University of Reading

Improvement of MD-Based Protocols for the Refinement of 3D Protein Models

A thesis submitted for the degree of
Doctor of Philosophy

School of Biological Sciences
University of Reading

Recep Adiyaman

March 2021

Declaration

I confirm that this is my own work and to the best of my knowledge, does not breach copyright law, and has not been taken from other sources except where such work has been cited and acknowledged within the text.

Recep Adiyaman

Date:17/03/2021

Abstract

Proteins are vital constituents of living cells with diverse structural and functional roles. Therefore, the study of functions and structures of proteins is key to our full understanding of living systems at the molecular level. X-ray crystallography and Nuclear Magnetic Resonance (NMR) are the main experimental techniques used to determine protein structures. However, such procedures are costly, labour intensive and time consuming, and many proteins are problematic to solve experimentally. *In silico* modelling of protein structures provides a potential solution to bridge the huge protein sequence-structure knowledge gap, which widening due to relative efficiency of cheap Next Generation Sequencing compared with experimental methods for determining structures. Nevertheless, the accuracy of predicted 3D structures may not always be adequate for further biological studies compared to experimental data. The refinement of the 3D protein models refers to process used for the improvement of predicted structures, by moving them closer towards experimental quality.

Since the 10th Critical Assessment Structure Prediction (CASP10), the usage of Molecular Dynamics (MD)-based refinement protocols has been found to be more effective compared with other protocols. However, the most successful MD-based protocols generally require supercomputer scale resources in order to refine a single 3D protein model. The ReFOLD server was developed by our group to rapidly refine 3D models with more modest computational resources. However, in CASP12 it was found that many of the 3D models from ReFOLD still contained structural flaws and some had drifted further away from the native structure during the refinement process.

Many restraint strategies have been used to prevent 3D models from the undesired deviations caused by force field inaccuracies. Here, we propose to use to prior predictive data to provide reliable guidance to the original MD-based protocol of ReFOLD, in order to direct the refinement of models towards the native basin. In the first part of this study, the predicted local model quality scores produced by the ModFOLD server were utilised to guide the original MD-based protocol of ReFOLD. A fixed threshold based on the predicted per-residue error was applied to determine

the poorly predicted regions in a 3D model, which could be targeted for refinement. The local quality assessment guided restraint strategy was successful in improving a higher number of 3D models, outperforming the original MD-based protocol according to observed scores. The local quality assessment guided MD-based protocol was also used to refine CASP13 targets, for the refinement and regular prediction categories, and it ranked among the top 10 approaches according to the official independent assessment.

Following the CASP13 experiment, the application of a fixed threshold based on the per-residue accuracy score was found to be less applicable for the multi-domain structures. Therefore, we proposed a novel gradual restraint strategy by considering the need of refinement for each residue according to the per-residue accuracy score. The gradual restraint strategy led to further increases in the population of the improved models compared to the fixed restraint strategy. We also applied our gradual restraint strategy for the refinement of the SARS-CoV-2 targets as a part of the CASP Commons COVID-19 initiative, in order to increase the accuracy of best-predicted 3D models, which were identified by our ModFOLD server. A significant number of the estimated top 10 models for each of the targets were generated by our group, according to the initial CASP Commons evaluation.

Residue-residue contact prediction methods have now reached up to 70% accuracy and the methods have proved to be useful in protein folding, model quality estimation and drug design. In this study, we describe the first attempt at applying contact predictions for refinement, where we use our Contact Distance Agreement (CDA) scores to apply gradual restraints to guide the MD protocol. The contact-assisted restraint strategy performed well, increasing the population of the improved models in comparison with the gradual restraint strategy based on the local quality estimation.

Finally, a binding site focused MD-based refinement protocol was also developed to improve the quality of the protein-ligand binding sites predicted by our FunFOLD server. This focused refinement protocol was successful at increasing the accuracy of all predicted binding sites that were tested as well as improving global model quality of some models.

This thesis has focused on exploiting prior predictive data for use in refinement pipelines, to direct the generation of 3D models closer towards the native basin. In the near future, each of the improvements described here will be integrated with new versions of our prediction servers, which will then be made freely accessible for use by general biologists.

Contents

List of Figures

List of Tables

List of Abbreviations

CASP	Critical Assessment of Structure Prediction
CDA	Contact Distance Agreement
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CPU	Central Processing Unit
CRYO-EM	Cryogenic Electron Microscope
DANNs	Deep Artificial Neural Networks
DBA	Disorder B-factor Agreement
DFIRE	Distance-Scaled, Finite-Ideal Gas Reference
DDFIRE	Dipolar Distance-Scaled, Ideal Gas Reference
DMP	DeepMetaPSICOV: A contact prediction method
DOIs	Digital Object Identifiers
DSSP	Dictionary of Secondary Structures of Proteins
EMA	Estimate of Model Accuracy
FM	Free Modelling
GDT	Global Distance Test
GDT-HA	Global Distance Test-High Accuracy
GDT-TS	Global Distance Test-Total Score
GPU	Graphics processing unit
LDDT	Local Distance Difference Test
MCC	Matthews correlation coefficient
MD	Molecular Dynamics
MLP	Multilayer Perceptron
MNIST	Modified National Institute of Standards and Technology
MSA	Multiple Sequence Alignment
ModFOLD	A model quality assessment program
MQAPs	Model Quality Assessment Programs
NAMD	Nanoscale Molecular Dynamics
NMR	Nuclear Magnetic Resonance

NN	Neural Network
PDB	Protein Data Bank
PKA	Protein Kinase A
PR	Precision and Recall
QA	Quality Assessment
QE	Quality Estimation
RMSE	Root Mean Squared Error
RWplus	Random Walk reference state Plus
RMSD	Root mean square deviation
ReFOLD	A refinement program
SG	SphereGrinder
SSA	Secondary Structure Agreement
SSEA	Secondary Structure Element Alignment
SV	Support Vector Machine
TBM	Template-Based Modelling
TM-score	Template Modelling Score

Acknowledgement

I would like to give my sincere thanks to Professor Liam McGuffin for all of the support and guidance he provided for me during my PhD. Without his patience, and humble support, I would not have been eager to learn and experience new horizons.

I also appreciate Professor Kim Watson, Professor Hugh Shanahan, and Professor David Leake for their sharing their experience and knowledge on our research. I would like to also thank my friends Emine, Ali, Naqib, Bajuna, Fahd, and Danielle.

I also would like to express my deepest gratitude to my family for their unconditional love, care, and understanding. They never let me feel lonely even when they are at a distance from me as I know their thoughts and prayers are always with me.

Chapter 1: General Introduction

Work presented in this chapter has been published in the following paper:

Recep Adiyaman, and Liam James McGuffin. "Methods for the Refinement of Protein Structure 3D Models." International journal of molecular sciences 20.9 (2019): 2301. (Both authors contributed equally to the paper as first authors. Figures and tables are adapted from Adiyaman and McGuffin 2019, unless otherwise indicated)

1.1 Protein Structure

Proteins have many crucial roles in biological reactions, perform many different functions, and are the second most abundant substance in living cells after water. Proteins consist of linear chains of various combinations of amino acids, which are covalently bonded with peptide bonds. Hydrogen bonds, ionic bonds and van der Waals interactions stabilise the folded proteins at different structural levels, and the bonds have a significant role in determining different functions, such as enzymes, messengers and other structural components (Brocchieri & Karlin, 2005; Rangwala & Karypis, 2010; Roche et al., 2013a; Stoker, 2013; Williamson, 2012a, 2012b). The structure of a protein determines its function, therefore, the elucidation of the 3D structures of proteins has always been a key factor in understanding their functions, ever since the first protein structure, myoglobin, was determined by John Kendrew and Max Perutz in 1958 (Brändén & Tooze, 1991; Fletterick, 1992; Williamson, 2012a). Fully understanding a protein's function depends on understanding the forces and interactions at its different structural levels. These 4 principal structural levels are hierarchical and described as the primary, secondary, tertiary and quaternary structures (see sections1.1.1-1.1.4).

Amino acids are the building blocks of proteins; each has a central carbon atom (C α) connected by an amino (NH_2), a carboxyl (COOH) group, and a side chain (R) group There are 20 standard amino acids that are coded for by the universal genetic code (Williamson, 2012a) (Table 1.1). Each amino acid is defined by its unique side chain (R) group (Table 1.1). The amino acids have different characteristics resulting from side chain (R) groups, and there can be interactions between the side chains that can affect protein functions (Brändén & Tooze, 1991; Stoker, 2013). The amino acids are categorised as nonpolar, polar neutral, polar acidic and polar basic amino acids (Brändén & Tooze, 1991; Stoker, 2013). Nonpolar side chains are the most commonly seen side chain, and other residues have different side chains in terms of their size, shape, acidity, chemical reactivity and charges that may be positive or negative (Brändén & Tooze, 1991; Stoker, 2013; Williamson, 2012a). Each of the twenty standard amino acids is denoted by a single letter (or less commonly three-letter) code to represent protein sequence (Table 1.1) (Brändén & Tooze, 1991; Brocchieri & Karlin, 2005; McGuffin & Roche, 2011; Stoker, 2013). The twenty amino acids can also be

combined to from protein chains of different lengths, which form different 3D-structures. Short chains of amino acids (less than 40 amino acids) are usually called peptides or oligopeptides, whereas longer chains are called proteins (Brändén & Tooze, 1991; Stoker, 2013; Williamson, 2012a).

Amino acid	Abbreviation		Side chain Polarity
	Three- letter	One-letter	
Aspartic acid ($C_4H_7NO_4$)	Asp	D	Polar
Glutamic acid ($C_5H_9NO_4$)	Glu	E	Polar
Arginine ($C_6H_{14}N_4O_2$)	Arg	R	Polar
Lysine ($C_6H_{14}N_2O_2$)	Lys	K	Polar
Histidine ($C_6H_9N_3O_2$)	His	H	Polar
Asparagine ($C_4H_8N_2O_3$)	Asn	N	Uncharged polar
Glutamine ($C_5H_{10}N_2O_3$)	Gln	Q	Uncharged polar
Serine ($C_3H_7NO_3$)	Ser	S	Uncharged polar
Threonine ($C_4H_9NO_3$)	Thr	T	Uncharged polar
Tyrosine ($C_9H_{11}NO_3$)	Tyr	Y	Uncharged polar
Alanine ($C_3H_7NO_2$)	Ala	A	Nonpolar
Cysteine ($C_3H_7NO_2S$)	Cys	C	Nonpolar
Glycine ($C_2H_5NO_2$)	Gly	G	Nonpolar
Isoleucine ($C_6H_{13}NO_2$)	Ile	I	Nonpolar
Leucine ($C_6H_{13}NO_2$)	Leu	L	Nonpolar
Methionine ($C_5H_{11}NO_2S$)	Met	M	Nonpolar
Phenylalanine($C_9H_{11}NO_2$)	Phe	F	Nonpolar
Proline ($C_5H_9NO_2$)	Pro	P	Nonpolar
Tryptophan ($C_{11}H_{12}N_2O_2$)	Trp	W	Nonpolar
Valine ($C_5H_{11}NO_2$)	Val	V	Nonpolar
Alanine ($C_3H_7NO_2$)	Ala	A	Nonpolar

Table 1. 1 List of the standard 20 amino acids.
Adapted from Williamson, (Williamson, 2012a, 2012b)

1.1.1 Primary Structure

The primary structure of protein consists of the amino acid sequence, with covalent peptide bonds between amino acids forming linear unbranched chains. The peptide bonds are the strongest bonds in proteins and form first during translation. A specific gene sequence in DNA specifies the primary structure via mRNA which is then translated into a protein. Frederick Sanger first discovered the linear amino acid sequence in insulin and defined the sequence of amino acids (Mandle et al., 2012; Sanger, 1959; Sanger & Tuppy, 1951). Determining the primary structure is possible with Edman degradation and mass spectrometry, and the genetic code can be translated to enable reading of the amino acid sequences from gene sequences (Mandle et al., 2012). Therefore, determining the primary structure of a protein, or its amino acid sequence, is relatively straightforward following high throughput next generation genome sequencing.

1.1.2 Secondary Structure

The next level of protein structure is the secondary structure which describes the specific sub-structures of a protein, namely the α-helices and β–strands. In between the helices and strands, there are also irregularly shaped ordered elements, called loops or random coil, which are also important in protein structure and functions. (Mandle et al., 2012; Pauling et al., 1951; Rangwala & Karypis, 2010). The α-helices and β–strands contribute to the stabilisation of a protein structure. The α–helix, is a common form of the secondary structure of right or left-handed twists, and the structure is kept by hydrogen bonding between N-H and C=O groups (Rangwala & Karypis, 2010; Stoker, 2013). β strands interact with each other forming hydrogen bonds also make up β–sheets, which are common in many structures. The direction of β-sheets depends on the β-strand positions, and they can be made up of either parallel or antiparallel strands (Brändén & Tooze, 1991; Stoker, 2013; Williamson, 2012a).

It is possible to determine the secondary structure by using spectroscopic methods like far-ultraviolet (Pelton & McLean, 2000). Although the infrared spectroscopy has been rarely used, an initially unassigned NMR spectrum can be used to get structural information of the protein

secondary structure (Meiler & Baker, 2003). Circular dichroism (CD) spectroscopy is also used to determine secondary structures, and the structures are obtained using characteristic spectrums (Greenfield, 2006).

Predicting secondary structure relies on accurately identifying the probabilities of α-helices and β–strand formation by consecutive amino acids, and the prediction is related to the calculation of the free energy of the structure. Chou–Fasman method (Chou & Fasman, 1978) is one of the first methods of secondary structure prediction (Chou & Fasman, 1978). The field of secondary structure prediction has since matured, and it is largely thought of as a solved problem - the accuracy of predictions (~80%) is at the same level as the discrepancy between the different methods for defining observed secondary structures. There are many accurate approaches to predict secondary structure, such as PSIPRED (McGuffin et al., 2000; Ward et al., 2003), SAM (Karplus, 2009), PORTER (Pollastri & McLysaght, 2005) and PROF (Adamczak et al., 2005).

1.1.3 Tertiary Structure

The three-dimensional description of atoms within a protein is referred to as the tertiary structure, and it comprises the primary and secondary structures and the amino acid side chain interactions. The interactions include electrostatic attractions, covalent disulphide bonds, hydrogen bonds, and hydrophobic interactions (Brändén & Tooze, 1991). Electrostatic interactions (salt bridges) are found between acidic and basic R groups. Disulphide bonds occur between SH-groups in cysteine residues, and they are the strongest and covalent bonds between side chains in the tertiary structure (Brändén & Tooze, 1991; Stoker, 2013; Williamson, 2012a). Hydrogen bonds also saturate interactions between amino acids in the sub-structures, and they can be weaker compared to other types of interactions. Non-polar R groups form hydrophobic interactions to stabilise the protein structure (Brändén & Tooze, 1991; Stoker, 2013; Williamson, 2012a). The side-chain interactions also form the final shape and function of the proteins (Baldwin, 2007).

The tertiary structure can be determined experimentally by using mainly X-ray crystallography and Nuclear Magnetic Resonance (NMR) and Cryogenic Electron Microscopy (Cryo-EM) (see

section 1.2) (Brändén & Tooze, 1991; Mandle et al., 2012; Rangwala & Karypis, 2010; Stoker, 2013; Williamson, 2012a). Tertiary structures can also be predicted by Template-Based Modelling (TBM) and Template-Free Modelling (FM) methods, which will be explained in section 1.4.

1.1.4 Quaternary Structure

Quaternary structure is the last level of protein structure which is represented by multiple-subunits of protein 3D chains and describes their non-covalent interactions in multimeric assemblies (Brändén & Tooze, 1991; Mandle et al., 2012; Rangwala & Karypis, 2010; Stoker, 2013; Williamson, 2012a). Thus, quaternary structures contain two or more separate protein chains that interact to form a multi-subunit complex or multimer. Electrostatic interactions, hydrogen bonds and hydrophobic forces provide stability for the interacting subunits forces in quaternary structures (Brändén & Tooze, 1991; Chiang et al., 2007; Mandle et al., 2012; Rangwala & Karypis, 2010; Stoker, 2013; Williamson, 2012a).

The prediction of quaternary structures is based on docking chains and symmetrical arrangement and RosettaDock (Gray et al., 2003; Sircar et al., 2010), DOT (Roberts et al., 2013),and HADDOCK (Kastritis et al., 2014) ZDOCK (Pierce et al., 2014) are among the popular methods (Seffernick & Lindert, 2020).

1.2. Protein Structure Determination and Prediction

Predicting and determining the structure of specific proteins enable a better understanding of the molecular mechanisms of diseases and allow us to infer protein function (McGuffin, 2008). Anfinsen demonstrated in the 1970s, that all information about the protein folding is contained within its own protein sequence (also known as Anfinsen's dogma or the thermodynamic hypothesis) (Anfinsen, 1973; Williamson, 2012a). This means that the prediction of the 3D structure is theoretically possible solely utilising the information from the amino acid sequence. Notwithstanding, some proteins which may perform their function in the unfolded or disordered

state. These proteins are known as intrinsically disordered proteins (IDP) and make up almost a third of the protein sequences (Dorn et al., 2014; Dunker et al., 2001; Tompa & Csermely, 2004; Uversky, 2002; Wright & Dyson, 1999). There have been many efforts by groups to understand protein folding and to develop accurate structure prediction methods that used the protein sequence information and our knowledge of known structures. In this section, the experimental and computational methods used to determine and predict protein structures will be introduced (Dorn et al., 2014; Guo et al., 2008).

