189,231,235,236	-0.031	-0.002	0.092	-0.018

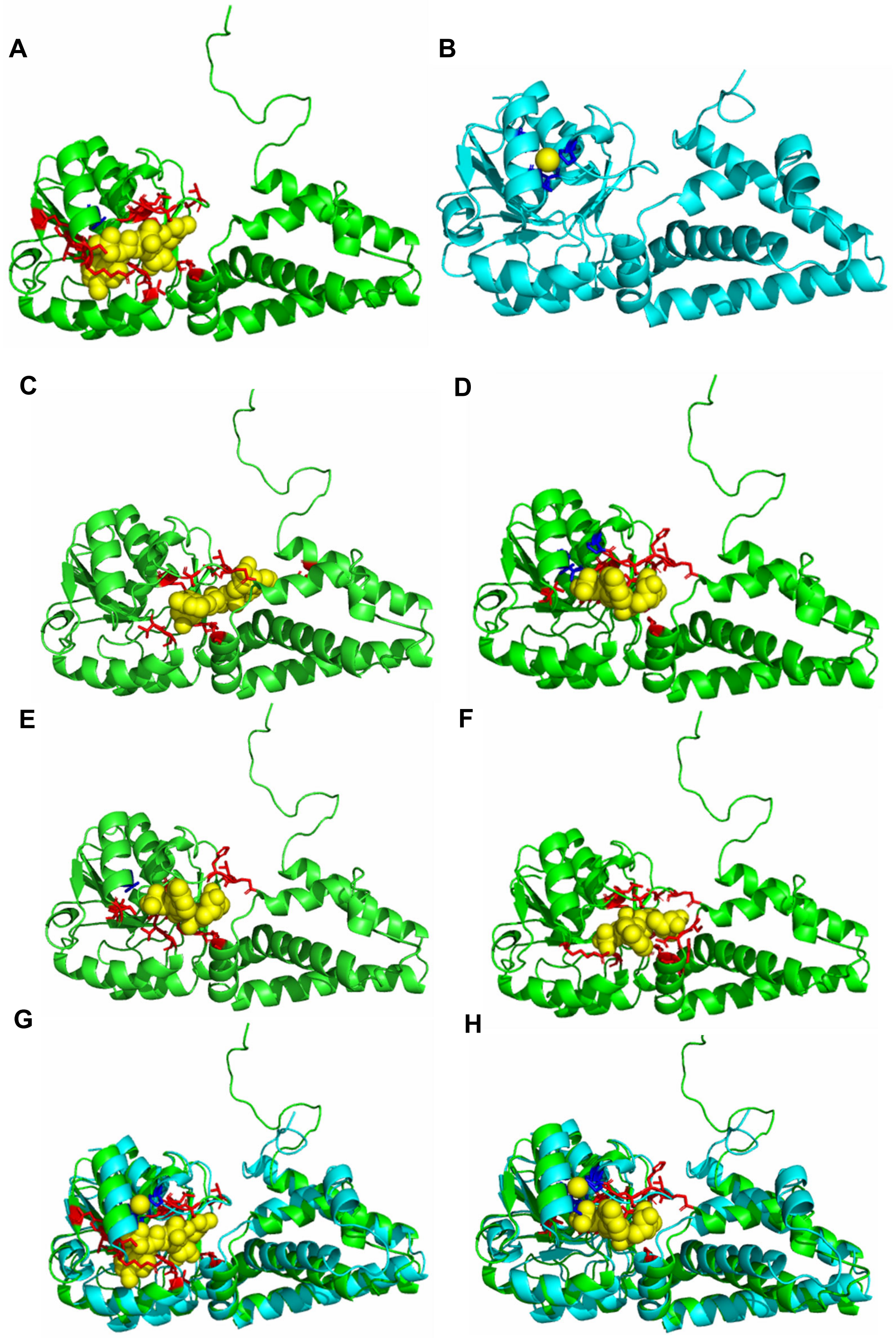


Figure S.80. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for NAP T0813 (PDB ID 4wji) **(A)** Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAP ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0813 (PDB ID 4w66) shown as sticks and coloured blue, the MG ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model complex 2 with 10% grid box calculation with the protein coloured green and the observed structure coloured cyan. BDT and MCC score of 0.27 and 0.34, respectively

Table S.48. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for NAP T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	16,73,74,99,100,123,126,128,129,131,235,284	-0.024	-0.103	**0.068**	**-0.132**
	2	16,73,74,75,99,100,123,126,128,129,131, 189,232,235	-0.026	-0.105	0.066	-0.134
2	1	188,192,254,257,261,280,284,287	-0.019	-0.098	0.009	-0.191
	2	188,192,195,254,257,261,280,284,287,288	-0.021	-0.1	0.009	-0.191
3	1	188,195,225,258,261,284,287,288	-0.019	-0.098	0.010	-0.19
	2	188,189,195,254,257,258,284	-0.018	-0.097	0.010	-0.19
4	1	192,195,224,225,253,276	-0.016	-0.095	0.011	-0.189
	2	**192,224,276,279,280**	**-0.015**	**-0.094**	**0.009**	**-0.191**
5	1	16,123,124,125,126,127,128,129,145,188, 193,235,236	-0.025	-0.104	0.045	-0.155
	2	16,123,124,125,126,127,129,184,188, 189,193,235,236	-0.025	-0.104	0.041	-0.159
6	1	127,187,188,193,227,254,284,287,288	-0.020	-0.099	0.017	-0.183
	2	188,193,254,257,284,287,288	-0.018	-0.097	0.011	-0.189
7	1	126,129,192,193,195,196,225,227,254,284	-0.021	-0.1	0.020	-0.18
	2	126,129,192,193,225,227,284	-0.018	-0.097	0.024	-0.176
8	1	191,192,195,253,254,257,272,276,279	-0.020	-0.099	0.008	-0.192
	2	191,192,253,254,275,276,279,280	-0.019	-0.098	0.008	-0.192
9	1	188,195,254,280,284,287	-0.016	-0.095	0.010	-0.19
	2	188,192,195,254,284,287	-0.016	-0.095	0.011	-0.189

Table S.49. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for NAP T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	**13,14,16,37,42,72,98,123,124,126,127,128,129, 130,131,132,133,235**	**0.34**	**0.261**	**0.27**	0.07
	2	14,16,37,39,42,72,73,74,98,99,100,128,129, 130,132,235	0.10	0.021	0.19	-0.01
2	1	11,12,16,37,38,39,73,74,75,98,99,100,123, 127,131,235	-0.027	-0.106	0.12	-0.08
	2	11,12,16,37,38,39,42,72,73,74,75,98,99,100, 123,127,128,131,132,235	0.086	0.007	0.18	-0.02
3	1	16,38,74,75,98,100,130,131,232,234,235	-0.023	-0.102	0.080	-0.12
	2	16,38,74,75,98,99,100,129,130,232,234,235	-0.023	-0.102	0.068	-0.132

Table S.50. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for NAP T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	12,16,37,72,73,123,124,128,129,235,236	-0.023	-0.102	0.11	-0.09
	2	12,16,37,38,72,73,74,100,123,128,235,236	-0.024	-0.103	0.12	-0.08
2	1	14,16,37,72,73,98,123,128,129,130,131,132,235	-0.025	-0.104	0.14	-0.06
	2	**16,37,39,42,72,73,74,98,99,100,128,129,130,235**	**0.11**	0.031	**0.16**	-0.04
3	1	16,38,74,75,98,130,131,232,234,235	-0.021	-0.1	0.085	-0.115
	2	16,38,74,75,98,100,129,130,131,232,234,235	-0.024	-0.103	0.078	-0.122
4	1	11,12,16,37,38,39,73,74,75,98,99,100,123,127,131,235	-0.027	-0.106	0.12	-0.08
	2	11,12,16,37,38,39,42,72,73,74,75,98,99,100,123,127,128,131,132,235	0.086	0.007	0.18	-0.02

Table S.51. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for NAP T0813 (PDB ID 4wji)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site with the predicted ligand NAP

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	**15,16,38,74,100,123,126,127,129,131,231,232,235,236**	**-0.026**	**-0.226**	**0.08**	-0.12
	2	15,16,38,74,100,123,124,126,128,129,231,232,235,236	-0.026	-0.226	0.07	-0.13

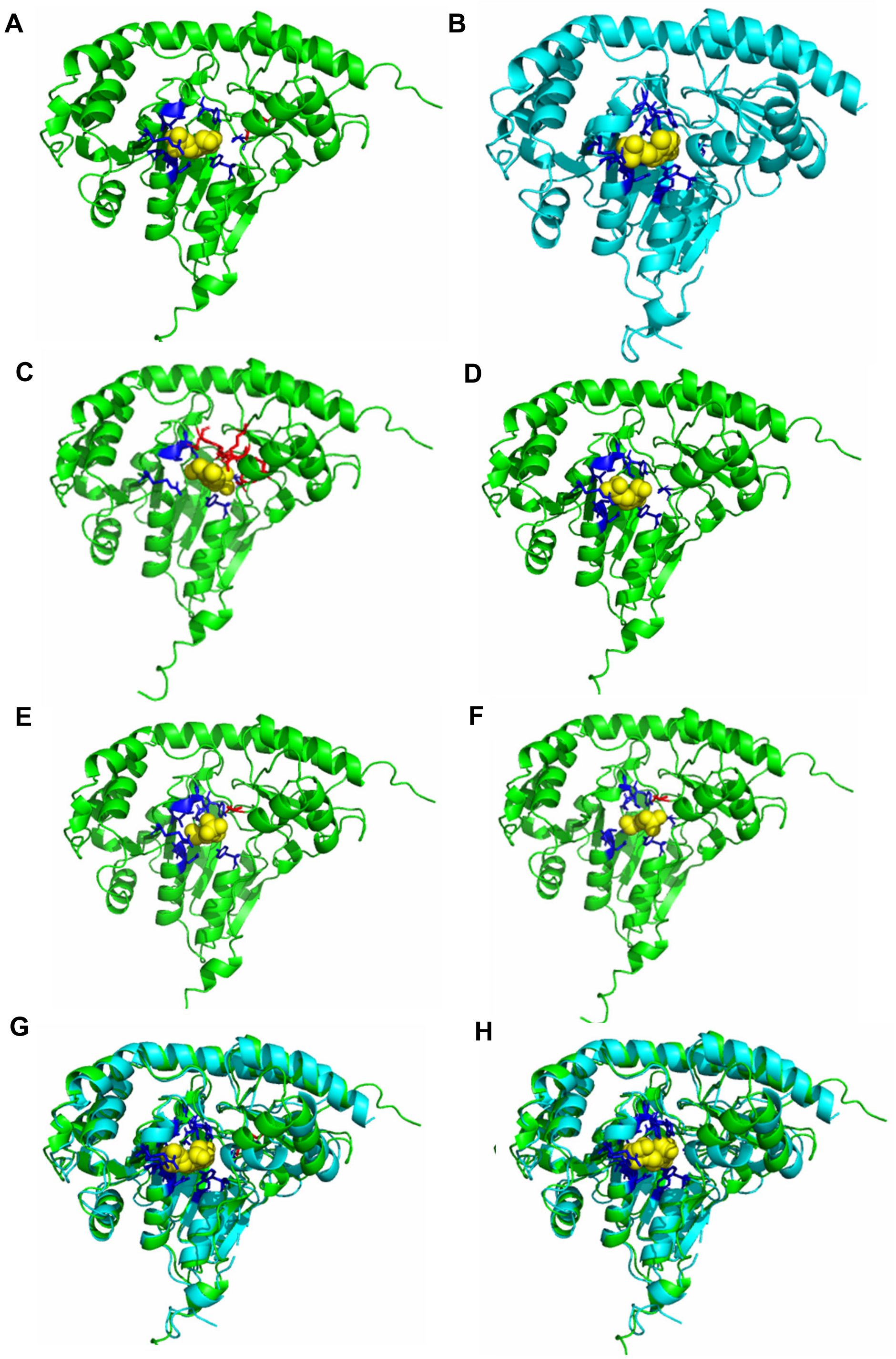


Figure S.81. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0819 (PDB ID 4wbt) **(A)** Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the PLP ligand is shown as sphere and coloured yellow **(B)** The observed ligand binding site residues for T0819 (PDB ID 4wbt) shown as sticks and coloured blue, the PLP ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct predictions are shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D for the best model (complex 8 with 10% grid box calculation with the protein coloured green and the observed structure coloured cyan. BDT and MCC score of 0.80 and 0.87, respectively

Table S.52. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0819 (PDB ID 4wbt)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	39,40,41,42,119,167,225,234,335,347	0.33	-0.547	0.36	-0.493
2	**37,39,40,41,42,119,167,225,226,234,335,347**	**0.38**	**-0.497**	**0.44**	**-0.413**
3	39,40,42,119,225,226,234,335	0.38	-0.497	0.33	-0.523
4	40,42,225,226,231,234,335	0.30	-0.577	0.29	-0.563
5	40,41,42,119,225,226,234,335,347	0.35	-0.527	0.35	-0.503
6	40,42,119,225,226,234,335	0.40	-0.477	0.33	-0.523
7	37,40,41,119,167,226,234,335	0.38	-0.497	0.33	-0.523
8	39,40,42,119,225,226,234,335,347	0.35	-0.527	0.35	-0.503
9	40,42,226,231,234,335	0.21	-0.667	0.21	-0.643

Table S.53. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0819 (PDB ID 4wbt)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	94,95,119,167,194,196,197,226,234	0.83	-0.047	0.69	-0.163
2	40,94,95,119,167,194,196,197,226	0.73	-0.147	0.62	-0.233
3	41,94,95,119,194,196,197,226,234	0.73	-0.147	0.63	-0.223
4	39,40,94,95,119,194,196,197,226,234,335	0.66	-0.217	0.63	-0.223
5	94,95,119,122,163,167,194,196,197,226	0.69	-0.187	0.66	-0.193
6	40,94,95,119,167,196,197,226,234	0.73	-0.147	0.62	-0.233
7	94,95,119,122,163,165,167,194,196,197,226,234	0.71	-0.167	0.75	-0.103
8	**94,95,119,167,194,197,223,225,226,234**	**0.87**	**-0.007**	**0.80**	**-0.053**
9	40,94,119,167,194,196,197,225,226,234	0.78	-0.097	0.70	-0.153

Table S.54. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0819 (PDB ID 4wbt)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	94,95,119,167,194,196,197,226,234	0.83	-0.047	0.69	-0.163
2	40,94,95,119,167,194,196,197,226	0.73	-0.147	0.62	-0.233
3	40,41,94,119,194,196,197,226,234	0.64	-0.237	0.56	-0.293
4	39,40,94,95,119,196,197,226,234,335	0.60	-0.277	0.55	-0.303
5	**40,94,95,119,196,197,223,225,226,234**	**0.78**	**-0.097**	**0.70**	**-0.153**
6	40,41,119,167,194,196,197,226,234	0.64	-0.237	0.56	-0.293
7	39,40,94,95,119,223,225,226,234,335,347	0.57	-0.307	0.56	-0.293
8	41,94,95,119,122,163,165,167,194,196,197,226,234	0.68	-0.197	0.76	-0.093
9	40,94,95,119,167,197,226,234,335	0.64	-0.237	0.55	-0.303

Table S.55. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0819 (PDB ID 4wbt)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	39,40,41,42,119,225,226,234,335	0.35	-0.527	0.35	-0.503
2	39,40,41,119,167,225,226,234,335	0.45	-0.427	0.41	-0.443
3	39,40,41,42,119,167,225,234,335,347	0.33	-0.547	0.36	-0.493
4	42,119,225,226,234,335,347	0.40	-0.477	0.33	-0.523
5	39,40,41,42,119,167,226,234,335,347	0.33	-0.547	0.36	-0.493
6	40,94,95,119,196,197,226	0.62	-0.257	0.47	-0.383
7	**40,94,95,119,167,194,196,197,226**	**0.73**	**-0.147**	**0.62**	**-0.233**
8	40,94,95,119,225,226,234,335	0.58	-0.297	0.47	-0.383
9	94,119,167,197,225,226,234	0.73	-0.147	0.54	-0.313

Table S.56. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for FRU ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å.