1.2.1 Experimental Methods for Protein Structure Determination

The experimental methods have played the key part in the elucidation of the protein structures since the structure of myoglobin was firstly solved using x-ray crystallography in 1958 (Brändén & Tooze, 1991; Fletterick, 1992). The protein-ligand and protein-protein interactions are determined but the shape or fold of the protein, so determining protein structures has a significant role in determining their functionality (Dorn et al., 2014; Dunker et al., 2001; Tompa & Csermely, 2004; Uversky, 2002; Wright & Dyson, 1999). A knowledge of the interactions occurring within and between proteins and ligands may enable further studies, such as new diagnostic methods, enzyme engineering, drug discovery and development and disease-related proteins (Dorn et al., 2014; Dunker et al., 2001; Tompa & Csermely, 2004; Uversky, 2002; Wright & Dyson, 1999). The determination of the protein structures may also provide alternative perspectives for therapy in disease caused by protein misfolding, including Parkinson's (Hughes et al., 2017) and Alzheimer's (Ashraf et al., 2014)

X-ray crystallography, Nuclear Magnetic Resonance spectroscopy (NMR) and Cryo-Electron Microscopy (Cryo-EM) are the main experimental methods that provide the most information for determining 3D protein structures. While the approaches have different procedures, they each give information about the relative 3D positions of each atom in a protein structure (Jöbstl et al., 2006; Petsko & Ringe, 2004; Williamson, 2012b). X-ray crystallography and Nuclear Magnetic Resonance spectroscopy are generally the preferred "gold-standard" methods for determining protein structures, while Cryo- EM is a more recent development, which is improving for the

elucidation of larger complexes albeit at lower resolution (Jöbstl et al., 2006; Petsko & Ringe, 2004; Williamson, 2012b).

1.2.1.1 X-ray Crystallography

X-ray crystallography is the most productive method in terms of determining the protein structures at atomic resolution (Jöbstl et al., 2006; Petsko & Ringe, 2004; Williamson, 2012b). In the first stage, X-ray crystallography requires purification to obtain the protein samples at a high enough concentration for crystallisation. Following crystallisation, in the second stage, X-rays are fired through the crystals, which interact with the electrons and then are scattered by them as defined by Braggs law (Bragg et al., 1913; Brändén & Tooze, 1991; Matthews, 1975; Petsko & Ringe, 2004; Williamson, 2012b). The 3-D atomic positions are then calculated from the electron density maps and are then refined using computational and chemical approaches to provide the resolved 3-D structures. For globular proteins X-ray crystallography is very successful, however, the method may not be applied to flexible or disordered regions of proteins (these regions cannot be resolved) and membrane proteins are often problematic (Drenth, 1999; Petsko & Ringe, 2004; Williamson, 2012b).

1.2.1.2 Nuclear Magnetic Resonance (NMR)

NMR can be used in solution without crystallisation, but the proteins should be soluble and its principle relies on the absorption of energy to provide a transition from one state to another state (Jöbstl et al., 2006; Petsko & Ringe, 2004; Williamson, 2012b). The states are two different nuclear spins, up and down, in a strong magnetic field and the magnetic resonance can be used to identify the distinctive resonance of each atom position (Heinemann et al., 2002; Petsko & Ringe, 2004). The observed characteristic resonances gives information about the position of the atoms in the structure (Drenth, 1999; Heinemann et al., 2002; Petsko & Ringe, 2004). If X-ray and NMR methods are applied to the same protein structure, then roughly similar structures should be found. Nevertheless, X-ray crystallography provides much more information about the structure and

higher resolution structures may be obtained compared to NMR (Drenth, 1999; Petsko & Ringe, 2004; Williamson, 2012b).

1.2.1.3 Cryogenic Electron Microscopy (Cryo-EM)

Recent developments in the electron microscopy and imaging technology enable determination of large protein complexes with Cryo-EM at relatively high resolutions (~1.25-1.5Å) (Yip et al., 2020). The determination process starts with the vitrification step, where samples are cooled at cryogenic temperature around –180 °C (Cabra & Samsó, 2015; Cho et al., 2013; Schmidt & Urlaub, 2017). The experimental stage is followed by image processing using advanced computational software to represent the 3D models (Murata & Wolf, 2018). The preparation of the samples may still be time-consuming, and the experimental procedure is relatively costly. Furthermore, determining high-resolution structures may not be possible depending on the thickness of the sample in ice (Cabra & Samsó, 2015; Cho et al., 2013; Costa et al., 2017; Jonic & Vénien-Bryan, 2009; Murata & Wolf, 2018; Schmidt & Urlaub, 2017)

1.3 Protein Sequence and Structure Databases

There are several core databases for protein sequences and three-dimensional structures, such as The Universal Protein Resource (UniProt), The National Center for Biotechnology Information (NCBI), the Protein Data Bank (PDB), the Structural Classification of Proteins (SCOP) and CATH (Class Architecture Topology Homology). Uniprot includes identified functional and sequence documentation of proteins and genomes and research literature used to collect the data (UniProt Consortium, 2008, 2015, 2011). More than 235 million protein sequences are also available in the Uniprot database (UniProt Consortium, 2008, 2015, 2011).

The Protein Data Bank (PDB) also contains protein structures obtained by X-ray crystallography, NMR, cryo-electron microscopy. (Brändén & Tooze, 1991; Fletterick, 1992; Sherry et al., 2001). By the end of September 2020, there have been 168,888 structures made available in the PDB; 149,528 structures determined by X-ray crystallography, 13,120 structures by NMR, and 5,740

structures by electron microscopy (Berman et al., 2000; Bernstein et al., 1977). The PDB is a critical resource in terms of structural biology enabling researchers worldwide to freely access protein structure data and the PDB enables scientists to obtain and upload new structures (Bernstein et al., 1977). The atomic positions and related information about the crystal structure of the experimentally determined protein structure are stored in the PDB file format as seen in Figure 1.1.

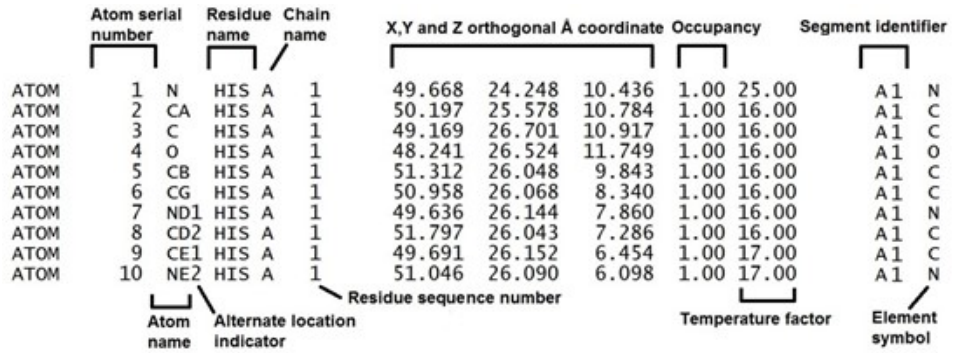

Figure 1. 1 PDB file format
Adapted from https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html.

Structural Classification of Proteins (SCOP) (Andreeva et al., 2008; Lo Conte et al., 2000) and CATH (Class Architecture Topology Homology) (Sillitoe et al., 2015) provide a hierarchical classification of proteins allowing researchers to study the similarities between protein structural domains, sequences and functions for all known structures.

1.3.1 Protein Sequence-Structure Gap

While the experimental methods for determining protein sequences are now automated, rapid and very cheap, the experimental methods for determining protein structure often encounter many difficulties in the cloning, expression and purification stages of the process. In addition, the experimental procedure can often lead to inadequate quality of crystals and reaching these steps can often be time consuming and expensive. Moreover, due to these time and cost constraints, the experimental methods are not yet efficient enough to determine the structures of all proteins, so there has been a significant widening gap between known protein sequences and solved 3D structures (Roche et al., 2014; Roche & McGuffin, 2016b).

Approximately 168,888 protein three-dimensional structures that have been determined by the experimental methods are available in the Protein Data Bank, whereas the UniProt database just hosts roughly 235 million protein sequences at the present time (Berman et al., 2000; UniProt Consortium, 2008, 2015; Westbrook et al., 2003). Recent developments in DNA sequencing, next-generation sequencing and proteogenomics have revealed millions of protein sequences. Just roughly one percent of the protein sequences in Uniprot have known structures, therefore the pace of the experimental methods for the determination of 3D protein structures is still far away from reaching that of sequencing methods (Kaján et al., 2014; Roche & McGuffin, 2016b). It is obvious that presently *in silico* modelling for predicting structures from the sequence is the only way to bridge the gap between known sequences and structures (UniProt Consortium, 2008, 2015, 2011).

1.4 Computational Studies on Protein Structures

In this section, the general computational methods for studying protein structures will be introduced and discussed. There has been a growing interest in predicting protein structures using computational methods as they enable scientists to predict 3D locations of atoms without spending the time and resources required for experimental methods. Nevertheless, the prediction of 3D structures at high accuracy has been a significant challenge for bioinformatics.

The process of computational modelling of protein structures includes steps starting from the prediction of the protein fold and ending with the assessment and refinement of 3D predicted models. Predicting protein structures with template free (FM) and template-based modelling (TBM) can produce many alternative models with different conformations (Adiyaman & McGuffin, 2019). These alternative models should then be evaluated using quality estimation tools to predict the most native-like structures and then they must be refined to bring them even closer to the actual structure (Bhattacharya & Cheng, 2013a; Roche et al., 2013a; Shuid et al., 2017). The accuracy of the predicted models is a key factor if they are to be usefully applied to help solve real world biological problems or be used in further computational studies such as, drug design, protein docking and the prediction of protein function.

1.4.1 Computational Protein Structure Prediction

There are two main strategies for predicting protein structures: Template-Based Modelling, or TBM (including homology modelling and fold recognition methods), and template-Free Modelling, or FM (including *ab initio* and fragment assembly methods). In general, the most successful methods have traditionally used a template-based approach, which relies on the usage of the experimental structures of related protein families as templates in order to predict the target structure. If the target sequence has similar protein families with the experimentally determined 3D structures, then template-based modelling methods can enable more accurate prediction compared to others (Dorn et al., 2014; Pavlopoulou & Michalopoulos, 2011; Roche & McGuffin, 2016b). Fold recognition and threading methods are also useful in case of low sequence similarity to build 3D models. Conversely, template-free modelling or *ab initio* prediction methods can be used to predict the structures in cases where suitable templates are unavailable. It is better to apply TBM to predict protein structures as a first step, but *ab initio* methods should be applied in the event where no suitable fold templates can be found (Dorn et al., 2014; Pavlopoulou & Michalopoulos, 2011; Roche & McGuffin, 2016b)..

1.4.1.1 Template-Based Modelling

Template-based modelling is a much more accurate approach to predict structures by utilising the evolutionary relationship based on the similarity with the protein structures having the similar protein families and sequences (Dorn et al., 2014; Guo et al., 2008; Pavlopoulou & Michalopoulos, 2011; Roche & McGuffin, 2016b). If the target sequences are derived from the same families as those with known structures, then the structures can be predicted by using the defined closest homolog as a template. Thus, the most simple method is called homology modelling, and this can predict protein structures by identifying experimentally determined structures with similar sequences (Bourne & Shindyalov, 2005; Dorn et al., 2014; Guo et al., 2008; Pavlopoulou & Michalopoulos, 2011; Rangwala & Karypis, 2010; Roche & McGuffin, 2016b). There is a growing

trend in applying the approach to proteins as the availability of protein structures determined by experimental methods gradually increases. The increase in known structures has led to an increase in the accuracy of predictions that can be obtained via homology modelling (Bourne & Shindyalov, 2005; Dorn et al., 2014; Guo et al., 2008; Pavlopoulou & Michalopoulos, 2011; Rangwala & Karypis, 2010; Roche & McGuffin, 2016b; Xiang, 2006).

Fold recognition is one of the widely preferred methods using similar folds or parts of folds (subdomains) as templates in order to predict protein structures in the absence of clear homologous protein sequences and structures (Peng & Xu, 2009). Proteins in different families can have similar structures, without having sequence similarity. Utilising fold recognition enables the prediction of target structures that are likely to share the same folds as determined structures (McGuffin, 2008b). There is no need for sequence similarity between queries and templates that have similar folding patterns, and there is no need to have identified similar protein families, as in homology modelling (Bourne & Shindyalov, 2005; Dorn et al., 2014; Guo et al., 2008; Pavlopoulou & Michalopoulos, 2011; Rangwala & Karypis, 2010; Roche & McGuffin, 2016b; Xiang, 2006). The availability of the Protein Data Bank and other 3D structure repositories, such as SCOP, has made it easier to study the similarity in fold patterns among templates with different sequences, in addition to identifying sequences with novel folds (Jones, 2000; McGuffin, 2008a). Fold recognition methods are generally more computationally intensive, and the predictions are based on the similarity with known folds. Therefore, these factors may limit the accuracy of the predictions made by the fold recognition methods (Jones, 2000; McGuffin, 2008a, 2008b).

1.4.1.2 Template Free Modelling

Template-free modelling (FM), also traditionally known as *ab initio* modelling is useful for predicting protein structures when a suitable template is unavailable. Such methods use physical, chemical, and thermodynamic principles to predict protein structure (Dorn et al., 2014; Pavlopoulou & Michalopoulos, 2011). *Ab initio* protein structure prediction is typically only used in the case of undetectable similarity between structures and sequences. These methods can be used to predict structures for small protein sequences having up to ~150 amino acids in order to

get accurate predictions without using known template structures (Dill et al., 2007; Lee et al., 2009). There is no direct use of known protein templates while predicting structures, and predictions are made by utilising energy functions and various modelling algorithms from its sequence as stated in Anfinsen's hypothesis (Anfinsen, 1973; Anfinsen et al., 1961; McGuffin, 2008a; Tramontano et al., 2008). *Ab initio* approaches can enable prediction of undetermined protein structures, but the accuracy of predicted models is often far lower when compared to those obtained via TBM (McGuffin, 2008a; Roche et al., 2013a). *Ab initio* also requires intensive computational resources and large conformational searches during the prediction of 3D models (McGuffin, 2008a; Roche et al., 2013a). Recent attempts focus on reducing the resources, the implications of the theoretical fundamentals and understanding of folding patterns (Allen et al., 2001; McGuffin, 2008a; Pande et al., 1998). Current applications of deep learning and machine intelligence have considerable potential for more accurate FM protein structure prediction since deep learning approaches were used to predict contacts to build 3D models in CASP13 (see Chapter 5) (Greener et al., 2019; Senior et al., 2019, 2020). AlphaFOLD which showed impressive performance in CASP13 employed different deep learning methods for FM prediction by not using knowledge of available structures, and its predictions were also based on deep distance prediction neural network utilising evolutionary covariation data (Senior et al., 2019, 2020). Some of the most widely used structure prediction tools and webservers are also listed in Table 1.2.

Name program ad servers	URL
PconsFold2 (Michel et al., 2017)	https://github.com/ElofssonLab/
RaptorX-Contact (Wang et al., 2017)	http://raptorx.uchicago.edu/ContactMap/
RBO_aleph (Mabrouk et al., 2015)	http://compbio.robotics.tu-berlin.de/rbo_aleph/
Rosetta (Das and Baker, 2008)	https://www.rosettacommons.org/
Seok-assembly (Ko et al., 2012)	http://galaxy.seoklab.org
QUARK (Xu and Zhang, 2012)	https://zhanglab.ccmb.med.umich.edu/QUARK/
HHpred (Hildebrand et al., 2009)	https://toolkit.tuebingen.mpg.de/#/tools/hhpred
I-TASSER (Roy et al., 2010)	https://zhanglab.ccmb.med.umich.edu/I-TASSER/
IntFOLD (McGuffin et al., 2019)	https://www.reading.ac.uk/bioinf/IntFOLD/
LOMETS (Wu and Zhang, 2007)	https://zhanglab.ccmb.med.umich.edu/LOMETS/
MODELLER (Fiser et al., 2000)	https://salilab.org/modeller/
PCONS (Wallner and Elofsson, 2006)	http://pcons.net/
SWISSMODEL (Arnold et al., 2006; Biasini et al.,2014; Bordoli et al., 2009)	http://swissmodel.expasy.org/workspace/
Rosetta (Das and Baker, 2008)	https://www.rosettacommons.org/
DMPfold (Greener et al., 2019)	https://github.com/psipred/DMPfold

Table 1. 2 List of some of the most popular server/programs for the prediction of 3D models.

1.5 The Critical Assessment of Techniques for Protein Structure Prediction

There has been a clear need to objectively test the reliability and capability of prediction programs and methods with sequences from unreleased experimentally determined structures. The Critical Assessment of Techniques for Protein Structure Prediction (CASP), which has been thought of as the "World Protein Structure Prediction Championships" or "Olympic Games of protein structure prediction" was firstly organised under the leadership of John Moult and his colleagues in 1994 (Fischer et al., 1999; Moult et al., 2005). CASP is a blind prediction experiment carried out every other year usually from May to September to effectively assess strategies aimed at modelling native-like structures from amino acid sequences. The competition involves a wide range of prediction groups from all around the world and evaluates improvements of methods in many prediction categories using a blind assessment process.

The CASP process starts with the determination of selected primary structure and the solution (or imminent solution) of the 3D protein structure by X-ray or NMR. The target sequences are given to predictor groups who will then predict structures and produce models prior to the release of the experimental data. Thus, predictors do not have any prior knowledge of the solved 3D structures before they are released. The ranking and assessment of prediction methods is made by independent assessors, who do not know the identities of the prediction groups, to ensure a reliable process performance. The predicted models and related data are evaluated by the CASP committee, who utilise mainly scoring methods based on the superposition of C-alpha atoms in the predicted and observed structures (Moult et al., 2014, 2016). The CASP experiments carry out a systematic and objective evaluation of the participating prediction methods, in the following prediction categories: tertiary structure prediction using template-based and template-free modelling, residue-residue contact prediction, disordered regions prediction, function prediction, model quality assessment, model refinement, protein-protein interactions; oligomerisation state and binding site prediction (Moult et al., 2009; Roche & McGuffin, 2016b). These prediction categories aim to provide a platform for the prediction groups to evaluate their performance strengths within the specific areas of their protein structure to function pipelines (Moult et al., 2014, 2016).

1.6 Model Evaluation

Predicting protein structures with either TBM or FM methods is challenging. Often many dozens or hundreds of alternative different 3D models are produced in the process of prediction, and it is often difficult to identify the most accurate model from among many alternatives. There is a need to assess the models to discover the most native-like models, thus various Model Quality Assessment Programs (MQAPs) have been developed in order to evaluate the alternative models, producing global and local quality scores for each model (McGuffin & Roche, 2010). Both the local and global errors, including unusual bonds and angles in alternative models, may be identified using MQAPs to evaluate the quality of predicted models. In the early years of CASP, basic stereochemical checks and simple energy functions were used for evaluation of models, however these initial approaches have been followed by more sophisticated methods to score the accuracy of 3D models.

The early methods were those which simply checked the stereochemical quality of the models, such as those used to validate structures obtained by NMR and X-ray crystallography, i.e. by using the Ramachandran plot which enables knowing possible combinations of φ, ψ dihedral angles (Ramachandran et al., 1963). The approach has been used to recognise unusual bonds in methods such as PROCHECK (Laskowski et al., 1993), WHAT-CHECK (Hooft et al., 1996), and MolProbity (Chen et al., 2010; Davis et al., 2004). PROCHECK helps to analyse global and local geometrical properties, and PROCHECK-NMR can be used to check protein structures acquired by NMR to know their reliability (Laskowski et al., 1993, 1996). Errors in protein crystal structure can also be found with WHAT-CHECK (Hooft et al., 1996). MolProbity is a more developed tool to optimise the hydrogen-bonding network and clashes (Chen et al., 2010; Davis et al., 2004). While these basic checks are important, such initial methods cannot be used to rank alternative folds and they do not provide a global score to distinguish correct models from incorrect ones, moreover, incorrect models can still have correct stereochemistry.

There are currently three main categories of quality assessment methods: the single-model methods, the clustering-based methods and the quasi single-model methods. The single model

methods are able to produce a global score for each predicted model separately in low availability of the models, and it is fair to say that the approach is fast, and such methods can produce reasonably consistent scores. However, the single model approach can be less accurate at producing global scores in cases where a wide variety of multiple 3D models are available for a given target (Maghrabi & McGuffin, 2017; McGuffin, 2007; McGuffin & Roche, 2010; Wallner & Elofsson, 2007). VERIFY 3D was one of the traditional MQAPs used to determine the local compatibility of the models by using their sequences (Eisenberg et al., 1997). PROSA-web was also later developed used for checking 3D structures (Wiederstein & Sippl, 2007), along with other single-model methods such as MetaMQAP (Pawlowski et al., 2008) and QMEAN, which used composite scoring functions (a combination of many scores) in order to assess the local and global quality of predicted models (Benkert et al., 2008, 2009).

The consensus, or clustering-based, multiple model approach is useful to produce quality scores in case of low similarity between many alternative models, and the approach can be used to compare multiple models simultaneously. The approach is more accurate, but computationally intensive compared to single-model methods. However, the selection of the best model is not always possible by using clustering-based methods compared to the single model methods (McGuffin & Roche, 2010; Roche et al., 2014). 3D-Jury was perhaps the initial approach program, which was based on grouping together models with similar conformations via all against all structural comparisons (Ginalski et al., 2003). The SPICKER method has also been developed using the k-means algorithm to select the most accurate models by structural clustering (Zhang & Skolnick, 2004a). The MULTICOM series of methods also use a clustering-based approach to assess predicted models with machine learning and hybrid techniques (Cheng et al., 2009). Other clustering-based tools include QMEANclust (Benkert et al., 2009), ModFOLDclust2 (McGuffin, 2009; McGuffin & Roche, 2010) and Pcons (see chapter 2) (Larsson et al., 2009; Lundström et al., 2001; Wallner & Elofsson, 2007)

The quasi-single model approach is an important development of model quality assessment programs. Such methods can produce scores with similar accuracy to the clustering-based methods in case of submission of multiple models, yet they also have the ability to make an assessment for a single model at a time (McGuffin et al., 2013; Roche et al., 2014). The quasi-single model

approach was first used in the development of the ModFOLD4 server and a more detailed description of the ModFOLD server will be given in Chapter 2.

1.7 Model Refinement

Further computational studies such as function prediction, drug design and protein-protein and protein-ligand docking often require highly accurate predicted 3D models for comprehensive *in silico* studies (Bonneau et al., 2001; Brylinski & Skolnick, 2008; Ekins et al., 2007; Feig, 2017; Laskowski et al., 2005; Mirjalili et al., 2014; Mirjalili & Feig, 2013; Oren et al., 2006; Roy et al., 2010; Wieman et al., 2004; Zhang, 2009). The process of improving the quality of 3D models is referred to as refinement and it is often used as the "last mile" for moving 3D models closer to the actual structures. Thus, refining predicted 3D models generated by either TBM or FM aims to increase their accuracy towards the native basin (Adiyaman & McGuffin, 2019). Despite the success in the improvement of TBM and FM over the years, predicted models may still include local and global errors including: unfavourable contacts, irregular hydrogen bonds, geometrical clashes and unrealistic bond lengths and angles (Bhattacharya & Cheng, 2013a; Hovan et al., 2018; Kryshtafovych et al., 2005; McGuffin et al., 2013; Moult, 2005). In some cases, errors may affect the utility of the 3D models, where near experimental accuracy is needed. To fix these types of errors, the refinement of the structures has been a crucial part of the prediction pipeline for further studies, what is also referred to as the "end-game" for structure prediction (Nugent et al., 2014; Shuid et al., 2017). Refinement of 3D models also includes the modification of the secondary structure elements, loops and sidechains. Importantly, it should be noted that these local and global errors may be successfully identified using Model Quality Assessment Programs (MQAPs), and the detected errors might be a useful guide for a targeted refinement process (Bhattacharya & Cheng, 2013a; McGuffin et al., 2013; Roche et al., 2013a).

Refinement of 3D models of proteins remains an unsolved problem. Recent attempts at increasing the accuracy of the predicted models have often resulted in the deterioration of accuracy; it is often challenging to improve upon the quality of the starting models (Adiyaman & McGuffin, 2019; Giorgetti et al., 2005; Meiler & Baker, 2003; Oren et al., 2006; Qian et al., 2007; Sliwoski et al.,

2014; Terashi & Kihara, 2018; Zhang, 2009). The refinement of a 3D model, particularly one predicted by TBM, is more likely to result in a deterioration in quality, compared to the initial structures, if the model is already based on a known structure and if it is already highly accurate. It is also worthy of note that refinement of 3D models predicted by FM is possibly more efficient, compared to refinement of TBM models, as there is more likely to be room for improvement (FM models are often less accurate) (Feig, 2017; Gront et al., 2012; MacCallum et al., 2009, 2011; Nugent et al., 2014). The success of the refinement approaches is based on two-important stages, which are both needed to consistently refine protein structures: the sampling stage and the scoring stage (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011). Firstly, the 3D models generated by the sampling approaches should be (on average) closer to the native state compared to the initial model. Secondly, the most improved models from among all generated models must be identified via the scoring approaches in order to complete a successful refinement process (Adiyaman & McGuffin, 2019). Different refinement approaches have been tried and tested over a decade of CASP experiments. Nonetheless, a refinement approach that always leads to consistent improvement has not yet been found. In CASP experiments, the best-predicted 3D models have often been identified and then assigned as the refinement targets, which makes the refinement process more difficult (less room for improvement, more opportunity to deteriorate), and so this may also be a limiting factor for the refinement category (Adiyaman & McGuffin, 2019).

1.7.1 Sampling Approaches in the Refinement Pipelines

There are two main categories of sampling tools: the fully-automated server-based programs and non-server-based highly CPU intensive approaches, also known in CASP as “human/manual” refinement methods (Adiyaman & McGuffin, 2019; Modi & Dunbrack, 2016; Nugent et al., 2014; Shuid et al., 2017). Different combinations of the knowledge-based methods (Bhattacharya et al., 2016; Bhattacharya & Cheng, 2013a, 2013b; Chopra et al., 2010; Jagielska et al., 2008; Lin & Head-Gordon, 2011; Lu & Skolnick, 2003; Mirjalili et al., 2014; Misura & Baker, 2005; Nugent et al., 2014; Rodrigues et al., 2012; Xu & Zhang, 2011; Zhang, 2009), Monte Carlo simulations (Han et al., 2008; Jagielska et al., 2008; Kim et al., 2009; Leaver-Fay et al., 2013; Lin & Head-Gordon, 2011; Misura & Baker, 2005; Ovchinnikov et al., 2018; Song et al., 2013), physics-based

potentials (Chen & Brooks, 2007; Fan & Mark, 2004; Gront et al., 2012; Ishitani et al., 2008; Jagielska et al., 2008; Kannan & Zacharias, 2010; Lin & Head-Gordon, 2011; Summa & Levitt, 2007), and MD simulations (Chen et al., 2008; Chen & Brooks, 2007; Fan & Mark, 2004; Feig, 2016; Gront et al., 2012; Ishitani et al., 2008; Kannan & Zacharias, 2010; Khoury et al., 2014, 2017; Lee et al., 2016, 2001; Lindorff-Larsen et al., 2010, 2011; Mirjalili et al., 2014; Mirjalili & Feig, 2013; Raval et al., 2012; Shuid et al., 2017; Zhu et al., 2008) are used in the sampling stage to generate near-native conformations.

Automated server-based refinement methods rely on the optimisation of side-chain and the minimisation of the structural energy utilising the knowledge of protein structures (Khoury et al., 2017; MacCallum et al., 2009, 2011; Shuid et al., 2017). Automated server-based approaches are also practical and scalable, but their changes to the initial predicted models may be seen as minor, or more conservative, in comparison with the less automated more intensive methods (MacCallum et al., 2009, 2011). The fully automated approaches performed much better in the earlier CASP experiments (CASP8, and CASP9), compared to other manual approaches as MD-based methods were not successful at directing the generation of 3D models towards the native basin (Khoury et al., 2014, 2017; MacCallum et al., 2009, 2011). It should be noted that dramatic structural deviations from the native basin have not been observed in the models generated by the automated approaches as much as in the manual sampling methods. Nevertheless, the more risk-averse approaches have not shown significant performance in terms of improving the quality of the poorly predicted initial structures, where there is plenty of room for improvement (Feig, 2017; Gront et al., 2012; Hovan et al., 2018; Modi & Dunbrack, 2016; Nugent et al., 2014). Despite the early relative successes of the automated server-based refinement programs, in the more recent CASP experiments they were less successful compared to the "human" refinement methods, which often use intensive MD approaches (Feig, 2017; Gront et al., 2012; Hovan et al., 2018; Modi & Dunbrack, 2016; Nugent et al., 2014).

Non-server-based highly CPU intensive approaches are based on the generation of models running MD simulations utilising physics-based force fields, and smart restraints by taking advantage of parallel computing processing units (GPUs) and/or CPUs (Feig, 2017; Heo & Feig, 2018b; Hovan et al., 2018; Modi & Dunbrack, 2016; Nugent et al., 2014). MD-based protocols allow for the

elimination of atomic clashes, the examination of the molecular geometries and the usage of the combination of templates by way of force fields (Feig, 2017). Such methods have also reached an important stage in terms of using the knowledge of the available structures in CASP experiments (Chen & Brooks, 2007; Feig & Mirjalili, 2016; Jagielska et al., 2008; Joo et al., 2007; Mirjalili et al., 2014; Mirjalili & Feig, 2013).

The Shaw group primarily developed an MD-based approach with the usage of a physics-based potential, and this protocol was also used for the refinement of CASP9 targets (Adiyaman & McGuffin, 2019; Lindorff-Larsen et al., 2011, 2012; Raval et al., 2012). However, the simulation time was found to be very long (100 μs) for a target, and flaws in the force field lead to substantial structural deviations from the native basin (Adiyaman & McGuffin, 2019; Lindorff-Larsen et al., 2011, 2012; Mirjalili & Feig, 2013; Raval et al., 2012).

The Feig group also developed an MD-based approach using C-alpha restraints to avoid structural deviations of refined models, the CHARMM force field; and an accuracy estimation to filter decoys (Chen & Brooks, 2007; Feig, 2017; Feig & Mirjalili, 2016; Mirjalili et al., 2014; Mirjalili & Feig, 2013). The method developed by Feig group was firstly tested in CASP10 and found to be the most successful refinement method due to the usage of an improved version of the force field, C-alpha restraints, and ensemble averaging stage in explicit solvent conditions (Chen & Brooks, 2007; Feig, 2017; Feig & Mirjalili, 2016; Mirjalili et al., 2014; Mirjalili & Feig, 2013). Nevertheless, the MD-based approach also requires highly intensive computational resources (75,000 core hours,12 days on 256 cores on average). This length of the simulation time may not be feasible for large-scale structures and refinement pipelines (Chen & Brooks, 2007; Feig, 2017; Feig & Mirjalili, 2016; Mirjalili et al., 2014; Mirjalili & Feig, 2013; Nugent et al., 2014).

MD-based methods are a practical approach in case of a low number of small protein targets, but the practicality and the success rate decrease when the number of targets and/or model size increases. A high number of protein structures brings with it a high computational cost for MD-based approaches, and so applying the approach routinely to multiple targets and models would be highly time-consuming and impractical. However, the increasing potential of optimised physics-based force fields, parallel computing on the GPU and CPU and smart constraints, enables more

effective MD-based refinement protocols to become practical (Adiyaman & McGuffin, 2019; Best et al., 2008, 2012; Feig, 2017; Hovan et al., 2018; Huang et al., 2017; Lindorff-Larsen et al., 2010; Maier et al., 2015; Mirjalili & Feig, 2013; Robertson et al., 2015). MD-based approaches are widely used by prediction groups participating in recent CASP experiments due to these improvements, and eight of the top ten groups attended were using MD-based approaches in CASP12 (Best et al., 2008, 2012; Hovan et al., 2018; Huang et al., 2017; Lindorff-Larsen et al., 2010; Maier et al., 2015; Mirjalili & Feig, 2013; Robertson et al., 2015). There is still a growing need to have faster and more consistent methods, which will enable the practical refinement of a large number of protein models. The accuracy of the force fields for simulating the structures is also an important factor to move 3D models towards experimental accuracy because force fields are used in MD-based protocols to determine atomic interactions in 3D models (Feig, 2017; MacKerell et al., 2001, 2004).

Chemistry at Harvard Macromolecular Mechanics (CHARMM) c22/CMAP (MacKerell et al., 2004) and c36 (Best et al., 2012) and the AMBER ff14SB (Maier et al., 2015) and 12SB (Ovchinnikov et al., 2016; Research Computing Documentation contributors, 2018) force fields are the most popular force fields used in the MD-based protocols to sample 3D models (Cheng et al., 2017; Heo & Feig, 2018b; Khoury et al., 2014; Ovchinnikov et al., 2018). The force field parameters are trained using the knowledge of native structures to model the structure. However, they still need further improvements in the parameterisation of potential energy functions to accurately simulate the interactions in the structures. Due to flaws in force fields, refinement processes may not always direct the models towards the native state (Adiyaman & McGuffin, 2019; Feig, 2017; Jagielska et al., 2008; Mirjalili et al., 2014; Mirjalili & Feig, 2013; Summa & Levitt, 2007).

Imperfect force fields may cause structural deviations from the native basin (Jagielska et al., 2008; Summa & Levitt, 2007). To avoid such deviations, different restraint strategies are applied to guide the MD-simulations towards the native basin (Mirjalili et al., 2014; Mirjalili & Feig, 2013). Unrestrained MD-based sampling strategies also result in quick deviations away from the native state (Chen & Brooks, 2007; Hovan et al., 2018; Mirjalili & Feig, 2013; Park et al., 2012; Raval et al., 2012; Summa & Levitt, 2007). It should be noted that the application of the restraint may

also limit the conformational space for the structures. Applying strong restraints on the initial structure may not allow as much improvement in the global quality as would be needed in order to reach near-experimental accuracy (Chen & Brooks, 2007; Hovan et al., 2018; Mirjalili & Feig, 2013; Park et al., 2012; Raval et al., 2012; Summa & Levitt, 2007). For this reason, the determination of the magnitude of the restraint force has been a crucial parameter for the extent of MD simulations (Chen & Brooks, 2007; Feig, 2016; Feig & Mirjalili, 2016; Hovan et al., 2018; Mirjalili et al., 2014; Mirjalili & Feig, 2013; Park et al., 2012; Raval et al., 2012; Summa & Levitt, 2007).

Different prior knowledge has been utilised to restrain either the whole structure, or specific regions, during the MD simulations to prevent 3D models from undesired deviations. The determination of which regions to restrain and which to leave unrestrained is likely to be a major determinant of success (Cao et al., 2003; Feig, 2017; Ishitani et al., 2008; Xu & Zhang, 2011). Various restraint strategies have made significant progress in successive CASP experiments to generate improved 3D models compared to the initial structures (Cao et al., 2003; Feig, 2017; Ishitani et al., 2008; Liu et al., 2018; Seemayer et al., 2014; Xu & Zhang, 2011).