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	462,492,525,554,556,557	-0.007	-0.00028	0.027	-0.0025
	2	238,239,240,556	-0.006	0.00072	0.017	-0.0125
2	1	205,238,239,240,556	-0.006	0.00072	0.017	-0.0125
	2	238,492,524,525,557	-0.006	0.00072	0.024	-0.0055
3	1	238,462,492,524	-0.006	0.00072	0.022	-0.0075
	2	153,205,462,465	-0.006	0.00072	**0.076**	**0.0465**
4	1	208,209,232,234,489,490,521	-0.008	-0.00128	0.023	-0.0065
	2	163,164,208,230,231,232	-0.007	-0.00028	0.038	0.0085
5	1	208,209,231,232,234,489,490,521	-0.008	-0.00128	0.022	-0.0075
	2	163,164,208,230,231	-0.007	-0.00028	0.042	0.0125
6	**1**	**210,238,240**	**-0.005**	**0.00172**	0.025	-0.0045
	2	238,465,492,524,525,554,557	-0.008	-0.00128	0.026	-0.0035
7	1	462,492,525,554,556,557	-0.007	-0.00028	0.027	-0.0025
	2	238,239,240,556	-0.006	0.00072	0.017	-0.0125
8	1	205,238,239,240,556	-0.006	0.00072	0.017	-0.0125
	2	238,492,524,525,557	-0.006	0.00072	0.024	-0.0055
9	1	238,462,492,524	-0.006	-0.00028	0.022	-0.0075
	2	*Same as Complex 3, pose 2*	-	-	-	-

Table S.57. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for FRU ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	163,208,232,459,489,490	-0.007	-0.00028	0.030	0.0005
	2	208,209,231,232,235,490	-0.007	-0.00028	0.026	-0.0035
2	1	208,209,231,232,234,235,490	-0.007	-0.00028	0.024	-0.0055
	2	163,208,459,489,490,521	-0.007	-0.00028	0.029	-0.0005
3	**1**	**163,208,489,521**	**-0.006**	**0.00072**	**0.036**	**0.0065**
	2	208,209,230,231,232,235,490,521	-0.008	-0.00128	0.023	-0.0065
4	1	208,209,231,232,234,235,238,490,521	-0.009	-0.00228	0.022	-0.0075
	2	163,208,232,489,490	-0.007	-0.00028	0.033	0.0035
5	1	208,209,231,232,234,235,490,521	-0.008	-0.00128	0.023	-0.0065
	2	163,208,459,489,490,521	-0.007	-0.00028	0.029	-0.0005
6	1	163,208,232,459,489,490	-0.007	-0.00028	0.030	0.0005
	2	208,209,231,232,235,490	-0.007	-0.00028	0.026	-0.0035
7	1	208,209,231,232,234,235,490	-0.007	-0.00028	0.024	-0.0055
	2	163,208,459,489,490,521	-0.007	-0.00028	0.029	-0.0005
8	1	*Same as complex 3 pose 1*	-0.006	0.00072	0.036	0.0065
	2	208,209,230,231,232,235,490,521	-0.008	-0.00128	0.023	-0.0065
9	1	208,209,231,232,234,235,238,490,521	-0.009	-0.00228	0.022	-0.0075
	2	163,208,232,489,490	-0.007	-0.00028	0.033	0.0035

Table S.58. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for FRU ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	163,232,489,521	-0.006	0.00072	0.031	0.0015
	2	208,209,230,231,232,235,490,521	-0.008	-0.00128	0.023	-0.0065
2	1	163,208,232,459,489	-0.007	-0.00028	0.032	0.0025
	2	208,209,230,231,232,235,490,521	-0.008	-0.00128	0.023	-0.0065
3	1	163,232,489,521	-0.006	0.00072	0.031	0.0015
	2	209,231,232,234,235,237,490,521	-0.008	-0.00128	0.020	-0.0095
4	1	232,489,490,521	-0.006	0.00072	0.017	-0.0125
	2	163,164,208,230,231,232	-0.007	-0.00028	0.038	0.0085
5	1	163,208,232,489	-0.006	0.00072	0.037	0.0075
	2	208,231,232,489,490,521	-0.007	-0.00028	0.021	-0.0085
6	1	208,209,231,232,234,235,490,521	-0.008	-0.00128	0.023	-0.0065
	2	163,208,459,487,489,490,521	-0.008	-0.00128	0.026	-0.0035
7	1	208,209,231,232,234,235,490,521	-0.008	-0.00128	0.023	-0.0065
	2	**163,208,489**	**-0.005**	**0.00172**	**0.043**	**0.0135**
8	1	*Same as complex 7, pose 2*	-	-	-	-
	2	208,209,231,232,233,234,235,490	-0.008	-0.00128	0.023	-0.0065
9	1	163,232,489,521	-0.006	0.00072	0.031	0.0015
	2	208,209,230,231,232,235,490,521	-0.008	-0.00128	0.023	-0.0065

Table S.59. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for FRU ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	**163,164,208,230,231,232**	-0.007	-0.00028	0.038	0.0085
	2	163,208,232,489,490,521	-0.007	-0.00028	0.030	0.0005
2	1	163,208,459,487,489	-0.007	-0.00028	0.030	0.0005
	2	208,209,232,234,235,489,490,521	-0.008	-0.00128	0.022	-0.0075
3	1	163,232,489,521	-0.006	0.00072	0.031	0.0015
	2	208,231,232,489,490,521	-0.007	-0.00028	0.021	-0.0085
4	1	163,208,230,231	-0.006	0.00072	0.039	0.0095
	2	208,232,489,490	-0.006	0.00072	0.023	-0.0065
5	1	163,208,231,232	-0.006	0.00072	**0.039**	**0.0095**
	2	208,209,232,234,235,237,490,521	-0.008	-0.00128	0.022	-0.0075
6	1	163,208,459,489	-0.006	0.00072	0.035	0.0055
	2	208,209,231,232,235,489,490,521	-0.008	-0.00128	0.023	-0.0065
7	1	163,208,231,232,459,489	-0.007	-0.00028	0.030	0.0005
	2	208,209,232,234,235,237,490,521	-0.008	-0.00128	0.022	-0.0075
8	1	208,209,230,231,232,234,489,490,521	-0.008	-0.00128	0.02	-0.0095
	2	232	-0.003	0.00372	0.007	-0.0225
9	**1**	**208,232,489**	**-0.005**	**0.00172**	0.024	-0.0055
	2	208,209,232,234,235,237,490,521	-0.008	-0.00128	0.022	-0.0075

Table S.60. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for MAV ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å.

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	468,533,535,569	-0.006	0.00292	0.015	-0.0063
	2	334,335,468,500	-0.006	0.00292	0.023	0.0017
	3	332,333,334,336,468,469,474,500	-0.008	0.00092	**0.026**	**0.0047**
2	1	468,533,535,569	-0.006	0.00292	0.015	-0.0063
	2	334,335,468,500	-0.006	0.00292	0.023	0.0017
	3	334,335,336,340,426,429,474,476,500	-0.009	-8E-05	0.015	-0.0063
3	1	535	-0.003	0.00592	0.003	-0.0183
	2	*Same as Complex 1, pose 2*	-0.006	0.00292	0.023	0.0017
	3	*Same as Complex 1, pose 3*	-		-	-
4	1	*Same as Complex 3, pose 1*	-		-	-
	2	*Same as Complex 1, pose 3*	-		-	-
	3	*Same as Complex 2, pose 3*	-		-	-
5	1	335,500,535	-0.005	0.00392	0.013	-0.0083
	2	*Same as Complex 1, pose 2*	-		-	-
	3	332,333,336,340,468,469,474,500	-0.008	0.00092	0.025	0.0037
6	1	468,474,497,499,500,533	-0.007	0.00192	0.013	-0.0083
	2	334,335,468,500,535	-0.007	0.00192	0.020	-0.0013
	3	334,468,533,569	-0.006	0.00292	0.019	-0.0023
7	**1**	**500**	**-0.003**	**0.00592**	0.003	-0.0183
	2	334,335,468,500,535	-0.007	0.00192	0.020	-0.0013
	3	332,333,334,335,340,468,469,474,500	-0.009	-8.0E-05	0.024	0.0027
8	1	*Same as Complex 6, pose 1*	-		-	-
	2	*Same as Complex 6, pose 2*	-		-	-
	3	*Same as Complex 6, pose 3*	-		-	-
9	1	334,335,468,533	-0.006	0.00292	0.023	0.0017
	2	334,468,499,500,535	-0.007	0.00192	0.017	-20%
	3	*Same as Complex 1, pose 3*	-	0.00292	-	-

Table S.61. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for MAV ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	332,340,399,423,426,428,469,474	-0.008	0.00092	0.017	-0.0043
	2	334,335,336,340,468,469,474,500	-0.008	0.00092	0.021	-0.0003
	3	334,335,468,500,533	-0.007	0.00192	0.020	-0.0013
2	1	334,468,533,535	-0.006	0.00292	0.019	-0.0023
	2	334,468,500	-0.005	0.00392	0.023	0.0017
	3	332,333,334,336,340,468,469,474,500	-0.009	-8E-05	**0.025**	**0.0037**
3	1	334,468,474,497,500	-0.007	0.00192	0.018	-0.0033
	2	334,335,468,535	-0.006	0.00292	0.023	0.0017
	3	**468,569**	**-0.004**	**0.00492**	0.014	-0.0073
4	1	468,469,474,497,500	-0.007	0.00192	0.018	-0.0033
	2	334,335,468,500,535	-0.007	0.00192	0.020	-0.0013
	3	334,468,533,569	-0.006	0.00292	0.019	-0.0023
5	1	468,533,535	-0.005	0.00392	0.017	-0.0043
	2	334,474,476,500	-0.006	0.00292	0.013	-0.0083
	3	329,332,336,340,399,423,426,428	-0.008	0.00092	0.020	-0.0013