1.7.2 Scoring Approaches in the Refinement Pipelines

A significant part of refinement pipelines is the identification of the most improved models in comparison with the initial structure from among tens or hundreds of 3D models generated by the sampling approaches. The similarity between alternative 3D models that are sampled in the refinement pipeline is quite high due to the same initial structural properties. Therefore, the consistent selection of the most near-native conformations remains elusive using current approaches, such as energy functions and quality estimation tools (Adiyaman & McGuffin, 2019; Alford et al., 2017; Chen & Brooks, 2007; Feig, 2017; Feig & Mirjalili, 2016; Gront et al., 2012; Kumar et al., 2015; Larsen et al., 2014; Lee et al., 2016; Lu & Skolnick, 2003; Mirjalili & Feig, 2013; Olson & Lee, 2014; Park et al., 2015; Rykunov & Fiser, 2010; Stumpff-Kane et al., 2007; Yang & Zhou, 2008a; Zhang et al., 2011; Zhang & Skolnick, 2004a, 2004b).

The concept of energy functions is based on Anfinsen's hypothesis which states that the native state of a protein structure has the lowest free energy (Anfinsen, 1973; Heo & Feig, 2018b). Therefore, it is assumed that the conformations with lowest potential/free energies are near the native state. Different energy function-based methods including DFIRE (Yang & Zhou, 2008a), DDFIRE (Yang & Zhou, 2008a), RW+ (Zhang & Zhang, 2010), and Rosetta energy functions (Adiyaman & McGuffin, 2019; Alford et al., 2017; DiMaio et al., 2009; Feig, 2017; Feig & Mirjalili, 2016; Tyka et al., 2011; Zhang & Zhang, 2010) have been used to identify the most-native like conformations generated in refinement pipelines (Adiyaman & McGuffin, 2019; Alford et al., 2017; DiMaio et al., 2009; Feig, 2017; Feig & Mirjalili, 2016; Tyka et al., 2011; Zhang & Zhang, 2010). Nevertheless, no energy function can be used to achieve clear and consistent selection of the improved models (Adiyaman & McGuffin, 2019; Alford et al., 2017; DiMaio et al., 2009; Han et al., 2008; Heo & Feig, 2018b; Kim et al., 2009; Kuhlman et al., 2003; Kuhlman & Baker, 2000; Leaver-Fay et al., 2013; Mirjalili & Feig, 2013; Park et al., 2016; Rohl et al., 2004; Tyka et al., 2011; Yang & Zhou, 2008a, 2008b; Zhou & Zhou, 2002)

Although quality estimation tools, such as ProQ (Wallner & Elofsson, 2003), ProQ2 (Uziela & Wallner, 2016), SELECTpro (Randall & Baldi, 2008), and ModFOLD6 (Maghrabi & McGuffin, 2017), have been designed to identify the most native-like conformations sampled by TBM and FM approaches, they have been also used to select improved models generated by the sampling approaches in refinement pipelines (Adiyaman & McGuffin, 2019; Cheng et al., 2017; Chopra et al., 2010; Shuid et al., 2017). While these tools are often successful at identifying the best 3D models from among alternative tertiary structure predictions servers, they have not shown as consistent performance when they are used to select the most improved refinement models. This is perhaps due to fact that the discrimination of the 3D refinement models is a much harder problem, due to the very high similarity among the alternative 3D models that are sampled in the refinement pipeline (Kryshtafovych et al., 2005, 2014; Larsson et al., 2009; McGuffin et al., 2013)

1.8. The Refinement Category in CASP

The CASP refinement category was introduced to encourage prediction groups to further increase the accuracy of predicted 3D models after they had been generated in the regular prediction category, and it has been operating as a separate category since CASP8 (MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016). The best-predicted 3D models (generated via either TBM or FM approaches) were selected as the refinement targets by CASP assessors. The aim was to improve the local and global quality of these and to move the models as close to the native structures as possible (MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). Increasing the accuracy of the best-predicted 3D models has been increasingly challenging as the 3D models may have already been once-refined in the prediction pipelines of the individual servers from which the models originate (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). Increasing the accuracy of the refinement targets is an already difficult problem, so it is arguably one of the hardest categories of the CASP experiment.

Up to five refined models can be submitted in preference order by each prediction group and then evaluated by CASP assessors (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). Although the submission of five models enables testing of different approaches for each model, 5 models may not be adequate for a proper evaluation of the sampling and scoring approaches (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). For instance, MD-based sampling approaches may generate hundreds of models, thereby it is essential to identify the most improved models among hundreds of models. If the prediction groups did not manage to select improved models, they would be failed even if they generated improved models. Therefore, we suggest that the sampling and scoring approaches should be assessed separately under the CASP refinement category as in Figure 1.2 (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014) .

Assessment criteria of the refinement protocols are established on the accuracy of backbone and side-chain contacts and the analysis of atomic clashes and geometry by using various

measurements determined by CASP (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). Since CASP9, assessors have been providing some useful information about the regions within the starting models that should be focused on, in order to assist prediction groups. However, in reality for automated approaches such information would not be available (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014).

There has been a significant increase in the success of refinement strategies since the category was introduced in CASP7. Although the success rate was very low in CASP8 and CASP9, the development of MD-based approaches has enabled considerable progress since CASP10 (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). In addition, the MD-based refinement protocols have outperformed the automated approaches since CASP10 (Adiyaman & McGuffin, 2019; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014). The numbers of targets and groups increased dramatically since CASP11, from 12 to 51 targets and 24 to 47 prediction groups participating in CASP14 (Adiyaman & McGuffin, 2019; Hovan et al., 2018; MacCallum et al., 2009, 2011; Modi & Dunbrack, 2016; Nugent et al., 2014).

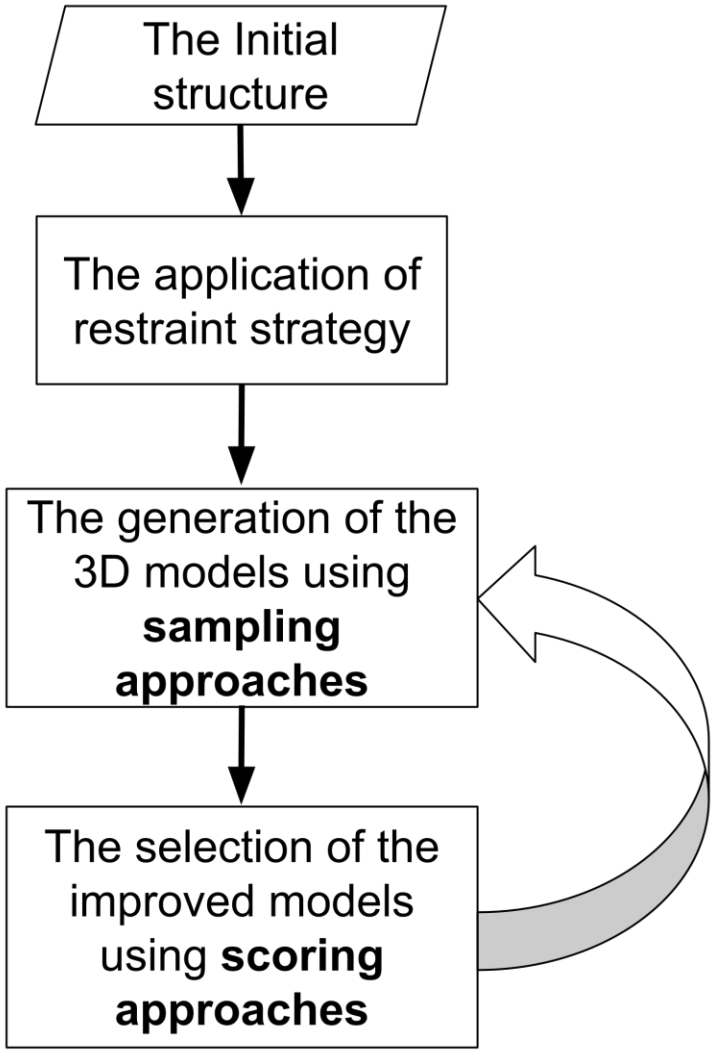

Figure 1. 2 Flowchart summarising a typical refinement protocol.
The restrained or unrestrained regions are determined before the sampling strategies. The 3D models are also generated using different sampling approaches. After the generation of the 3D models, the 3D models are ranked using various scoring methods for the selection of improved models compared to the initial structure. The sampling and scoring methods can also be applied in an iterative cycle.

1.9 Refinement Tools and Webservers

Different refinement tools and web servers have been developed to increase the accuracy of the initial structures for biologists. Feig (Feig, 2017) and Adiyaman (Adiyaman & McGuffin, 2019) have also provided a comprehensive review of the protocols. A few of the best freely available methods were described in Table 1.3.

Name	Method Definition	URL
PREFMD (Heo & Feig, 2018a)	MD-based protocol tested in CASP11	http://feiglab.org/prefmd
GalaxyRefine (Heo et al., 2013)	Rebuilding side chains and structure relaxation by molecular dynamics simulation	http://galaxy.seoklab.org/refine
KoBaMIN (Rodrigues et al., 2012)	Energy minimization using knowledge-based potential	http://csb.stanford.edu/kobamin
Princeton_TIGRESS 2.0 (Khoury et al., 2017)	Monte Carlo and molecular dynamics simulations	http://atlas.engr.tamu.edu/refinement/
ModRefiner (Xu & Zhang, 2011)	Molecular dynamics simulation using a composite of physics- and knowledge-based force field	http://zhanglab.ccmb.med.umich.edu/ModRefiner
3DRefine (Bhattacharya et al., 2016; Bhattacharya & Cheng, 2013a, 2013b)	Optimisation of hydrogen bonds and energy minimisation	http://sysbio.rnet.missouri.edu/3Drefine/
ReFOLD (Shuid et al., 2017)	A modest MD-based protocol and the usage of quality estimation tool for the selection	http://www.reading.ac.uk/bioinf/ReFOLD/

Table 1. 3 Summary list of popular refinement web servers.

1.10. Project Aims and Objectives

The major goal of the project is the development of an efficient refinement pipeline with the integration of quality estimation, protein-ligand binding site prediction and contact prediction tools. An overview of protein structure prediction approaches including the prediction of 3D models, the assessment of the predicted 3D models and the model refinement stage has been presented. The ReFOLD (Shuid et al., 2017) protocol including the hybrid combination of

i3Drefine, MD-based protocol, and ModFOLD6 (Maghrabi & McGuffin, 2017) is a promising refinement strategy. The MD-based protocol is an important and distinctive part of the refinement process, and it enables simulation of large biological systems and can be run on normal desktop computers along with supercomputers (Phillips et al., 2005). The McGuffin group has managed to develop ReFOLD to refine predicted protein structures with far less computational effort (Shuid et al., 2017). Our aim is to develop ReFOLD in order to get higher accuracy in refined models by using the different restraint strategies, so that it may help direct the models towards the native basin, so that models will have great utility in further *in silico* studies.

1.10.1 The Restraint Strategy Based on the Local Quality Estimation

For the first objective of the study, we aimed to use the per-residue accuracy score produced by ModFOLD6 to guide the original MD-based protocol of ReFOLD by applying a threshold according to the distribution of the per-residue accuracy score. The predicted per-residue accuracy score of a 3D model is precious information and could be better utilised to help us decide which regions of the predicted models require the most refinement. The local quality assessment guided restraint strategy based on the per-residue accuracy score was developed to avoid structural deviations from the native basin. The performance of the local quality assessment guided restraint strategy was also compared with the original MD-based protocol of ReFOLD using many scoring methods. In Chapter 2, considerable progress is reported due to the usage of this new local quality assessment restraint strategy.

1.10.2 The Application of the Gradual Restraint Strategy Based on the Local Quality Estimation

Chapter 3 focuses on the development and investigation of the performance of the local quality assessment guided MD-based protocol. The local quality assessment guided restraint strategy was upgraded using ModFOLD7 (Maghrabi & McGuffin, 2019) for the selection of the improved models and providing the per-residue accuracy score. The performance of the upgraded version in

the CASP13 experiment was also analysed, presenting CASP13 official results. The local quality assessment guided strategy has consolidated the performance of our refinement protocol as in the top ten approaches in CASP13 among roughly a hundred prediction groups.

After the investigation of CASP13 performance, a gradual restraint strategy based on the per-residue accuracy score was developed as the application of a threshold based on the per-residue accuracy score was not found to be practical for large targets. A detailed comparison of the local quality assessment guided restraint strategy developed in CASP13 with the gradual restraint strategy based on the per-residue accuracy score is also described in the chapter. The gradual restraint strategy was also used to refine SARS-2-CoV protein structures as a part of the CASP Commons COVID-19 initiative. The performance of the gradual restraint strategy was investigated in terms of generating the improved models according to CASP initial results, and the gradual restraint strategy has provided a significant proportion of the top ten 3D models for the ten SARS-2-CoV targets. The development of the gradual restraint strategy also enhanced the competitiveness of our refinement pipeline.

1.10.3 The Binding Site-Focused Restraint Strategy

The FunFOLD3 (Rhizobium, 2013; Roche & McGuffin, 2016a) server was developed by McGuffin group to predict protein-ligand interactions. For the fourth chapter, we aimed to increase the accuracy of the binding sites predicted by FunFOLD3 (Roche & McGuffin, 2016a) according to the observed scores such as GDT-HA (Zhang & Skolnick, 2005), the BDT (Roche et al., 2010) and MCC scores. The binding site-focused MD-based protocol, which is similar to the local quality assessment guided MD- based protocol developed in Chapter2, was proposed for the refinement of the predicted binding sites. The effectiveness of the binding site-focused MD-based protocol was also investigated using CASP12 and CASP13 targets in terms of improving the quality of the predicted binding sites. The integration of the binding site-focused MD-based protocol with the FunFOLD server may also provide highly accurate binding site predictions.

1.10.3 The Contact-Assisted MD-Based Protocol for the Refinement of Protein 3D Models

The predicted residue-residue contacts have also made significant improvements to protein structure prediction strategies, particularly during the CASP13 experiment. This valuable information has helped to increase the accuracy of the predicted 3D models. Furthermore, accurate information regarding predicted pairwise distances might also provide very valuable guidance for a more consistent refinement. Highly accurate residue-residue contact predictions, using the DeepMetaPSICOV (Kandathil et al., 2019a) data from CASP13, were used to determine gradual restraints according to the distribution of the Contact Distance Agreement (CDA) scores (Maghrabi & McGuffin, 2017) These gradual restraints were applied during the MD simulations to improve the quality of the predicted structures to meet our fourth objective. The performance of the contact-assisted MD-based protocol was also compared with the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol in Chapter 5.

A unique combination of the contact-assisted and gradual restraint strategies was used to increase the accuracy of the CASP14 targets in our refinement pipeline. The comparison of ModFOLD6, ModFOLD7 and ModFOLD8 in terms of the selection of the improved models generated by the combined protocols was also investigated for further improvements in the final part of the analysis.

Chapter 2 The Usage of Local Model Quality Estimates to Guide the MD-Based Protocol

Work presented in this chapter has been submitted in the following paper:

Recep Adiyaman and Liam James McGuffin, 2020. Using Local Protein Model Quality Estimates to Guide a Molecular Dynamics Based Refinement Strategy. Submitted to Springer Nature.

2.1 Background

The prediction of 3D models via TBM and/or FM methods may generate hundreds of 3D models in alternative conformations. The global scores produced by the Model Quality Assessment Programs (MQAPs) can be used for ranking models but may not always be adequate for the assessment of the decoys and the identification of the most native-like structures, especially in cases where models are very close in structure, such as alternative refinement models. For this reason, the local (or per-residue) quality assessment scores are also required in order to discover more about the finer details of predicted models, and many model quality estimation tools are also able to produce the per-residue accuracy scores besides the global scores.

2.1.1 The Local Quality Estimation of 3D Models

Early single-model based quality assessment programs producing the per-residue accuracy score were inaccurate in comparison with the actual distance among residues in the observed structures (McGuffin, 2010; Wallner & Elofsson, 2006). The clustering-based methods, which rely on comparing many varied models, were found to be more successful in producing the per-residue accuracy score compared to the single-model based approaches (McGuffin, 2010). Therefore, in the CASP experiments, most of the higher performance quality assessment programs producing the per-residue accuracy scores have been based on the clustering-based approach, because in CASP multiple varied models are always available. Pcons-local was one of the initial programs developed to calculate the per-residue accuracy scores (Wallner & Elofsson, 2006). The ModFOLDclust method (McGuffin, 2009, 2008c) has been also a leading per-residue clustering-based program in CASP experiments, and its calculation relied on the S-score (see below). The ModFOLDclust method remains an important component method for the per-residue accuracy score calculation in the last version of ModFOLD server (Maghrabi & McGuffin, 2019, 2017).

The S-score used by ModFOLDclust is defined using the following equation (McGuffin, 2009, 2010; McGuffin & Roche, 2010):

$$S_i = \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2}$$

Where S_i is the S-score for residue *i* in a predicted model, d_i is the Euclidean distance between residues according to the TM-score (Zhang & Skolnick, 2004b) superposition and d_0 is a distance threshold (3.9Å). The *mean* S-score is calculated:

$$S_r = \frac{1}{N - 1} \sum_{a \in A} S_{ia}$$

Where S_r is the residue accuracy for the assessed model, N is the number of assessed models, A is the set of alignments and S_{ia} is the S_i score for a residue in a structural alignment (a). The size of set A is also equal to N-1 (McGuffin, 2009, 2010; McGuffin & Roche, 2010). The *mean* S-score is then used to convert to the per-residue error score or the distance of the residue from the native structure in Ångströms:

$$d_r = d_0 \sqrt{\left(\left(\frac{1}{S_r}\right) - 1\right)}$$

An upper limit of 15 Å is applied for d_r. The calculated per-residue accuracy score are inserted to the B-factor columns of each set of ATOM records, and the files including the per-residue accuracy score in the B-factor column in the PDB format are available for download on the ModFOLD server result page (Maghrabi & McGuffin, 2017; McGuffin, 2009, 2010; McGuffin et al., 2013; McGuffin & Roche, 2010, 2011; Roche et al., 2014).

2.1.2 The ModFOLD Server

The ModFOLD method was firstly introduced by the McGuffin group in CASP7, and the approach included a neural network with the combination of ProQ (Wallner et al., 2003), MODCHECK (Pettitt et al., 2005) and ModSSEA methods (McGuffin, 2007). The original ModFOLD server accommodated two different options of methods: ModFOLD which was able to produce a global score for a single model, and ModFOLDclust which enabled local and global scoring for multiple

models (McGuffin, 2008c). ModFOLDclust was the best quality assessment program in producing global scores in CASP8 (McGuffin, 2010).

The quasi-single approach was later pioneered by the McGuffin group to improve versions 3 and 4 of the server, in order to assess both single and multiple models (McGuffin et al., 2013). The quasi-single model approach worked by making use of reference sets of models, which were generated from the sequence by the IntFOLD server (McGuffin et al., 2015, 2019), for comparison with the model to be scored. The idea being that you could gain the accuracy of clustering-based methods, but you only needed to submit a single submitted model, rather than many varied models. Successive MoldFOLD versions have continued to use the quasi-single approach and have been the top performers program in the recent CASP experiments (Cheng et al., 2019). The ModFOLD6 server was released as a novel hybrid combination of pure-single and quasi-single methods, and the approach has succeeded in distinguishing accurate models in the CASP12. The 3 different ModFOLD6 variants (ModFOLD6, ModFOLD6_rank and ModFOLD6_cor) by the McGuffin group were among the top few in the local scoring methods, and they were also among the top 3 prediction groups for global scoring in CASP12 (Maghrabi & McGuffin, 2017; McGuffin, 2009, 2010; McGuffin et al., 2013; McGuffin & Roche, 2010, 2011; Roche et al., 2014).

The accuracy of the local scores produced by ModFOLD6 was increased by using six alternative local scoring methods, and the methods were also combined by using a neural network in the ModFOLD6 protocol (Maghrabi & McGuffin, 2017). The first method was a new pure-single method based on the Contact Distance Agreement (CDA) between the contact predictions, which were made using the MetaPSICOV method (Jones et al., 2015), and the contacts measured by the Euclidean distance between residues in the 3D models. The second component of the ModFOLD6 local scoring protocol was utilising the Secondary Structure Agreement (SSA) between the residues in the secondary structures predicted by using PSIPRED and those in secondary structures of the 3D model according to Dictionary of Secondary Structures of Proteins (DSSP) (D. W. A. Buchan et al., 2013; Kabsch & Sander, 1983; Maghrabi & McGuffin, 2017). Local ProQ2 scores were also added as a part of the local scoring method (Uziela & Wallner, 2016). Another component method was the generation of the ModFOLD5_single local QA score, which comparing each model against the reference set of 130 models built by IntFOLD4, using quasi-

single approach (Maghrabi & McGuffin, 2017; McGuffin et al., 2013). The ModFOLDclustQ_single local quality score was also calculated in comparison with the reference IntFOLD4 set, but this time employing the local Q-score approach (McGuffin, 2008c; McGuffin & Roche, 2010; Roche et al., 2012b; Torchala et al., 2013) and included in the combination of local quality estimation methods. Finally, the Disorder B-factor Agreement (DBA) score was additionally generated by a new quasi-single model method. The DBA score is a function of the agreement between the disordered residues predicted by DISOPRED3 (Jones & Cozzetto, 2015) and the per-residue accuracy score from ModFOLD5_single (Jones & Cozzetto, 2015; Maghrabi & McGuffin, 2017). A simple multilayer Neural Network (NN) was used to combine the six local scoring methods and then produce the final ModFOLD6 per-residue score. The *mean* per-residue accuracy scores produced by the six local scoring methods were also part of the calculation of the global score in ModFOLD6 (Maghrabi & McGuffin, 2017).

2.1.3 Pcons, ProQ, ProQ2 and ProQ3

Many MQAPs are available to score predicted models with local and global approaches. Pcons is one of the local quality assessment programs improved by means of a neural network, and the clustering-based methods were used to estimate the accuracy of a set of alternative models (Lundström et al., 2001; Wallner & Elofsson, 2007). The assessment of predicted models with Pcons was more accurate in comparison with the single-model methods in CASP7 (Larsson et al., 2009). The ProQ method (which originated from the same group as the Pcons approach) was able to produce global and local (per-residue) quality assessment scores, and ProQ was one of the more successful single model approaches at recognising the correct models in CASP7 and CASP8. Machine learning and descriptive features of protein modelling were also utilised to get an accurate scoring function (Wallner & Elofsson, 2003, 2007). A support vector machine (SVM) was later used in the development of the next version, ProQ2, and the approach provided the evaluation of local and global error, using a single-model approach. The ProQ2 was integrated with Rosetta and performed well in CASP11 (Uziela & Wallner, 2016). ProQ3 was the subsequent upgrade, this time deploying a deep neural network instead of the super vector machine, which has performed better than other machine learning approaches tested. The approach has been improved by

combining different Rosetta energy functions implemented in ProQRosCen, ProQRosFA and ProQ2. Significant progress has been observed in the assessments of the same models via ProQ3 compared to ProQ2 (Uziela, Hurtado, Wallner, & Elofsson, 2016; Uziela, Shu, Wallner, & Elofsson, 2016)

2.1.4 Estimation of Model Accuracy in CASP Experiments

The blind assessment of MQAPs was first conducted in the QA category introduced in CASP7, with the aim of encouraging the development of model quality assessment methods (Cozzetto et al., 2007; Moult et al., 2007). In the QA category, the predicted 3D models generated by tertiary structure prediction servers for each CASP target are then submitted to MQAP servers and standalone methods for quality scoring, prior to the availability of the solved structures (Cozzetto et al., 2007; Kryshtafovych et al., 2011; Roche et al., 2014). In CASP12, based on a benchmark of models for 70 protein targets, 42 MQAPs were evaluated in two subcategories: 1) QAglob: used for global assessments, 2) QAloc: introduced for the per-residue accuracy. 42 prediction groups produced scores in the QAglob category and 24 prediction groups produced scores in the QAloc category (Kryshtafovych et al., 2017). These categories have helped in evaluating performance of participating model quality assessment groups, and a significant progress was observed in the CASP12, particularly in the usage of the deep learning and residue-residue contacts. In CASP13, 52 prediction groups including 41 automated servers were benchmarked in the quality estimation category. Again, in CASP13, the further development of the deep learning-based methods along with the utilisation of contact predictions has revealed the potential of the quality estimation tools for the identification of the highly accurate 3D models.

2.1.5 ReFOLD

The ReFOLD method was developed to refine predicted 3D models by the McGuffin group as a fully automated refinement server, and the program was firstly tested in CASP12. ReFOLD is a unique hybrid method including a combination of rapid iterative refinement with i3Drefine,

scalable molecular dynamics with Nanoscale Molecular Dynamics (NAMD) (Phillips et al., 2005), and ModFOLD6 (Maghrabi & McGuffin, 2017) have been used for the refinement of protein structures (Shuid et al., 2017). The approach aims to fix errors identified by ModFOLD6 in predicted 3D models, with modest computational resources required for the simulation of the protein structures, compared with other MD-based protocols (Maghrabi & McGuffin, 2017). The refinement process includes three steps to increase the local and global quality of the models to generate native-like structures. The first step is using a rapid iterative method by way of i3Drefine in 20 refinement cycles (Bhattacharya et al., 2016). After the iterative refinement, a MD-based (NAMD) protocol, inspired by that of Feig and Mirjalili (Feig & Mirjalili, 2016), is used to refine each predicted model. The second protocol includes the application of C_α restraints, the ensemble averaging of the models and interpolation between starting and refined models. NAMD is designed to do parallel high-performance simulations using AMBER (Götz et al., 2012; Lindorff-Larsen et al., 2010; Research Computing Documentation contributors, 2018) and CHARMM (Best et al., 2012; Huang et al., 2017; MacKerell et al., 2001) functions and force fields, and the simulation program is primarily written in the C++ language and uses Charmm++ (MacKerell et al., 2001; Maier et al., 2015; Phillips et al., 2005). NAMD can be used in parallel platforms that include GPUs, and it can even be run on a desktop, as a low-cost approach (Phillips et al., 2005). In the last step, the 3D models generated by the two protocols are ranked by ModFOLD6 according to global quality scores. It is also possible to assess the local quality of all 3D models via ModFOLD6 (Maghrabi & McGuffin, 2017; Shuid et al., 2017).

The complex protocol can be run on the ReFOLD server via a simple web form and the results produced are also available via a user-friendly web interface. The only inputs required by the server are the amino acid sequence and a 3D model (in PDB format) of the target (Shuid et al., 2017). The output from the server showed a high-ranked performance and significantly improved the global quality scores of the models submitted by the McGuffin group in the TS category of CASP12 (Shuid et al., 2017).

2.2 Aims and Objectives

MD-based methods are the most preferred approach by many prediction groups to refine predicted models. However, the approaches are highly computationally intensive, time-consuming, and often may result models drifting away from the native structure due to the lack of optimised force fields. Although, our group has managed to refine 3D models spending modest computational resources at the MD simulation stage after the development of the ReFOLD, undesired structural deviations from the native basin were still observed for the refinement of the of predicted 3D models, especially for TBM structures in CASP12. (Shuid et al., 2017). To avoid such deviations, different restraints could be applied in MD-based methods, to mitigate the effects of the imperfect force fields, but the problem is deciding which part of the structure should be restrained.

The local quality estimation of model accuracy was proposed as a possible future guide for the original MD-based protocol of ReFOLD (Shuid et al., 2017), so that it may help direct the models towards the native structure. The ModFOLD6 method developed by the McGuffin group has been consistently ranked among the top MQAPs in CASP experiments in terms of accurately predicting the per-residue accuracy score in 3D models. Here, we aim to make use of the per-residue accuracy score identified by ModFOLD6, which are typically shown as the predicted C-alpha distances from the native structure and stored in the B-factor column of PDB formatted model files. It is also proposed that the distances can be used to determine and select poorly predicted regions, which may then be further refined to improve the quality of predicted 3D models.

In the first stage, the study includes the identification of the per-residue accuracy score via ModFOLD6, and the determination of a threshold based on the per-residue accuracy score. The determined threshold is then applied to the predicted models in order to restrain the well-predicted regions, so that only the poorly predicted regions in the 3D models will be targeted for refinement, during the subsequent MD simulations. The 3D models are then refined with the application of the new local quality assessment guided MD-based protocol

The last step involves the detailed comparison of the new local quality assessment guided MD-based protocol developed here versus the MD-based protocol used in the original ReFOLD pipeline, by using different metrics such as the GDT-HA (Zhang & Skolnick, 2005) and Molprobity (Chen et al., 2010; Davis et al., 2004) scores. Lastly, the performance of ModFOLD6 was then analysed in terms of its ability to select the best-refined 3D model.

2.3. Materials and Methods

2.3.1 Data Collection

The CASP12 regular (T0) and refinement (TR) targets were used to test the newly developed local quality assessment guided MD-based protocol. 42 refinement targets and 60 regular targets are publicly available from the CASP website (http://predictioncenter.org/download_area/CASP12/). Both the number of residues and starting model quality (high accuracy score - GDT-HA) of the refinement targets varies widely from 54 to 396 residues and from 0.23 to 0.76 respectively (Hovan et al., 2018). We also used the models refined by the original MD-based protocol of ReFOLD during CASP12 to compare with the new local quality assessment guided MD-based protocol developed in this study. The refined models were downloaded from our ReFOLD server results pages for each CASP12 target.

2.3.2 Computational Design

The computational protocol consists of the three main stages: i) the identification and the determination of the threshold (predicted C-alpha distance from native structure) based on the per-residue accuracy score using ModFOLD6 for each target; ii) the application of the local quality assessment guided MD-based protocol; iii) the assessment of the local quality assessment guided MD-based protocol and its comparison with the original MD-based protocol of ReFOLD. The stages of the protocol were detailed in the flow chart (Figure 2.1*)*. Starting models were downloaded from CASP12 website (http://predictioncenter.org/download_area/CASP12/), and then were submitted to ModFOLD6

(http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD6_form.html) to get the per-residue accuracy score for each starting model. The range of the per-residue accuracy scores, reported in the B-factor (temperature factor) column of each PDB file produced by ModFOLD6, is important for determining the thresholds which were applied to the 3D models in the MD simulation stage. The per-residue accuracy scores in the models assessed by ModFOLD6 were the predicted distances in Ångströms of each residue from the native structure. (Maghrabi & McGuffin, 2017).

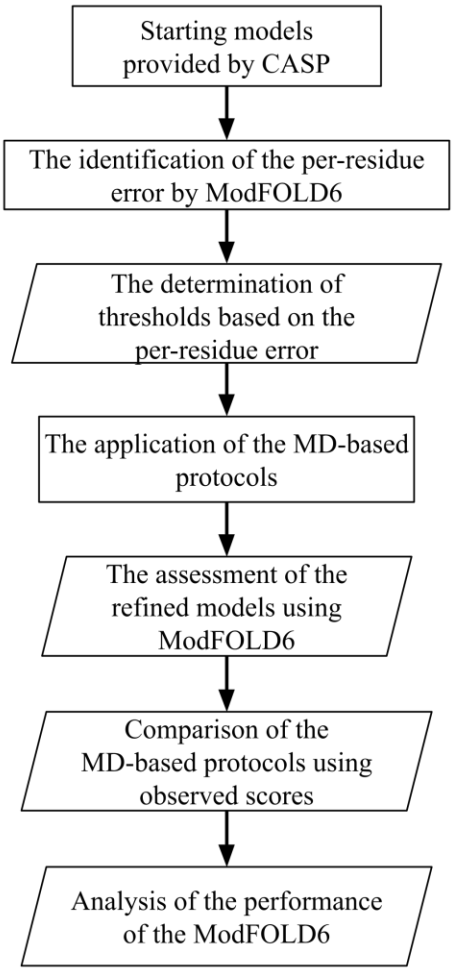

Figure 2. 1 Flow of data and methods developed in this chapter.
The starting models were submitted to ModFOLD6 to get the predicted per-residue accuracy score and then the thresholds based on the predicted per-residue accuracy score were determined to selectively refine. The starting models were refined using the new local quality assessment guided MD-based protocol. The refined models were also evaluated using observed structures and then compared with models generated by the original MD-based protocol of ReFOLD in the following step. Finally, the 3D models generated by the local quality assessment guided MD-based protocol were ranked using ModFOLD6 and the performance of ModFOLD6 was analysed in terms of selecting optimal models.

The question is how to avoid deviations from the native structure or minimise the effect of undesired deviations from the native basin in MD simulations. The use of the restraint is a widely preferred method to prevent refined models generated by the MD-based protocol from structural drifts. In this stage, the per-residue accuracy score produced by ModFOLD6 may provide a reliable guidance towards the native structure. We observed that if the predicted per-residue accuracy score is below 3 Ångströms, then those particular residues should have been well-predicted in the model, thereby further refining the residues may not be needed (or indeed, be wise, as it may lead to degradation). However, if the per-residue accuracy score is predicted to be above 8 Ångströms, then the residues are more likely to be further away from the native state, so they should be refined further in order to improve the quality of the predicted 3D model. For this reason, we decided to apply three thresholds based on the per-residue accuracy scores at 3, 5 and 8 Ångströms and we aimed to restrain the residues below each determined threshold in the MD simulation. Varying the different thresholds and then repeating the simulations, helps us to determine the optimal distance cut-off in a systematic way.

The thresholds based on the per-residue accuracy score were applied to the MD simulation as the major test variable in the study. Because of this, the same simulation parameters that were originally optimised in ReFOLD (Shuid et al., 2017) were used to control for the effect of using the local quality assessment guided restraint strategy on refined 3D models during the MD simulation. Thus, simulations were conducted using NAMD (Phillips et al., 2005), version 2.10 in GPU accelerated mode. To maintain normal cellular behaviour, the conditions were defined as a temperature of 298K and a pressure of 1 bar. The combination of the set of CHARMM22/27 force field parameters (MacKerell et al., 2001) and default TIP3P water model (Jorgensen et al., 1983) was used to simulate a water model (Feig & Mirjalili, 2016; Shuid et al., 2017). The system was also neutralised by inserting Na^+ or Cl^- ions to balance the net charge using Particle Mesh Ewald (PME) (Götz et al., 2012). The non-bonded interactions (mostly van der Waal's) were cut off by 12Å to the exclusion of bonded interactions by using CHARMM27 default parameter file with the switching distance of 10Å. (Shuid et al., 2017). Using pairlistdist function with 14Å distance between atom pairs for inclusion in pair lists enabled making the switching function more efficient. The rigidBonds functions were also used to rigidify hydrogen bonds with a 2fs timestep (Shuid et al., 2017). The system's electrostatics and temperature was calculated by PME with the

temperature control using Langevin dynamics under the NTP conditions (constant number of particles, temperature, and pressure) (Loncharich et al., 1992). The maintenance of the biological system ensured by periodic boundary conditions to rationalise the simulation process.