6	1	*Same as Complex 5, pose 1*	-	-	-	-
	2	334,468,500	-0.005	0.00392	0.023	0.0017
	3	334,335,340,468,469,474,500	-0.008	0.00092	0.020	-0.0013
7	1	332,340,423,426,428,469,474	-0.008	0.00092	0.017	-0.0043
	2	335,468,469,474,500	-0.007	0.00192	0.020	-0.0013
	3	*Same as Complex 4, pose 2*	-	-	-	-
8	1	340,426,429,474,500	-0.007	0.01592	0.011	-0.0103
	2	334,335,468,474,500	-0.007	0.00192	0.020	-0.0013
	3	335,468,500,533,535	-0.007	0.00192	0.016	-0.0053
9	1	468,500,533,535	-0.006	0.00292	0.015	-0.0063
	2	468,471,474,497,499,500,533	-0.008	0.00092	0.014	-0.0073
	3	332,333,334,335,336,340,468,469,474,500	-0.009	-8E-05	0.024	0.0027

Table S.62. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for MAV ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	334,468,499,500,535	-0.007	0.00192	0.017	-0.0043
	2	334,335,340,468,469,474,500	-0.008	0.00092	0.020	-0.0013
	3	329,340,374,399,402,423,426,428,429,474	-0.009	-8E-05	0.014	-0.0073
2	1	329,332,340,374,399,402,423,426,428,474	-0.009	-8E-05	0.016	-0.0053
	2	427,474,476,500	-0.006	0.00292	0.009	-0.0123
	3	*Same as Complex 1, pose 1*	-	-	-	-
3	1	334,468,500,535	-0.006	0.00292	0.019	-0.0023
	2	334,335,468,474,500	-0.007	0.00192	0.020	-0.0013
	3	329,332,340,374,399,423,426,428,429,474	-0.009	-8E-05	0.016	-0.0053
4	1	468,474,497,499,500,533	-0.007	0.00192	0.013	-0.0083
	2	334,335,468,500,535	-0.007	0.00192	0.020	-0.0013
	3	334,468,569	-0.005	0.00392	0.023	0.0017
5	1	*Same as Complex 4, pose 2*	-	-	-	-
	2	334,335,336,340,468,474,500	-0.008	0.00092	**0.021**	**-0.0003**
	3	332,340,423,426,427,428,429,474	-0.008	0.00092	0.014	-0.0073
6	**1**	**468,533,535**	**-0.005**	**0.00392**	0.017	-0.0043
	2	334,474,476,500	-0.006	0.00292	0.013	-0.0083
	3	329,332,340,399,423,426,428,429	-0.008	0.00092	0.017	-0.0043
7	1	334,468,474,497,499,500	-0.007	0.00192	0.016	-0.0053
	2	334,335,500,535	-0.006	0.00292	0.016	-0.0053
	3	*Same as Complex 4, pose 3*	-	-	-	-
8	1	334,468,499,500,533,535	-0.007	0.00192	0.015	-0.0063
	2	334,468,535,569	-0.006	0.00292	0.019	-0.0023
	3	468,529,533,567,569	-0.007	0.00192	0.014	-0.0073
9	1	468,527,533,567,569	-0.007	0.00192	0.015	-0.0063
	2	*Same as Complex 4, pose 3*	-	-	-	-
	3	*Same as Complex 8, pose 1*	-	-	-	-

Table S.63. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for MAV ligand for T0912 (PDB ID 5mqp)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	334,335,468,500,535	-0.007	0.00192	0.020	-0.0013
	2	334,335,468,474,500	-0.007	0.00192	0.020	-0.0013
	3	329,332,340,399,423,426,428,429,474	-0.009	-8E-05	0.016	-0.0053
2	1	468,533,535,569	-0.006	0.00292	0.015	-0.0063
	2	334,335,468,499,500	-0.007	0.00192	0.020	-0.0013
	3	332,333,334,336,340,468,469,474,500	-0.009	-8E-05	0.025	0.0037
3	1	468,533,535	-0.005	0.00392	0.017	-0.0043
	2	334,335,468,500	-0.006	0.00292	0.023	0.0017
	3	334,335,336,340,426,429,474,476,500	-0.009	-8E-05	0.015	-0.0063
4	1	329,332,340,374,399,402,423,426,428,474	-0.009	-8E-05	0.017	-0.0043
	2	427,474,476,500	-0.006	0.00292	0.009	-0.0123
	3	334,335,468,499,500,535	-0.007	0.00192	0.018	-0.0033
5	1	332,335,340,423,429,469,474,500	-0.008	0.00092	0.018	-0.0033
	2	*Same as Complex 3, pose 2*	-	-	-	-
	3	468,500,533,535	-0.006	0.00292	0.015	-0.0063
6	1	329,332,340,426,428,474	-0.007	0.00192	0.017	-0.0043
	2	468,474,475,476,499,500	-0.007	0.00192	0.013	-0.0083
	3	468,499,500,533,535	-0.007	0.00192	0.014	-0.0073
7	1	468,474,497,499,500,533	-0.007	0.00192	0.014	-0.0073
	2	334,335,468,500,535	-0.007	-6.99108	0.020	-0.0013
	3	**334,468,569**	**-0.005**	**0.00392**	**0.023**	**0.0017**
8	1	*Same as Complex 5, pose 3*	-	-		-
	2	*Same as Complex 1, pose 2*	-	-		-
	3	332,340,423,426,428,469,474,500	-0.008	0.00092	0.016	-0.0053
9	1	340,426,427,474,476,500	-0.007	0.00192	0.0100	-0.0113
	2	*Same as Complex 3, pose 2*	-	-		-
	3	*Same as Complex 5, pose 3*	-	-		-

Table S.64. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0913
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	**99,103,171,266**	**-0.02**	**0.0167**	**0.009**	**-0.001**
	2	98,99,103,171,222,263,264,266	-0.03	0.0067	0.020	-0.071
	3	98,99,103,171,263,266	-0.03	0.0067	0.014	-0.077
2	1	220,221,224,226,250,252,253	-0.03	0.0067	0.02	-0.071
	2	220,221,224,225,252,253	-0.03	0.0067	0.02	-0.071
	3	220,221,224,225,226,250,253,262	-0.03	0.0067	0.02	-0.071
3	1	96,103,171,222,263,264,266	-0.03	0.0067	0.02	-0.071
	2	96,103,171,222,262,263,264,266	-0.03	0.0067	0.02	-0.071
	3	96,103,171,222,263,264,266	-0.03	0.0067	0.02	-0.071
4	1	98,99,103,171,263,264,266	-0.03	0.0067	0.02	-0.071
	2	98,103,171,263,264	-0.02	0.0167	0.01	-0.081
	3	98,99,103,171,263,264,266	-0.03	0.0067	0.02	-0.071
5	1	98,99,103,263	-0.02	0.0167	0.01	-0.081
	2	91,96,98,99,103,104,263	-0.03	0.0067	0.01	-0.081
	3	96,98,99,102,103,171,263	-0.03	0.0067	0.01	-0.081
6	1	220,221,224,226,250,251,253	-0.03	0.0067	0.02	-0.071
	2	220,224,226,250,253	-0.02	0.0167	0.02	-0.071
	3	220,221,224,226,250,251,253	-0.03	0.0067	0.02	-0.071
7	1	95,96,99,103,171,222,262	-0.03	0.0067	0.01	-0.081
	2	96,99,103,171,222,263	-0.03	0.0067	0.01	-0.081
	3	95,96,99,103,171,222	-0.03	0.0067	0.01	-0.081
8	1	96,98,99,103,104	-0.02	0.0167	0.01	-0.081
	2	96,98,99,103	-0.02	0.0167	0.01	-0.081
	3	98,99,102,103,104	-0.02	0.0167	0.01	-0.081
9	1	96,98,99,103,104,171,263	-0.03	0.0067	0.01	-0.081
	2	91,98,99,103,263	-0.02	0.0167	0.01	-0.081
	3	96,98,99,103,171,263	-0.03	0.0067	0.01	-0.081

Table S.65. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0913
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	100,103,156,171,266,267,270,315,318,359,360,364	-0.04	-0.0033	0.10	0.009
	2	**100,103,156,171,266,267,270,318,359,360,361,364**	**-0.04**	**-0.0033**	**0.11**	**0.019**
	3	100,103,156,171,266,267,270,318,359,360,364	-0.04	-0.0033	0.10	0.009

Table S.66. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0913
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	1	100,103,156,171,266,267,270,315,318,359, 360,361,364	-0.04	-0.0033	0.12	0.029
	2	**100,103,156,171,266,267,269,270,315,318, 359,360,361,364**	**-0.04**	**-0.0033**	**0.13**	**0.039**
	3	100,103,171,266,267,270,318,359,360,361,364	-0.04	-0.0033	0.11	0.019

Table S.67. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0913
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Pose	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**1**	100,103,171,266,267,270,315,318,359,360,364	**-0.04**	**-0.0033**	**0.10**	**0.009**
	2	100,103,266,267,270,315,318,359,360,364	-0.03	0.0067	0.09	-0.001
	2	100,103,171,266,267,270,359,360,361,364	-0.03	0.0067	0.09	-0.001