A weak harmonic positional restraint on all atoms below each determined threshold (based on the per-residue accuracy score) was applied with a force constant of 0.05kcal/mol/A^2. As a different application from other restraint strategies, we applied the restraint on all atoms including C-alpha as the C-alphas above the determined threshold may need to be refined. The occupancy column of each atom below the determined threshold was assigned to a value of 1 to restrain, values of 0 were assigned to indicate that the atoms should not be restrained during NAMD simulation stage (Phillips et al., 2005).

The correction of clashes and the minimisation of the system were carried out by 1000 steps in the first step of the MD-based protocol. The minimisation step was followed by the implementation of the defined MD simulation to refine each target. Four parallel simulations were run for 2 ns, making 8 ns in total for a target as in the original MD-based protocol of ReFOLD (Shuid et al., 2017). Four short trajectories (one million steps for each trajectory) were preferred rather than one long trajectory as optimized for the original version of ReFOLD (Shuid et al., 2017), and the same trajectory length was used for all MD-based protocols developed in this thesis. After the completion of the simulation run, 164 refined models were generated per target by taking a snapshot every 50 ps. The refinement protocol was performed on a machine using Intel® Core™ i7 processors and NVIDIA GeForce GTX 1070 Graphics Cards by taking advantage of GPU computing with 16GB RAM. The simulation time for a protein structure with roughly 100 residues takes almost 10 hours (Shuid et al., 2017)

2.3.3 Evaluation Methods

We used a wide range of assessment tools to evaluate the refinement approach developed here, in terms of both the global and local accuracies. We used scores based on comparisons of the refined 3D models with the observed score (GDT-HA) and an analysis-based assessment tool

(Molprobity). The TM-score tool (Zhang & Skolnick, 2004b) was used to produce GDT-HA scores to evaluate refined models by comparing them with the released observed experimental structures by CASP. We also used the Molprobity method, which is a free assessment program to score refined models via a multi-criterion chart (Chen et al., 2010). The GDT-HA and Molprobity scores were used to measure the quality of the 3D models both before and after the refinement stage. These scores have also been used by the official CASP assessors in order to evaluate the performance of the prediction groups in both the regular prediction and refinement categories (Hovan et al., 2018; Modi & Dunbrack, 2016). We also used ModFOLD6 developed by McGuffin group to evaluate our ability to assess the quality of the refined models prior to the availability of the experimental structure, which will be critical for the practical application of the approach as a whole (Maghrabi & McGuffin, 2017).

The GDT-HA score is a global score based on the multiple global positioning of C-alpha atoms using the superposition of the predicted or refined models with the observed structure (Zhang & Skolnick, 2004b). The TM-score tool uses a smaller cut off distance to calculate the GDT-HA score which allows for higher accuracy (hence, HA) in measurements (compared to the standard GDT-TS score) after the superposition. While there is a strong correlation between GDT-HA and GDT-TS scores, we preferred to use GDT-HA score which is more sensitive to evaluate the smaller changes in refined models. The GDT-HA score ranges from 0 to 1, with *higher* values indicating more accuracy (Zhang & Skolnick, 2004b).

The Molprobity score is a measurement of the physical properties and local model quality of the 3D models and it provides an all-atom contact analysis (Chen et al., 2010; Davis et al., 2004). The calculation of the Molprobity score is a non-native dependant scoring method that does not require the observed/solved structure to score predicted models. The Molprobity scores range from 0 upwards, and models with *lower* scores are more stereochemically accurate than higher-scored models (Chen et al., 2010).

For the statistical comparison of the MD-based protocols based on the GDT-HA score and Molprobity scores, Wilcoxon tests were run for the analysis of the data in Tables 2.1- 2.8, with R statistical package.

2.4. Results and Discussion

Twenty-two regular (T0) and thirty-four refinement (TR) CASP12 targets were refined using our new MD-based protocol. 29 out of the 56 refined targets were predicted by TBM, 14 targets were predicted by FM and 13 targets were predicted by both prediction methods (TBM/FM). These results were compared in three ways: (1) according to the comparison of the local quality assessment guided MD-based protocol with the original MD-based protocol of ReFOLD, (2) according to the target prediction methods (i.e., the methods used to produce the starting models), and (3) according to the performance of ModFOLD6 on refined models, which was evaluated in terms of its ability to select improved models.

Our initial aim is to develop the original MD-based protocol of ReFOLD by applying a new restraint strategy based on the per-residue accuracy score. ReFOLD was found to be less successful in the refinement of the structures predicted by TBM methods than for those using FM methods in CASP12, which we postulate was due to lack of a reliable guidance during the MD simulation (Shuid et al., 2017). Moreover, our group has also been one of the leading groups in terms of producing predicted per-residue accuracy score via the ModFOLD server in both the CAMEO and CASP experiments (Haas et al., 2018; Hovan et al., 2018; Maghrabi & McGuffin, 2017). Therefore, it is worth our while to attempt to utilise the accurate predicted per-residue accuracy scores to guide MD simulations, so as not to allow the refined models to drift from the native structure. The application of the local quality assessment restraint strategy is also summarised in Figure 2.2

The refinement of the predicted protein structures aims to bring the refined model closer to the experimental accuracy. For this reason, the GDT-HA score was used as a major measurement score to benchmark the protocols, as it relies on the C-alpha atoms superposition of the refined model with the experimentally determined structure. The Molprobity score is also used to analyse the refined models as it is calculated taking account of all-atoms, not just C-alpha atoms which are used in the GDT-HA score.

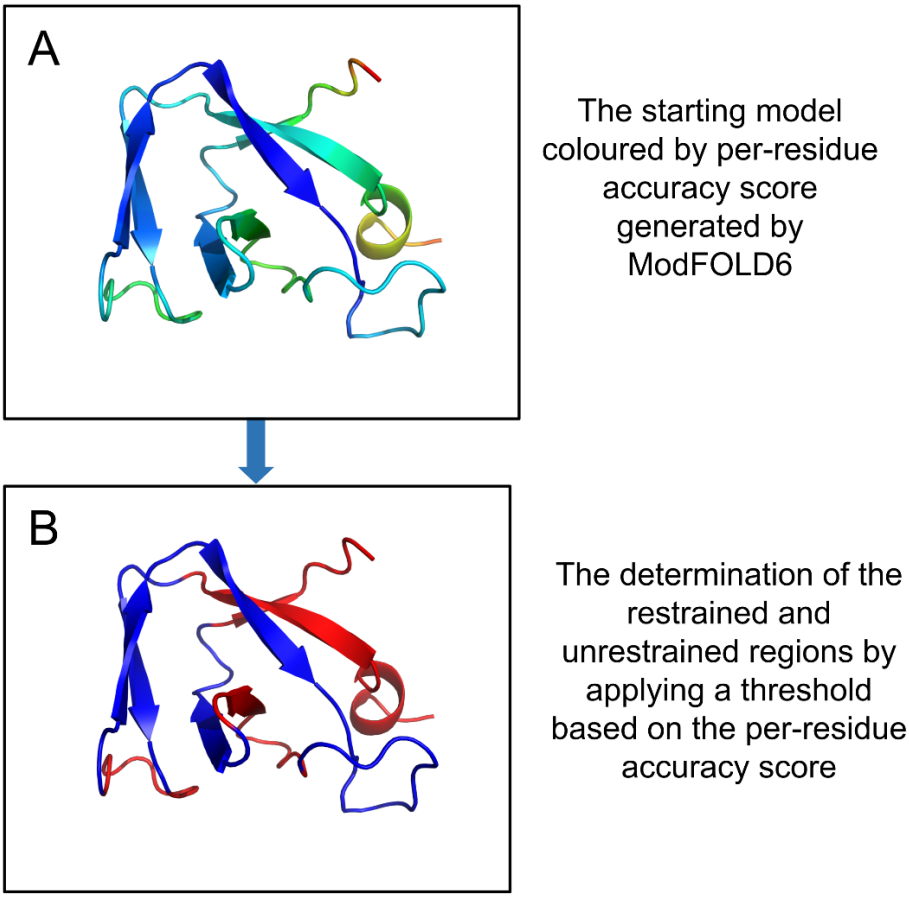

Figure 2. 2 The application of the local quality assessment guided MD-based protocol on an FM/TBM CASP12 target.
(A) The starting model for the CASP12 target TR896 coloured by the per-residue accuracy score produced by ModFOLD6. (B) The starting model coloured by the occupancy column, which based on the ModFOLD6 distance threshold, where blue regions represent the restrained regions and red regions represent the unrestrained regions during the MD simulation.

The analysis of the results starts from the comparison of the local quality assessment guided MD-based protocol versus the original MD-based protocol of ReFOLD. This is so we can quantify the usefulness of including restraints based on the predicted per-residue accuracy score. We also observed that the prediction methods that are used to generate the starting models are an important factor and should be taken into account for the comparison. The prediction methods that are used to generate starting models are based on the target difficulty. The target difficulty is determined by the CASP assessors based on fold similarity between the target structure and the known available structures. TBM based methods are generally favoured by predictors of 3D models in cases where good known template structures are available. Conversely, FM methods are generally

used when there is very low fold similarity or where the target has a novel fold. If the target sequence has some similar structural fragments that can be taken from experimentally determined structures, then a combination of TBM and FM methods may be used for the prediction of the target structure. The reliability of the predicted models by TBM is generally much higher than FM, because using knowledge from all available structures has historically been the better strategy and methods for this are more mature. There is a direct relationship between the accuracy of the per-residue error score and the target difficulty/methods used (Maghrabi & McGuffin, 2017) . Therefore, the accuracy of the per-residue accuracy score is also related to the availability of known structures. Thus, the predicted per-residue accuracy scores of the TBM models are likely much more accurate compared to those for FM models.

It is clear that the original MD-based protocol of ReFOLD performed better at refining the FM starting models compared to the local quality assessment guided MD-based protocol (denoted below as "local") in terms of the cumulative *maximum* GDT-HA score for FM targets (∑ΔGDT-HAmax(ReFOLD)=0.3972, ∑ΔGDT-HAmax(local)=0.1766) (Appendix 1). However, the cumulative *mean* GDT-HA and cumulative *minimum* GDT-HA scores of the FM models refined by the local quality assessment guided MD-based protocol is higher than those refined by the original MD-based protocol of ReFOLD (∑ΔGDT-HAmean (ReFOLD)=-0.0906224, and ∑ΔGDT-HAmin (ReFOLD)=-0.5127 versus ∑ΔGDT-HAmean(local)=-0.0510092, and ∑ΔGDT-HAmin(local)=-0.2495) (Appendix 1).

The effect of the usage of the per-residue accuracy score is clear in Figure 2.3-2.4 as the GDT-HA scores of the models refined by the local quality assessment guided MD-based protocol are observed to distribute in a narrower range with the majority greater than the starting model. In comparison, the GDT-HA scores for original MD-based protocol of ReFOLD produced models are distributed in a wider range compared to those produced by the local quality assessment guided MD-based protocol. The smaller ranges of GDT-HA scores resulted from the restraint strategy as a stricter restraint, which was applied in the local quality assessment guided MD-based protocol to avoid structural deviations. It is promising to see that both protocols showed a considerable increase in quality scores following the refinement of the FM targets, as there is plenty of room for improvement (Figure2.3-2.4, and Appendix 1-3).

A similar trend is observed for the FM/TBM targets; the local quality assessment guided MD-based protocol has a narrower range of GDT-HA scores in comparison with the original MD-based protocol of ReFOLD (Figure 2.5-2.6 and Appendix 4-6). Again, the cumulative *maximum* GDT-HA score of the original MD-based protocol of ReFOLD is also higher than for the local quality assessment guided MD-based protocol (∑ΔGDT-HAmax(ReFOLD)=0.284 versus ∑ΔGDT-HAmax(local)=0.1495) (Appendix 4), but the cumulative *mean* and cumulative *minimum* GDT-HA scores of the models refined by local quality assessment guided MD-based protocol is higher than the original MD-based protocol of ReFOLD (∑ΔGDT-HAmean(ReFOLD)=-0.209902, and ∑ΔGDT-HAmin(ReFOLD)=-0.6187 versus ∑ΔGDT-HAmean(local)=-0.097524, and ∑ΔGDT-HAmean(local)= -0.3045) (Appendix 4). Both protocols performed well on the FM/TBM CASP12 targets in terms of improving the starting models. Nonetheless, it can be said that the local quality assessment guided MD-based protocol showed a better performance on the FM/TBM targets due to the higher cumulative *mean* and *minimum* GDT-HA scores (Figure 2.5-2.6, and Appendix 4-6).

Historically it has been more challenging to improve the quality of TBM targets and highly accurate models in CASP experiments, which is understandable as there is less room for improvement and, therefore more potential to make those models worse. The original MD-based protocol of ReFOLD was also less successful on TBM targets during CASP12. Here we aim to increase the accuracy of the TBM targets with the guidance of the per-residue accuracy score using the philosophy “if it is not broken, then don’t fix it”. Thus, our new protocol will only attempt to improve the parts of a model that are likely to need improving, while the rest of the model is kept more or less fixed.

The deterioration rate of the TBM models refined by the original MD-based protocol of ReFOLD is much higher than the local quality assessment guided MD-based protocol, and this can be seen from the *mean* , *minimum* GDT-HA scores, and Figure 2.7-2.8. (∑ΔGDT-HAmean(ReFOLD)=-1.262351, and ∑ΔGDT-HAmin(ReFOLD)=-2.9414 versus ∑ΔGDT-HAmean(local)=-0.558148, and ∑ΔGDT-HAmin(local)=-1.3293) (Appendix 7). While, the local quality assessment guided MD-based protocol shows a similar behaviour on the TBM targets as for the FM and FM/TBM targets, most of the TBM models refined by the original MD-based protocol of ReFOLD have

deviated from the starting model (Figure 2.7-2.8, and Appendix 7-10). This is evidence that the use of the per-residue accuracy score has helped to prevent models from deviating further away from the native structure. However, again the original MD-based protocol of ReFOLD performed relatively better than the local quality assessment guided MD-based protocol according to the *maximum* GDT-HA score, as seen for the other targets (∑ΔGDT-HAmax(ReFOLD)=0.50170.284, ∑ΔGDT-HAmax(local)=0.2801) (Appendix 7). The important criterion is improving the overall quality of refined models, so it is obvious from the higher cumulative *mean* GDT-HA score that the local quality assessment guided MD-based protocol performed better than the original MD-based protocol of ReFOLD. The success of the local quality assessment guided MD-based protocol is also apparent in TBM targets, which may be due to the availability of more accurate per-residue accuracy scores compared with other targets. Despite the relatively high success of the local quality assessment guided MD-based protocol, a consistent improvement across all TBM targets is not observed (Figure 2.7-2.8, and Appendix 7-10).

According to the Molprobity score, both protocols showed a significant improvement in comparison with the starting model. Note that *lower* Molprobity scores indicate *higher* accuracy, which is the opposite of the GDT_HA score. The local quality assessment guided MD-based protocol is found to be an improvement on the original MD-based protocol of ReFOLD according to the *minimum* and *mean* Molprobity scores for all targets as in Figure 2.9 and Appendix 11-14 (∑ΔMolprobity (the starting models) =115.14, ∑ΔMolprobitymean (ReFOLD)=99.32419, and ∑ΔMolprobitymin (ReFOLD)=71.38, versus ∑ΔMolprobitymean (local)=74.546925, and ∑ΔMolprobitymin (local)=54.46) (Figure 2.9 and Appendix 11-14). The Molprobity score is based on the consideration of all atoms, and the experimentally determined structure is not required knowledge of native structure. It is evident that the restraint strategy applied in the local quality assessment guided MD-based protocol is successful at improving Molprobity scores, perhaps due to the fact that all atoms below the threshold based on the per-residue accuracy score are restrained, not just C-alpha atoms. In addition to this, the clearer difference indicates that the GDT-HA score, which just considers the C-alphas, may not be an adequate measurement alone in order to assess refined models, and scores that consider all atoms should also be taken into consideration. It should be noted that the native structure is not considered in the calculation of Molprobity score, so the GDT-HA score is used as the main measurement by CASP and in this study.

The refinement targets that are used to measure the performance of the refinement approaches are carefully selected by the CASP assessors to be particularly difficult (i.e., already well modelled) or interesting cases (e.g., models with incorrectly folded regions). However, 3D models built using standard servers for the regular (T0) targets are perhaps a more appropriate "real-world" test of refinement, as in reality, the general biologist would be obtaining similar quality starting models from automated servers. It should be noted that the performance of the local quality assessment guided MD-based protocol on refinement (TR) targets is nevertheless relatively better than for the regular (T0) targets (Figure 2.3- 2.9, and Appendix 1-14).

Three different thresholds, based on the per-residue accuracy scores were applied (3, 5 and 8 Ångströms) in order to determine which threshold should be applied to result in an optimal protocol. However, we observed that the difference among the thresholds is negligible according to the *minimum* and *maximum* GDT-HA and Molprobity scores (Appendix 15-18). The 3 Ångströms threshold appears to be applicable across all target difficulty categories (Appendix 15-18).

The results were reported using the observed structure to evaluate the local quality assessment guided MD-based protocol. ModFOLD6 was used to predict the global quality of the models refined by the local quality assessment guided MD-based protocol in the absence of the native structures, as previously used in the original ReFOLD (Shuid et al., 2017). The performance of ModFOLD6 was also evaluated using GDT-HA and Molprobity scores in terms of selecting the best model (Table 2.1). 38 out of 55 "best" models selected by ModFOLD6 were in fact deteriorated in quality compared to the starting models according to GDT-HA (Table 2.1). The cumulative GDT-HA score of the best model selected by ModFOLD6 is lower than the cumulative GDT-HA score of the starting models and *maximum* GDT-HA scores of the models refined by the local quality assessment guided MD-based protocol (Table 2.1). However, it is better than the cumulative *mean* GDT-HA score of the models refined by the local quality assessment guided MD-based protocol ($\sum$GDT-HAmodFOLD6(best model)=23.0221, $\sum$GDT-HAmax(local)=24.0583, $\sum$GDT-HAmean(local)=22.7597788 , and $\sum$GDT-HA(the starting models)=23.4559) (Table 2.1). Such results show that ModFOLD6 failed to select improved

models compared to the starting models. Nevertheless, using ModFOLD6 for selection resulted in better performance than the cumulative *mean* score of refined models according to GDT-HA, indicating that it is worthwhile and better than random.

It is worthy of note that the MolProbity score of the best models selected by ModFOLD6 is much lower than the starting models, and lower Molprobity scores are likely to be more physically realistic. (∑ΔMolprobity (the best model selected by ModFOLD6) =42.95, and ∑ΔMolprobity (the starting models) =121.93) (Table 2. 1). ModFOLD6 was found to be highly successful in terms of the selection of the improved models according to Molprobity scores, and this also demonstrates that the improvement of all atoms can be detected by the method (Table 2.1).

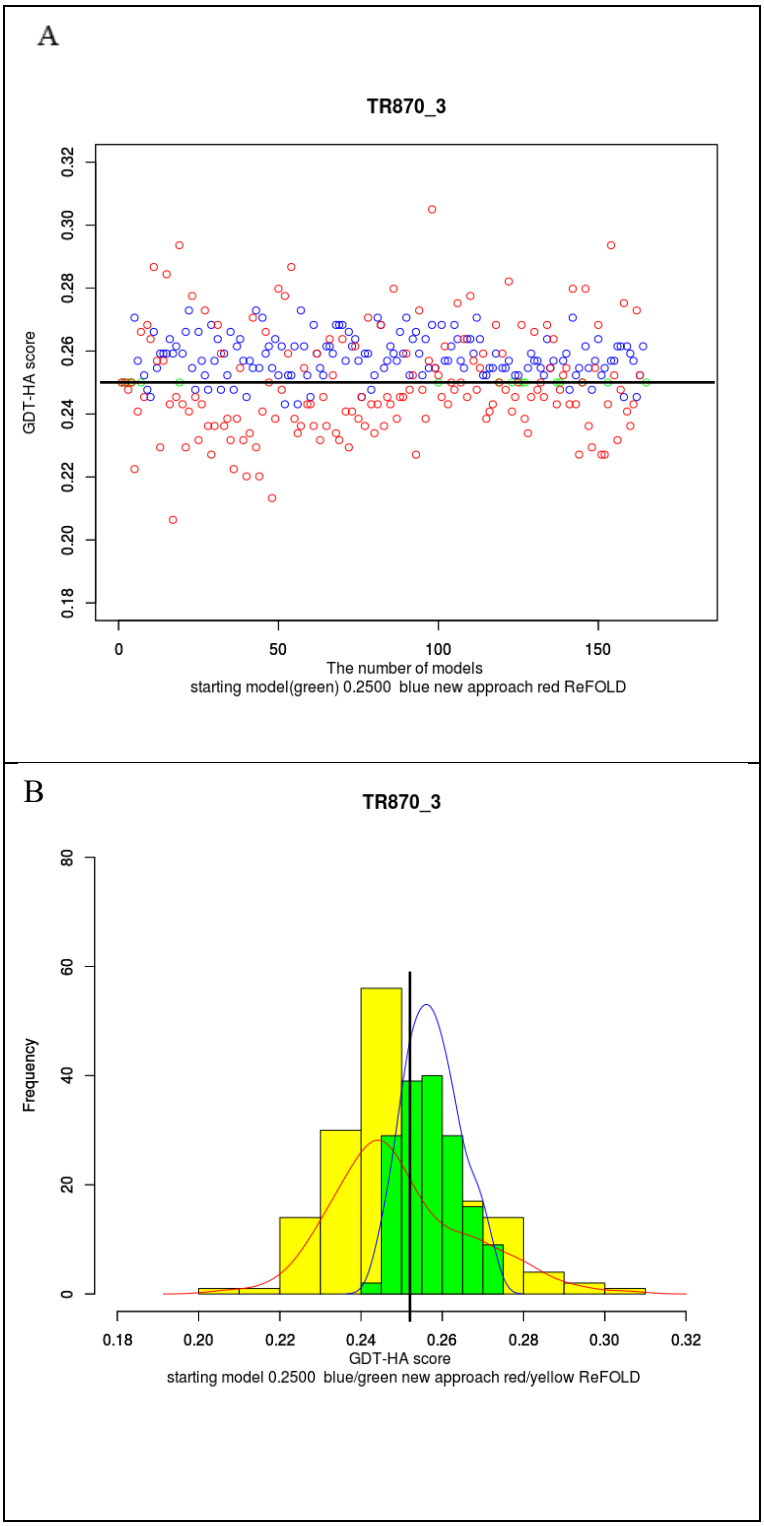

Figure 2. 3 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM target

Performance of methods on TR870 (an FM category CASP12 refinement target) according to GDT-HA score. The GDT-HA score of the starting model is 0.25 (with an applied threshold of 3 Ångströms). (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better)

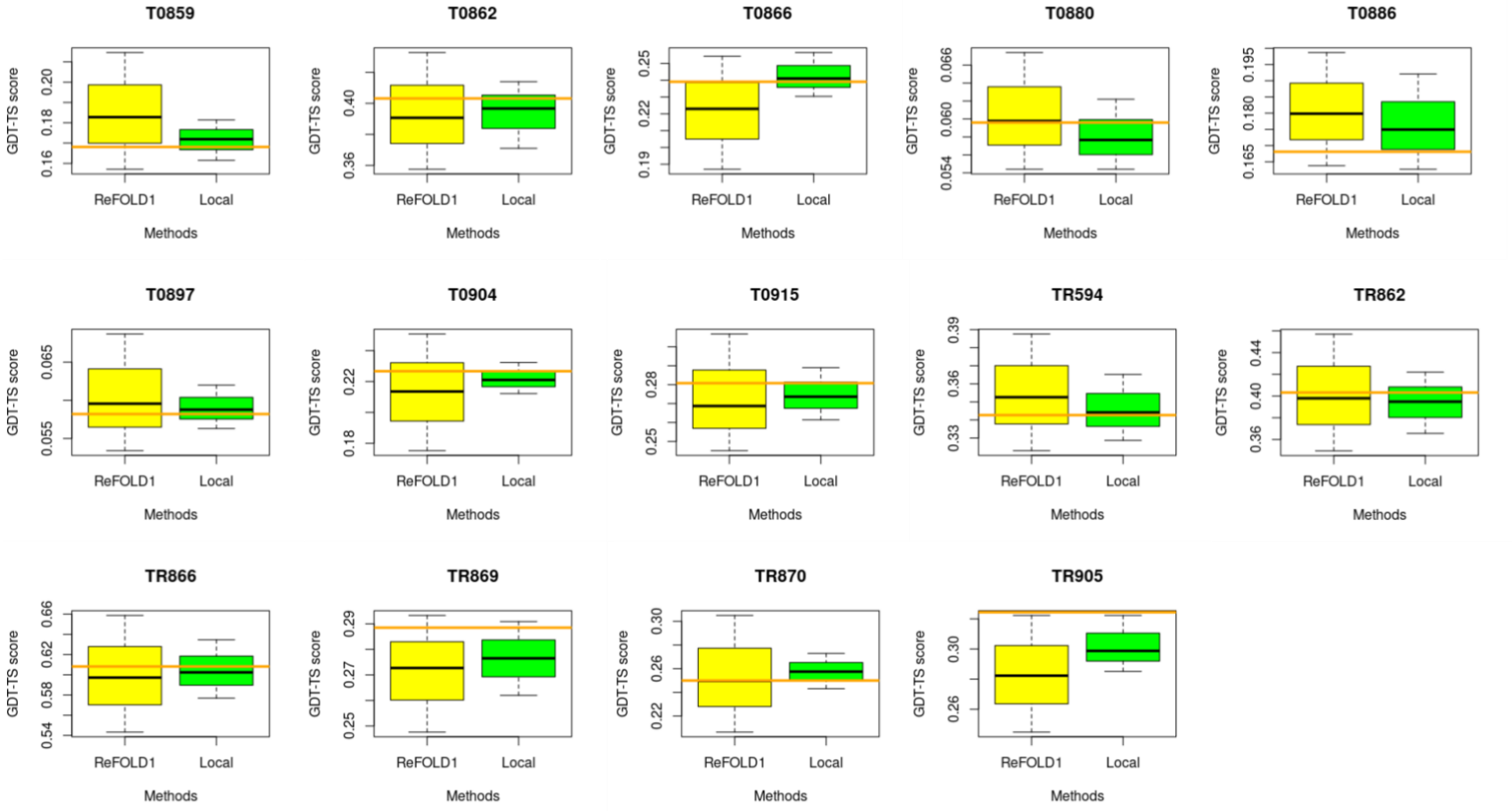

Figure 2. 4 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on the CASP12 FM targets according to the GDT-HA score.

The green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, yellow bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

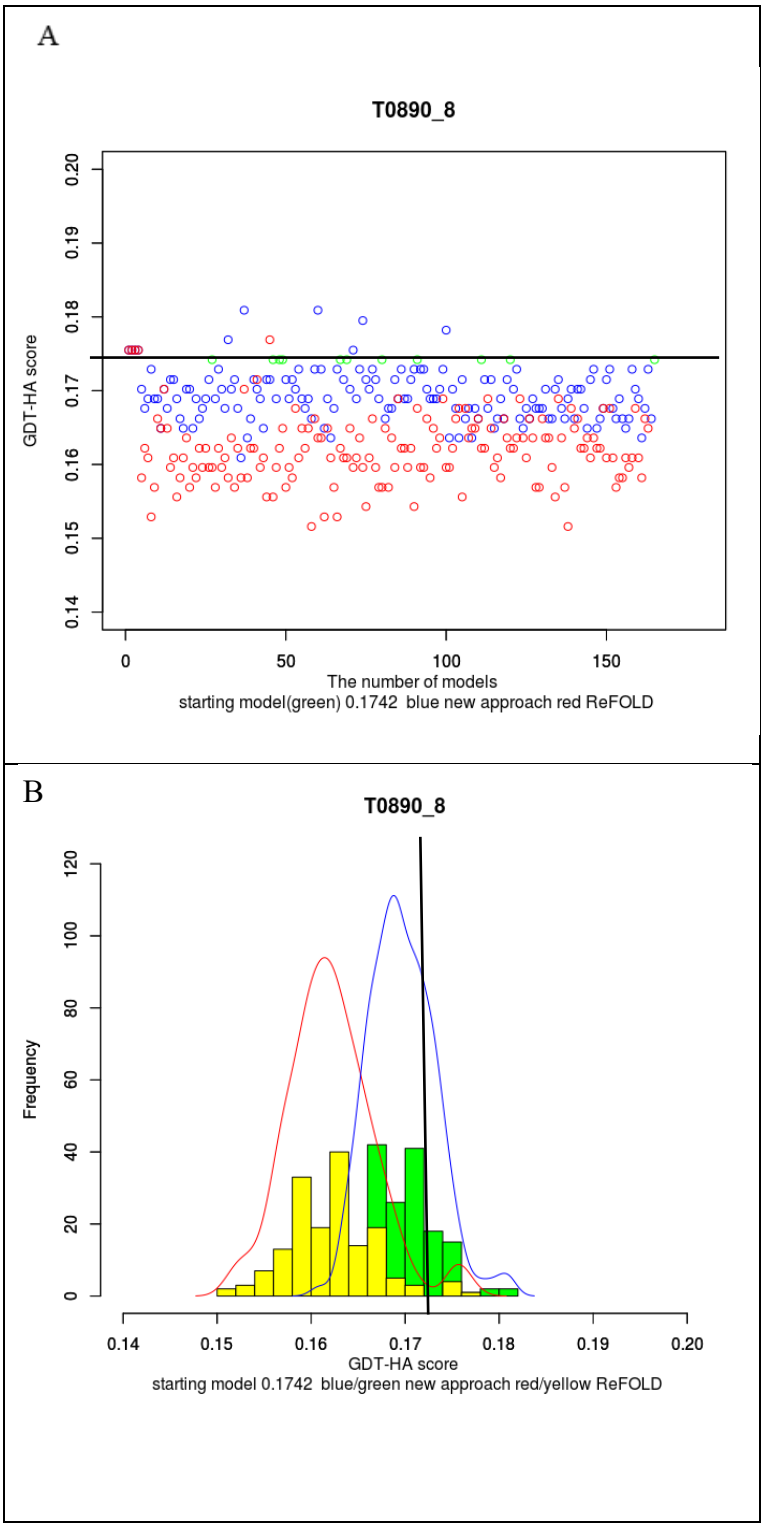

Figure 2. 5 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM/TBM target

Performance of methods on TR890 (an FM/TBM category CASP12 refinement target) according to GDT-HA score (with an applied threshold of 8 Ångströms). The GDT-HA score of the starting model is 0.1742. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better)

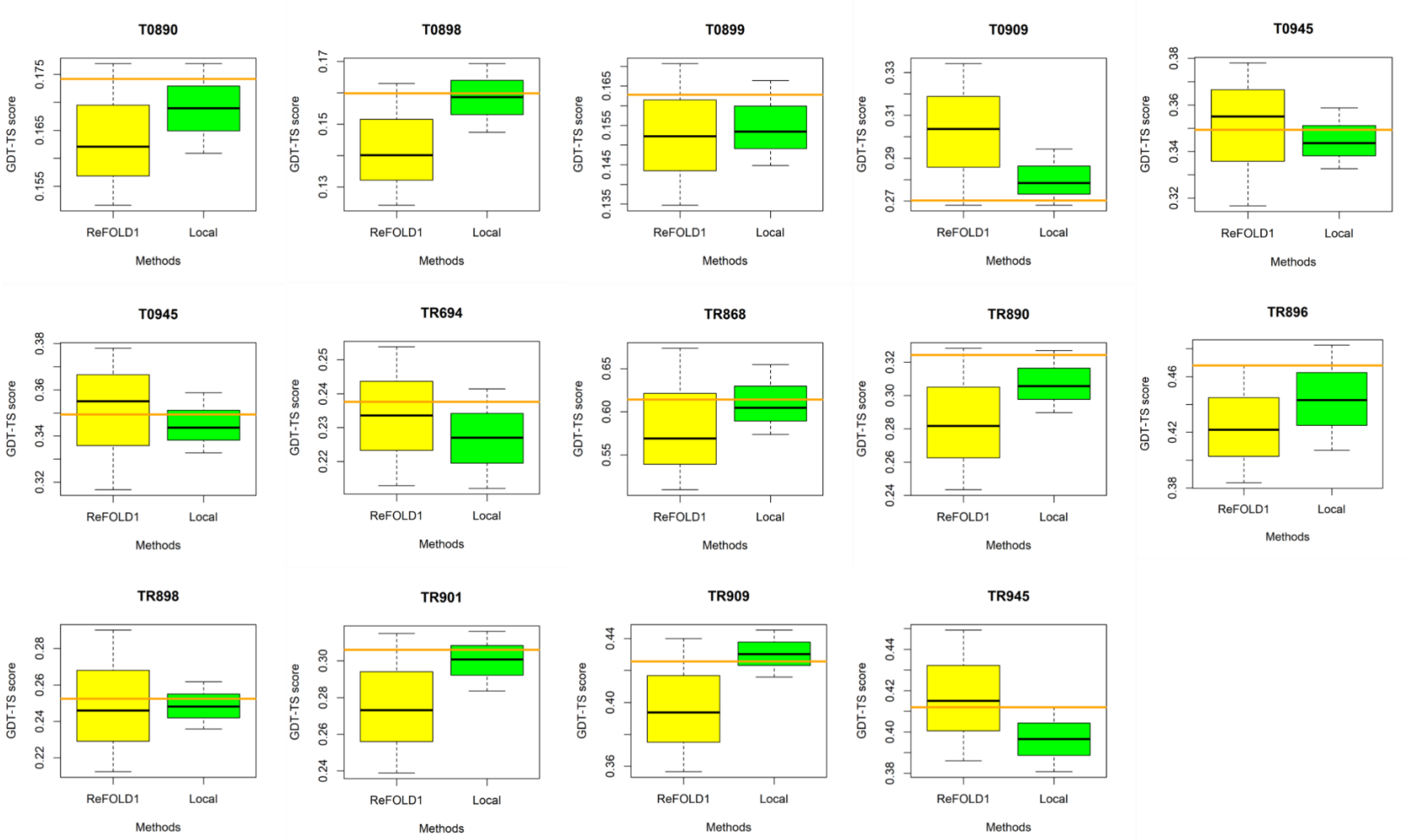

Figure 2. 6 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on the CASP12 FM/TBM targets according to the GDT-HA score.