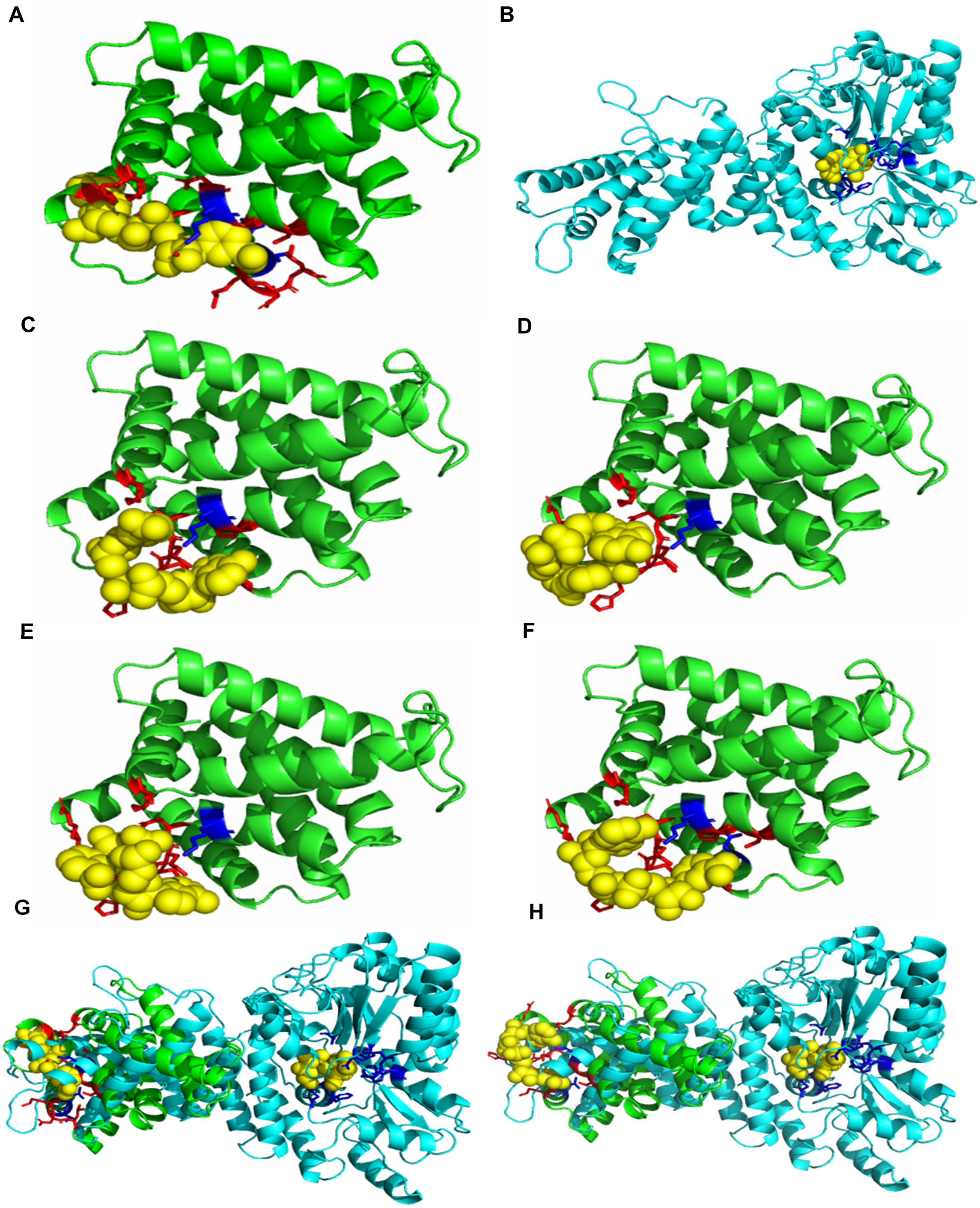


Figure S.82. Comparison of FunFOLD3 and FunFOLD3-D ligand-binding site predictions for T0916 (PDB ID 5tj4)
(A) Predicted ligand-binding site residues shown as sticks with incorrect predictions shown in red, the NAD ligand is shown as sphere and coloured yellow. BDT score of 0.370 and MCC score of 0.263 was achieved. **(B)** The observed ligand binding site residues for T0916 (PDB ID 5tj4) shown as sticks and coloured blue, the GLC(2) ligand is shown as sphere and coloured yellow **(C)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 22.5Å. Correct prediction is shown as sticks and coloured blue and incorrect predictions are shown as sticks and coloured red **(D)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 10% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(E)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 20% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(F)** Predicted ligand-binding site residues following docking with AutoDock Vina and using 50% grid box calculation. Incorrect predictions are shown as sticks and coloured red **(G)** Comparison of the ligand binding site for predictions made by FunFOLD3 with the protein coloured green and the observed structure coloured cyan **(H)** Comparison of the ligand binding site for predictions made by FunFOLD3-D with the predicted structure coloured green and the observed structure coloured cyan. BDT and MCC score of 0.32 and 0.16, respectively was achieved for 50% box calculation for GLC ligand (2)

Table S.68. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0916 (PDB ID 5tj4)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	MCC	Score change	BDT	Score change	BDT	Score change
		GLC (1)		GLC (2)		GLC (1)		GLC (2)	
1	**16,57,58,59,60,63,64,107**	**0.093**	**-0.069**	**0.081**	**-0.182**	**0.25**	**-0.024**	**0.30**	**-0.07**
2	31,34,35,78,79,82,83,86,90,141,142,143	-0.56	-0.722	-0.06	-0.323	0.052	-0.222	0.04	-0.33
3	28,31,32,79,82,83,86,90,142	-0.05	-0.212	0.059	-0.204	-0.05	-0.324	0.047	-0.323
4	28,31,32,35,82,83,86,90	-0.047	-0.209	-0.05	-0.313	0.06	-0.214	0.049	-0.321
5	15,16,18,23,58,59,64,68,71,72	-0.053	-0.215	0.06	-0.203	0.14	-0.134	0.27	-0.1
6	28,31,32,78,79,82,83,86	-0.047	-0.209	0.059	-0.204	-0.051	-0.325	0.043	-0.327
7	16,54,56,57,58,59,60,64	-0.047	-0.209	-0.051	-0.314	0.13	-0.144	0.17	-0.2
8	28,31,32,78,79,80,82,83,86,142,143	-0.056	-0.218	-0.060	-0.323	0.069	-0.205	0.042	-0.328
9	28,31,34,35,78,79,82,83,86,87,90	-0.056	-0.218	-0.060	-0.323	0.061	-0.213	0.046	-0.324

Table S.69. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0916 (PDB ID 5tj4)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	MCC	Score change	BDT	Score change	BDT	Score change
		GLC (1)		GLC (2)		GLC (1)		GLC (2)	
1	53,54,55,56,57,58,60	-0.044	-0.206	-0.047	-0.31	0.068	-0.206	0.065	-0.305
2	**54,57,58,59,60,63,107**	**0.11**	**-0.052**	**0.11**	**-0.153**	**0.23**	**-0.044**	**0.20**	**-0.17**
3	54,56,57,59,60,107	-0.041	-0.203	-0.044	-0.307	0.085	-0.189	0.076	-0.294
4	54,55,56,57,59	-0.037	-0.199	-0.040	-0.303	0.046	-0.228	0.045	-0.325
5	54,56,57,58,59,60 63	0.11	-0.052	0.092	-0.171	0.23	-0.044	0.20	-0.17
6	54,55,56,57,58,59,60,107	-0.047	-0.209	0.088	-0.175	-0.051	-0.325	0.094	-0.276
7	53,54,55,57,58,59,60	-0.044	-0.206	-0.047	-0.31	0.086	-0.188	0.078	-0.292
8	54,56,57,58,59,60	-0.041	-0.203	-0.044	-0.307	0.083	-0.191	0.077	-0.293
9	54,56,57,58,60	-0.037	-0.199	-0.040	-0.303	0.060	-0.214	0.056	-0.314

Table S.70. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0916 (PDB ID 5tj4)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	MCC	Score change	BDT	Score change	BDT	Score change
		GLC (1)		GLC (2)		GLC (1)		GLC (2)	
1	50,51,54,57,58,59,60,107	-0.047	-0.209	-0.051	-0.314	0.094	-0.18	0.095	-0.275
2	**54,56,57,58,59,60,63,107**	**0.093**	**-0.069**	**0.081**	**-0.182**	**0.21**	**-0.064**	**0.21**	**-0.16**
3	54,55,56,57,58,60	-0.041	-0.203	-0.044	-0.307	0.064	-0.21	0.061	-0.309
4	53,54,55,56,57,58,60	-0.044	-0.206	-0.047	-0.31	0.068	-0.206	0.065	-0.305
5	54,55,56,57,59,60	-0.041	-0.203	-0.044	-0.307	0.076	-0.198	0.068	-0.302
6	54,55,56,57,58,59,60,107	-0.047	-0.209	-0.051	-0.314	0.088	-0.186	0.094	-0.276
7	53,54,55,56,57,59,60,107	-0.047	-0.209	-0.051	-0.314	0.082	-0.192	0.084	-0.286
8	54,56,57,58,59,60	-0.041	-0.203	-0.044	-0.307	0.083	-0.191	0.077	-0.293
9	53,54,56,57,58,59,60	-0.044	-0.206	-0.047	-0.31	0.087	-0.187	0.081	-0.289

Table S.71. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0916 (PDB ID 5tj4)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	MCC	Score change	BDT	Score change	BDT	Score change
		GLC (1)		GLC (2)		GLC (1)		GLC (2)	
1	50,54,56,57,58,59,60,63,107	0.082	-0.08	0.070	-0.193	0.19	-0.084	0.20	-0.084
2	54,56,57,58,59,60	-0.041	-0.203	0.083	-0.18	-0.044	-0.318	0.077	-0.318
3	50,54,56,57,58,59,60,107	-0.047	-0.209	-0.051	-0.314	0.091	-0.183	0.096	-0.183
4	54,57,58,59,60,107	-0.041	-0.203	-0.044	-0.307	0.091	-0.183	0.082	-0.183
5	51,54,57,59,60,107	-0.041	-0.203	-0.044	-0.307	0.088	-0.186	0.075	-0.186
6	50,51,52,53,54,55,56,57,58,59,60,107	-0.06	-0.222	-0.06	-0.323	0.074	-0.2	0.077	-0.2
7	**15,16,54,57,58,59,60,63,64,68,107**	**0.064**	**-0.098**	**0.16**	**-0.103**	**0.22**	**-0.054**	**0.32**	**-0.054**
8	16,50,54,56,57,58,59,60	-0.047	-0.209	-0.051	-0.314	0.090	-0.184	0.13	-0.184
9	54,56,57,58,59,60,107	-0.044	-0.206	-0.047	-0.31	-0.047	-0.321	0.087	-0.321

Table S.72. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0953s2 (PDB ID 6f45)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	153,154,155,167,168,207,208,226	-0.016	-0.136	0.031	-0.079
2	153,154,155,207,208,209,210	-0.015	-0.135	0.022	-0.088
3	153,154,155,167,168,207,208,209,210,226	-0.018	-0.138	0.029	-0.081
4	**117,118,119,156,166,167,207**	**-0.015**	**-0.135**	**0.054**	**-0.056**
5	153,154,155,167,168,207,226	-0.015	-0.135	0.033	-0.077
6	117,118,119,154,155,156,167,207	-0.016	-0.136	0.032	-0.078
7	118,119,154,155,156,167,207	-0.015	-0.135	0.033	-0.077
8	116,153,154,155,167,207,226	-0.015	-0.135	0.030	-0.08
9	*Same as complex 6*	-	-	-	-

Table S.73. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0953s2 (PDB ID 6f45)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	118,119,156,165,166,167	-0.014	-0.134	0.12	0.01
2	119,165,166,167	-0.012	-0.132	0.16	0.05
3	119,155,156,165,166,167	-0.014	-0.134	0.12	0.01
4	**165,166,167**	**-0.010**	**-0.13**	**0.21**	**0.1**

Table S.74. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0953s2 (PDB ID 6f45)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	119,156,165,167	-0.011	-0.131	0.13	0.02
2	117,118,119,156,166,167	-0.014	-0.134	0.058	-0.052
3	119,120,156,165,166,167	-0.014	-0.134	0.12	0.01
4	119,165,166,167	-0.012	-0.132	0.16	0.05
5	119,120,165,166,167	-0.013	-0.133	0.14	0.03
6	117,118,119,156,165,166,167	-0.015	-0.135	0.10	-0.01
7	118,119,120,165,166,167	-0.014	-0.134	0.12	0.01
8	**165,166,167**	**-0.010**	**-0.13**	**0.21**	**0.1**

Table S.75. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0953s2 (PDB ID 6f45)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	117,118,119,120,155,156,165,166,167	-0.017	-0.137	0.09	-0.02
2	117,118,119,156,166,167,207	-0.015	-0.135	0.05	-0.06
3	118,119,156,165,166,167	-0.014	-0.134	0.12	0.01
4	**119,165,166,167**	**-0.012**	**-0.132**	**0.16**	**0.05**
5	118,119,120,165,166,167	-0.014	-0.134	0.12	0.01
6	*Same as complex 5*	-	-	-	-
7	118,119,155,156,167,207	-0.014	-0.134	0.036	-0.07
8	117,118,119,155,156,166,167	-0.015	-0.135	0.053	-0.06
9	119,120,156,165,166,167	-0.014	-0.134	0.12	0.01

Table S.76. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T0954 (PDB ID 6cvz)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	25,27,43,44,45,75,76,77	-0.019	-0.004	0.022	-0.006
2	Same as complex 1	-	-	-	-
3	**77,119,187,213,214,231,275**	**-0.018**	**-0.003**	**0.027**	**-0.001**
4	77,119,187,213,231,273,275	-0.018	-0.003	0.026	-0.002
5	119,187,213,231,275	-0.015	0	0.025	-0.003
6	187,213,231,275	-0.013	0.002	0.016	-0.012
7	25,27,43,75,76,77,274,323,340	-0.020	-0.005	0.022	-0.006
8	119,187,213,231,273,275	-0.016	-0.001	0.023	-0.005
9	25,27,43,75,76,77	-0.016	-0.001	0.025	-0.003

Table S.77. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T0954 (PDB ID 6cvz)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	27,77,119,231,273,274,275	-0.018	-0.003	0.025	-0.003
2	Same as complex 1	-	-	-	-
3	**77,119,273,274,275**	**-0.015**	**0**	**0.028**	**0**
4	27,77,119,273,274,275	-0.016	-0.001	0.027	-0.001
5	77,119,187,231,273,274,275	-0.018	-0.003	0.026	-0.002
6	Same as complex 4	-	-	-	-
7	Same as complex 4	-	-	-	-
8	27,77,274,275	-0.013	0.002	0.020	-0.008
9	27,77,273,274,275	-0.015	0	0.023	-0.005

Table S.78. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T0954 (PDB ID 6cvz)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	27,119,187,231,274,275	-0.016	-0.001	0.025	-0.003
2	27,77,119,231,274,275	-0.016	-0.001	0.027	-0.001
3	**77,119,187,273,274,275**	**-0.016**	**-0.001**	**0.028**	**0**
4	25,27,43,77,273,274,275	-0.018	-0.003	0.021	-0.007
5	77,119,187,231,273,274,275	0.018	0.033	0.026	-0.002
6	27,43,77,273,274,275	-0.016	-0.001	0.022	-0.006
7	25,27,77,274,275,323	-0.016	-0.001	0.023	-0.005
8	25,27,77,94	-0.013	0.002	0.022	-0.006
9	27,77,274,275	-0.013	0.002	0.020	-0.008

Table S.79. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T0954 (PDB ID 6cvz)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	25,27,43,76,77,274,323,340	-0.019	-0.004	0.022	-0.006
2	25,27,43,75,77,274,323,340	-0.019	-0.004	0.021	-0.007
3	25,27,43,75,77,94	-0.016	-0.001	0.025	-0.003
4	25,27,43,75,76,77	-0.016	-0.001	0.025	-0.003
5	25,27,43,75,76,77,274,323,340	-0.020	-0.005	0.022	-0.006
6	**27,43,75,77,94**	**-0.015**	**0**	**0.027**	**-0.001**
7	25,27,43,76,274,340	-0.016	-0.001	0.020	-0.008
8	77,119,187,231,273,274,275	-0.018	-0.003	0.026	-0.002
9	25,27,43,75,77,274,323,340	-0.019	-0.004	0.021	-0.007

Table S.80. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T1003 (PDB ID 6hrh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	46,47,368,369,370,371,399,401,445	-0.03	0.01	0.05	-0.01
2	172,173,219,247,395,473,474	-0.02	0.02	0.-03	-0.03
3	56,57,58,59,91	-0.02	0.02	0.01	-0.05
4	106,109,110,120,121,122,312	-0.03	0.01	0.01	-0.05
5	72,173,219,220,395,404,473,474	-0.02	0.02	0.04	-0.02
6	172,173,219,247,278,395,474	-0.02	0.02	0.03	-0.03
7	106,109,110,111,120,122,312	-0.06	-0.02	0.14	0.08
8	**50,146,172,219,247,389,390,391,404,473,474**	**0.06**	**0.1**	**0.15**	**0.09**
9	106,110,120,121,122,312	-0.02	0.02	0.01	-0.05

Table S.81. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T1003 (PDB ID 6hrh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	143,144,145,146,277,283,284,473,474	-0.03	0.01	**0.11**	**0.05**
2	146,172,219,247,278,395,404,474	-0.03	0.01	0.05	-0.01
3	146,172,173,174,177,473,474	-0.02	0.02	0.02	-0.04
4	146,147,150,172,173,174,177	-0.03	0.01	0.03	-0.03
5	**144,146,172,284,473,474**	**-0.02**	**0.02**	0.06	0
6	146,277,282,283,284,472,473,474	-0.02	0.02	0.06	0
7	109,110,111,112,308,309,311,313	-0.03	0.01	0.02	-0.04
8	144,146,147,172,473,474	-0.02	0.02	0.04	-0.02
9	147,150,151,154,181,302,303	-0.03	0.01	0.03	-0.03

Table S.82. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T1003 (PDB ID 6hrh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	143,144,145,146,277,283,284,473,474	-0.03	0.01	**0.11**	**0.05**
2	146,172,173,219,395,474	-0.02	0.02	0.03	-0.03
3	146,172,173,219,247,278,395	-0.03	0.01	0.04	-0.02
4	172,173,247,395,404,473,474	-0.02	0.02	0.03	-0.03
5	146,172,219,247,394,395,404	-0.03	0.01	0.05	-0.01
6	144,145,146,277,283,284,472,473,474	-0.02	0.02	0.10	0.04
7	146,172,173,174,177,473,474	-0.02	0.02	0.02	-0.04
8	143,144,277,283,284,473,474	-0.02	0.02	0.08	0.02
9	144,146,277,283,284,473,474	-0.02	0.02	0.07	0.01

Table S.83. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T1003 (PDB ID 6hrh)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	143,144,146,277,282,283,284,472,473,474	-0.03	0.01	**0.10**	**0.04**
2	170,171,172,173,219,247,395,404	-0.03	0.01	0.04	-0.02
3	146,172,395,473,474	-0.02	0.02	0.02	-0.04
4	171,172,173,219,220,395	-0.02	0.02	0.03	-0.03
5	146,172,219,247,394,395,404	-0.03	0.01	0.05	-0.01
6	146,172,173,219,395,474	-0.02	0.02	0.03	-0.03
7	172,173,219,247,395,473,474	-0.02	0.02	0.03	-0.03
8	171,172,173,218,219,220,395,404	-0.03	0.01	0.05	-0.01
9	172,219,395,404,473	-0.02	0.02	0.03	-0.03

Table S.84. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T1009 (PDB ID 6dru)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	287,289,291,292,356,357,359,360	-0.013	-0.923	0.09	-0.85
2	291,292,356,357,358,359,360	-0.012	-0.922	0.04	-0.9
3	287,289,290,291,292,329,356,357	-0.013	-0.923	0.10	-0.84
4	287,289,291,292,357,360	-0.011	-0.921	0.08	-0.86
5	158,493,496,637,641	-0.010	-0.92	0.016	-0.924
6	490,491,492,526,533,536,537	-0.012	-0.922	0.05	-0.89
7	490,492,526,533,536,537,540	-0.012	-0.922	0.04	-0.9
8	490,492,533,536,537,540	-0.011	-0.921	0.04	-0.9
9	109,110,112,113,142,145,147	-0.012	-0.922	0.005	-0.935

Table S.85. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T1009 (PDB ID 6dru)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	679,681,682,695,697,700	-0.011	-0.921	0.007	-0.933
2	257,258,260,527,560	0.13	-0.78	0.16	-0.78
3	257,258,527,560	0.14	-0.77	0.15	-0.79
4	258,260,527,560	-0.009	-0.919	0.07	-0.87
5	257,258,560	0.17	-0.74	0.15	-0.79
6	286,325,395,396,487	0.53	-0.38	0.40	-0.54
7	**286,325,395,396,487,520**	**0.61**	**-0.3**	**0.49**	**-0.45**
8	286,325,395,396	0.45	-0.46	0.31	-0.63
9	257,258,527,560	0.14	-0.77	0.15	-0.79

Table S.86. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T1009 (PDB ID 6dru)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	286,287,292,325,355	0.26	-0.65	0.23	-0.71
2	Same as complex 1	-	-	-	-
3	173,325,355,396,400,403	0.11	-0.8	0.16	-0.78
4	257,258,260,527,560	0.13	-0.78	0.16	-0.78
5	286,287,292,325,355,356	0.24	-0.67	0.24	-0.7
6	Same as complex 5	-	-	-	-
7	679,681,682,697,700	-0.01	-0.92	0.006	-0.934
8	173,396,487	0.17	-0.74	0.14	-0.8
9	**173,325,396,403,487**	**0.26**	**-0.65**	**0.24**	**-0.7**

Table S.87. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T1009 (PDB ID 6dru)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	289,290,292,356,357,359,360	-0.012	-0.922	0.05	-0.89
2	292,329,354,355,356,360,362	-0.012	-0.922	0.04	-0.9
3	292,354,356,357,360	-0.010	-0.92	0.03	-0.91
4	287,289,291,292,329,356,357	-0.012	-0.922	0.08	-0.86
5	287,291,292,356,357,360	-0.011	-0.921	0.07	-0.87
6	287,291,292,329,356,357	-0.011	-0.921	0.07	-0.87
7	173,396,400,401,403	-0.010	-0.92	0.07	-0.87
8	325,348,353,355,396,400,401,402,403	0.087	-0.823	0.17	-0.77
9	256,264,528,560,568,570,571	-0.012	-0.922	0.09	-0.85

Table S.88. Predicted ligand-binding site residues and MCC and BDT scores with box calculation 22.5Å for T1014 (PDB ID 6qrj)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The box calculation is 22.5Å

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	100,101,102,103	-0.03	0.02	0.006	-0.044
2	53,57,120,123	-0.03	0.02	0.01	-0.04
3	103,104,105,115,116	-0.03	0.02	0.01	-0.04
4	116,117,118,119	-0.03	0.02	0.007	-0.043
5	103,104,114,115,116,117	-0.03	0.02	0.01	-0.04
6	110,113,115	-0.02	0.03	0.009	-0.041
7	89,90,91,107	-0.03	0.02	0.04	-0.01
8	53,57,119	-0.02	0.03	0.01	-0.04
9	124,133,134	-0.02	0.03	0.01	-0.04

Table S.89. Predicted ligand-binding site residues and MCC and BDT scores with 10% box calculation for T1014 (PDB ID 6qrj)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 10% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	**61,85,88,89,90,98,103,104,105,110,115,116,144,146**	**-0.05**	**0**	**0.13**	**0.08**
2	61,85,87,88,89,90,98,103,104,105,110,113,115,116, 117,120,144	-0.06	-0.01	0.12	0.07
3	61,88,90,98,104,105,110,113,115,120,146	-0.05	0	0.06	0.01
4	61,88,89,90,98,103,104,105,107,110,113,114,115, 116,117,120	-0.06	-0.01	0.07	0.02
5	61,88,89,90,91,98,104,105,107,110,113,114,115,120	-0.05	**0**	0.09	0.04

Table S.90. Predicted ligand-binding site residues and MCC and BDT scores with 20% box calculation for T1014 (PDB ID 6qrj)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 20% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	61,88,89,90,98,103,104,105,107,110,113,114, 115,116,120	-0.06	-0.01	0.08	0.03
2	61,88,89,90,91,98,103,104,105,107,110,113,115,116, 117,120,140	-0.06	-0.01	**0.10**	**0.05**
3	61,88,89,90,91,94,98,103,104,105,107,110,113,115,116	-0.06	-0.01	0.09	0.04
4	61,88,90,91,98,104,105,107,110,113,114,115,116	-0.05	0	0.07	0.02
5	61,88,89,90,98,103,104,105,107,110,113,114,115,116,120	-0.06	-0.01	0.08	0.03

Table S.91. Predicted ligand-binding site residues and MCC and BDT scores with 50% box calculation for T1014 (PDB ID 6qrj)
Models are listed in numerical order following docking with change in MCC and BDT given as a percentage increase or decrease. The model with the best MCC and BDT is in bold. The grid box calculation was based 50% of the ligand-binding site

Model number	Predicted ligand-binding site residues	MCC	Score change	BDT	Score change
1	53,57,60,114,116,117,118,119,122	-0.04	0.01	0.02	-0.03
2	52,53,56,57,60,102,116,117,118,119	-0.04	0.01	0.02	-0.03
3	57,104,114,116,117,118,119	-0.04	0.01	0.02	-0.03
4	60,102,104,114,115,116,117,118,119	-0.04	0.01	0.02	-0.03
5	53,60,114,115,116,117,118,119	-0.04	0.01	0.02	-0.03

Appendix 5

Comparison to C-I-TASSER

Contact-guided Iterative Threading ASSEmbly Refinement or C-I-TASSER is a method extended from I-TASSER for high accuracy protein structure and function predictions.(Zheng, Li, *et al.*, 2019) Starting from a query sequence, C-I-TASSER first generates inter-residue contact maps using multiple deep neural-network predictors.(Zheng, Li, *et al.*, 2019)C-I-TASSER predicted 3D structural models and function annotations for all proteins encoded by the genome of SARS-CoV-2, for comparison the predicted models from ReFOLD were compared against predictions by C-I-TASSER. Additionally, using the C-I-TASSER models ligand-binding site predications will be made to determine if potentially improvements in the 3D structure could have improved the predictions. A comparison of the 3D models selected by ModFOLD8 and then refined ReFOLD3 using and C-I-TASSER are given below in Figure S.83, only models which were included as CASP Commons targets are being compared and comparisons the second round of the full structure, where applicable.

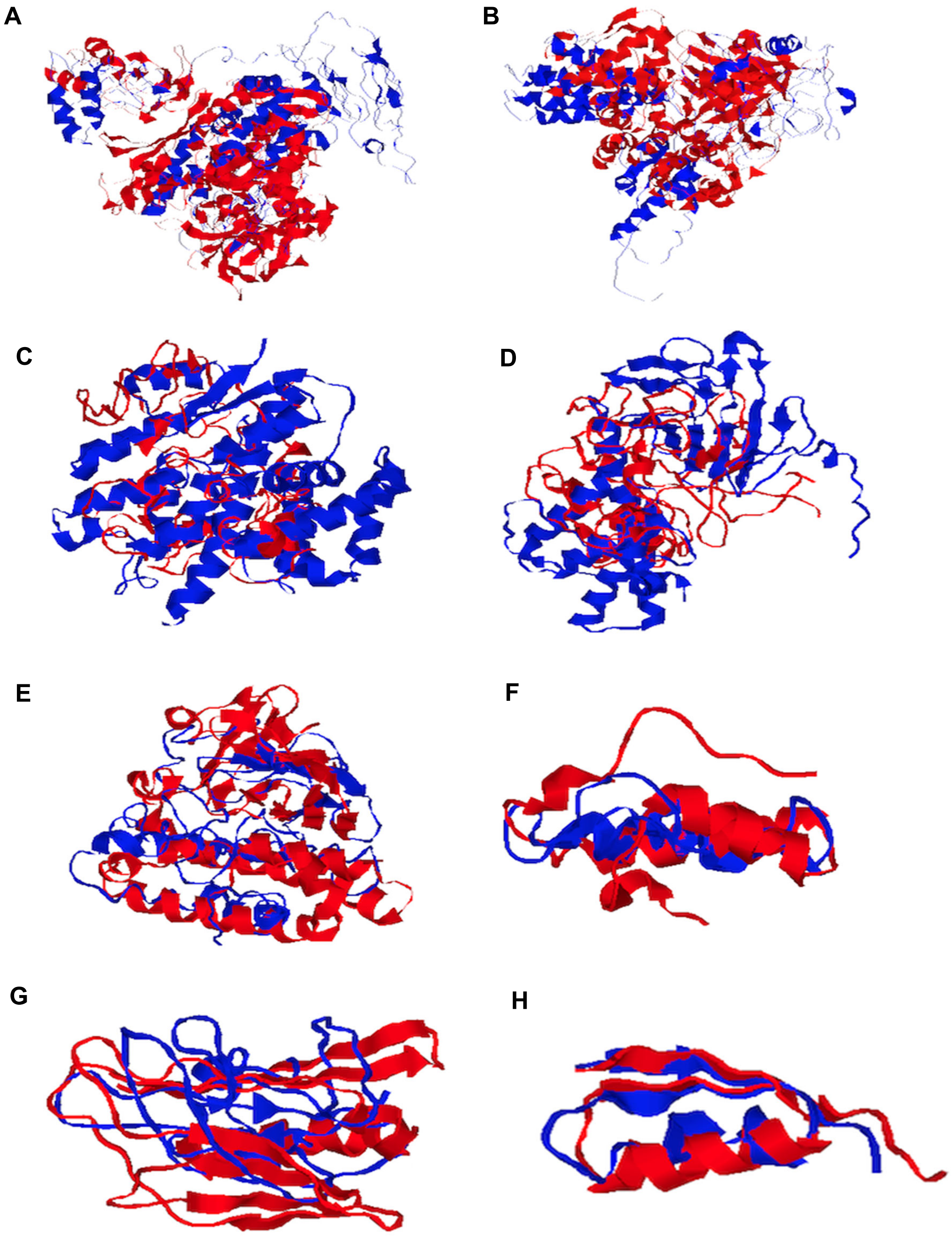


Figure S.83. Comparison of 3D predicted structures by C-I-TASSER and ReFOLD
(A) The structure in blue is the nsp2 structure from C-I-TASSER and the predicted structure for C1901 from ReFOLD is shown in red. A TM-score of 0.26184 was achieved for the protein structures. **(B)** The structure in blue is the nsp4 structure from C-I-TASSER and the predicted structure for nsp4 (C1902) from ReFOLD is shown in red. A TM-score of 0.26293 was achieved for the protein structures. **(C)** The structure in blue is the nsp6 structure from C-I-TASSER and the predicted structure for C1903 from ReFOLD is shown in red. A TM-score of 0.28611 was achieved for the protein structures. **(D)** The structure in blue is the 0RF3a structure from C-I-TASSER and the predicted structure for C1905 from ReFOLD is shown in red. A TM-score of 0.25072 was achieved for the protein structures. **(E)** The structure in blue is the Membrane protein structure from C-I-TASSER and the predicted structure for C1906 from ReFOLD is shown in red. A TM-score of 0.42324 was achieved for the protein structures. **(F)** The structure in blue is the ORF6 structure from C-I-TASSER and the predicted structure for C1907 from ReFOLD is shown in red. A TM-score of 0.37708 was achieved for the protein structures. **(G)** The structure in blue is the ORF8 structure from C-I-TASSER and the predicted structure for C1908 from ReFOLD is shown in red. A TM-score of 0.33229 was achieved for the protein structures. **(H)** The structure in blue is the ORF10 structure from C-I-TASSER and the predicted structure for C1909 from ReFOLD is shown in red. A TM-score of 0.45423 was achieved for the protein structures.

As can be seen from the comparison in Figure S.83, there was a poor overall structural superposition between the 3D structures from C-I-TASSER and top server models selected by ModFOLD8 and then refined using ReFOLD3, the structures with better structural homology were the proteins with smaller residues lengths e.g. ORF10, ORF6 and Membrane protein. The next stage, was to utilise the C-I-TASSER 3D structure models for FunFOLD3 ligand-binding site predictions. Of the eight proteins, only one has ligand predicted and is shown below in Figure S.84. Similarities can be seen with the predictions from FunFOLD3, with HEM ligand being predicted with both structures.

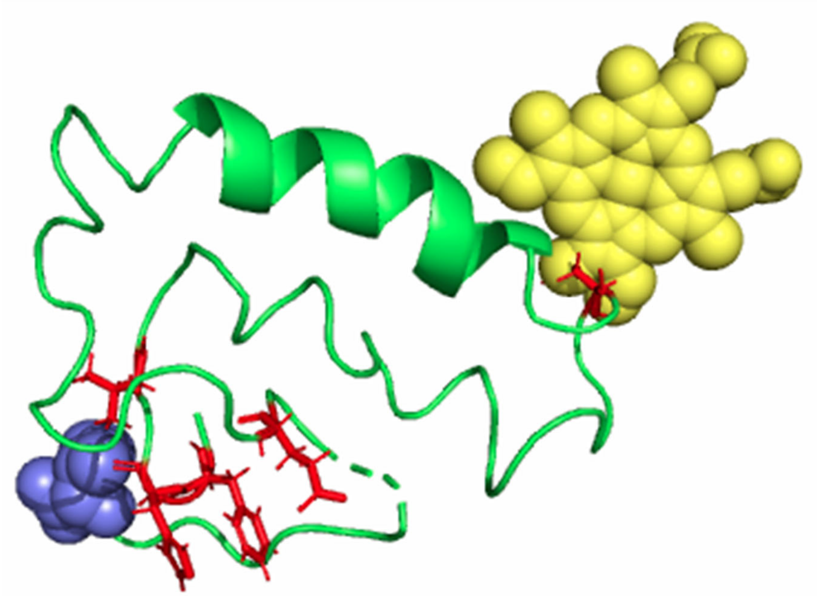


Figure S.84. Predicted ORF6 protein structure from C-I-TASSER
Predicted structure of ORF6 using the software C-I-TASSER. Predicted structure is shown as cartoon and coloured green. The predicted ligand HEM is shown as sphere and coloured yellow and the predicted ligand ILE is shown as sphere and coloured blue