The green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, yellow bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

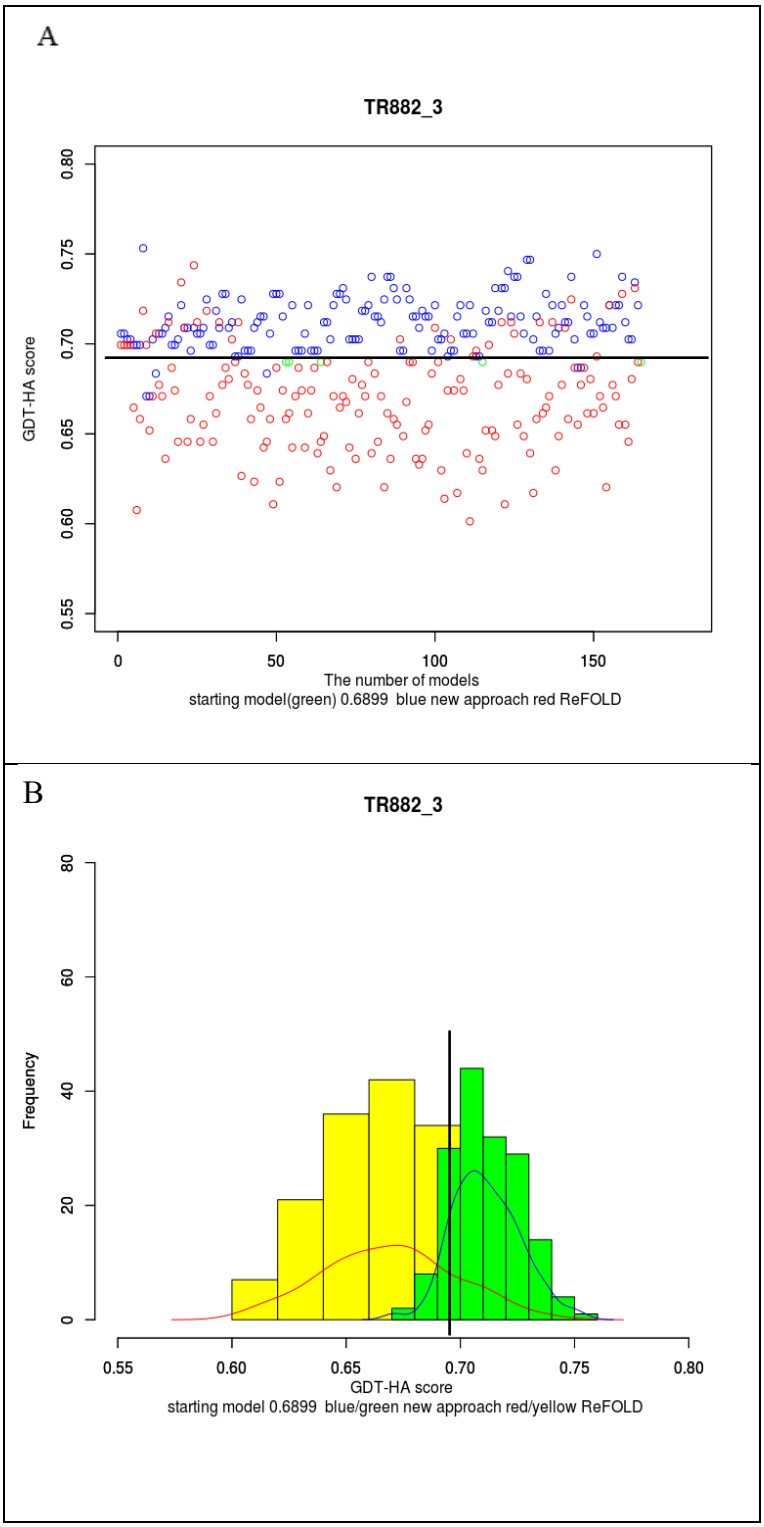

Figure 2. 7 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on a TBM target

Performance of methods on TR882 (a TBM category CASP12 refinement target) according to GDT-HA score (with an applied threshold of 3 Ångströms). The GDT-HA score of the starting model is 0.6899. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better)

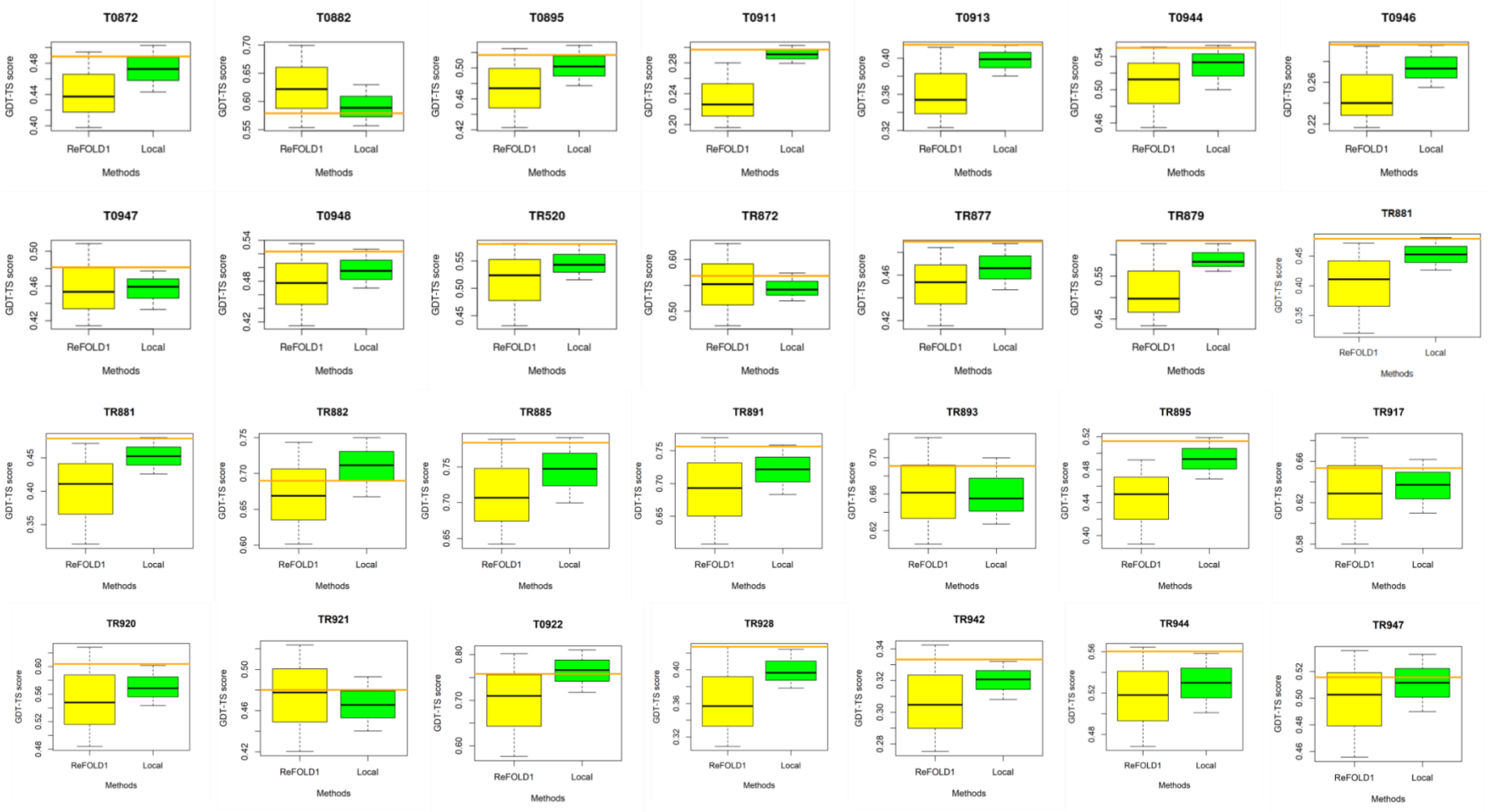

Figure 2. 8 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on the CASP12 TBM targets according to the GDT-HA score.

The green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, yellow bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

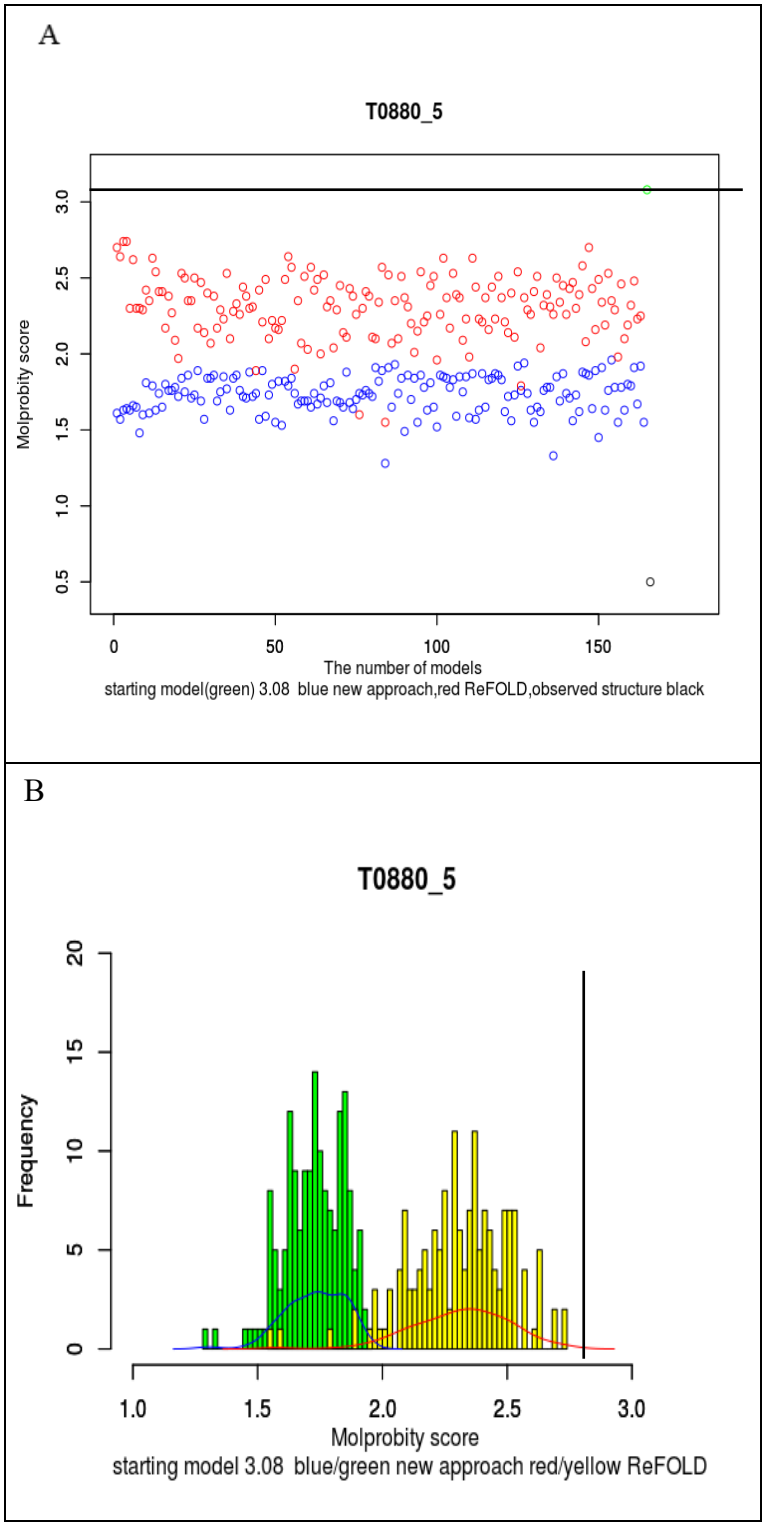

Figure 2. 9 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol according to Molprobity score.

Performance of methods on TR880 (an FM category CASP12 refinement target) according to Molprobity score (with an applied threshold of 5 Ångströms). The Molprobity score of the starting model is 3.08. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points below the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (lower Molprobity scores are better)

Target info.				GDT-HA score					Molprobity Score			
Target ID by domain	CASP Category	Prediction Method	selected top model ID	Starting model	Score of selected top model's score	Diff. score	Maximum score	Mean score	Starting model	selected the best model's score	Minimum Score	Mean Score
T0859	Regular	FM	2000-TR859_8-tra1.pdb	0.1681	0.1659	-0.0022	0.1814	0.17192	2.88	1.72	1.17	1.65713
T0862	Regular	FM	2800-TR862_1-tra1.pdb	0.4032	0.4032	0	0.414	0.396678	2.72	1.93	1.56	1.71963
T0866	Regular	FM	900-TR866_3-tra1.pdb	0.2391	0.2413	0.0022	0.2565	0.241027	3.29	1.88	1.21	1.76933
T0880	Regular	FM	2700-TR880_3-tra2.pdb	0.0596	0.057	-0.0026	0.0622	0.0576659	3.08	1.90	1.26	1.74982
T0886	Regular	FM	1000-TR886_5-tra3.pdb	0.1681	0.1747	0.0066	0.1921	0.174943	3.23	1.78	1.52	1.78854
T0897	Regular	FM	3500-TR897_3-tra3.pdb	0.0582	0.0601	0.0019	0.062	0.0587799	1.03	1.21	0.89	1.23537
T0904	Regular	FM	0-TR904_3-tra1.pdb	0.2267	0.2267	0	0.2323	0.221043	3.16	1.91	1.44	1.6375
T0915	Regular	FM	2200-TR915_5-tra1.pdb	0.2808	0.2695	-0.0113	0.289	0.273628	1.82	1.49	0.76	1.17085
T0890	Regular	FM/TBM	3900-TR890_3-tra3.pdb	0.1742	0.1715	-0.0027	0.1769	0.168949	2.3	1.84	0.89	1.27024
T0898	Regular	FM/TBM	3300-TR898_8-tra3.pdb	0.1599	0.1584	-0.0015	0.1693	0.158659	3.24	1.56	1.28	1.61866
T0909	Regular	FM/TBM	1000-TR909_5-tra1.pdb	0.2703	0.2815	0.0112	0.2943	0.278384	3.18	1.86	1.21	1.76707
T0872	Regular	TBM	1100-TR872_3-tra1.pdb	0.4886	0.4688	-0.0198	0.5028	0.472564	2.62	1.38	1.21	1.76707
T0882	Regular	TBM	2800-TR882_5-tra1.pdb	0.5791	0.5918	0.0127	0.6297	0.588816	1.73	1.40	0.92	1.3414
T0895	Regular	TBM	1000-TR895_3-tra2.pdb	0.5167	0.5125	-0.0042	0.5292	0.501923	2.24	1.57	0.66	1.0686
T0911	Regular	TBM	0-TR911_5-tra2.pdb	0.2972	0.2947	-0.0025	0.3027	0.291071	3.24	1.55	0.88	1.50561
T0913	Regular	TBM	0-TR913_5-tra4.pdb	0.4149	0.4135	-0.0014	0.4142	0.398787	3.09	2.68	1.23	1.48238
T0944	Regular	TBM	0-TR944_8-tra1.pdb	0.5504	0.5524	0.002	0.5534	0.533065	1.91	1.54	1.62	1.86073
T0946	Regular	TBM	1900-TR946_3-tra3.pdb	0.2962	0.2808	-0.0154	0.2954	0.27321	3.74	1.84	1.13	1.43963
T0948	Regular	TBM	0-TR948_8-tra3.pdb	0.5235	0.5151	-0.0084	0.5268	0.495054	2.72	1.84	1.63	1.88628
TR520	Refinement	TBM	0-TR520_1-tra1.pdb	0.581	0.581	0	0.581	0.543085	1.91	1.32	0.9	1.28591

TR872	Refinement	TBM	200-TR872_5-tra3.pdb	0.5682	0.5284	-0.0398	0.5739	0.541712	0.5	0.73	0.5	0.93372
TR877	Refinement	TBM	3500-TR877_1-tra2.pdb	0.4894	0.4648	-0.0246	0.4877	0.466093	1.41	1.33	0.77	1.23122
TR879	Refinement	TBM	2000-TR879_3-tra1.pdb	0.633	0.5977	-0.0353	0.6261	0.583705	3.07	1.82	1.31	1.65665
TR881	Refinement	TBM	300-TR881_3-tra1.pdb	0.479	0.4641	-0.0149	0.4802	0.452469	2.68	1.46	1.14	1.5203
TR882	Refinement	TBM	0-TR882_3-tra4.pdb	0.6899	0.7025	0.0126	0.75	0.71137	0.5	0.64	0.5	0.681037
TR885	Refinement	TBM	2600-TR885_3-tra4.pdb	0.7837	0.7909	0.0072	0.7909	0.747322	0.82	0.56	0.56	0.93878
TR891	Refinement	TBM	0-TR891_1-tra1.pdb	0.7567	0.7478	-0.0089	0.7589	0.721636	1.56	1.28	0.81	1.21762
TR893	Refinement	TBM	2500-TR893_3-tra1.pdb	0.6908	0.6642	-0.0266	0.6997	0.655179	1.51	0.96	0.68	1.05195
TR895	Refinement	TBM	0-TR895_5-tra2.pdb	0.5146	0.5083	-0.0063	0.5188	0.492889	2.22	1.17	0.87	1.21744
TR913	Refinement	TBM	2700-TR913_5-tra4.pdb	0.4534	0.4408	-0.0126	0.4586	0.436608	1.34	1.54	1.02	1.28902
TR917	Refinement	TBM	0-TR917_5-tra1.pdb	0.6535	0.6471	-0.0064	0.6618	0.637371	1.36	0.95	0.87	1.16061
TR920	Refinement	TBM	0-TR920_5-tra4.pdb	0.6039	0.5993	-0.0046	0.6016	0.568527	1.61	1.39	0.71	1.32872
TR921	Refinement	TBM	2600-TR921_1-tra4.pdb	0.4801	0.4728	-0.0073	0.4928	0.465543	1.61	1.42	0.9	1.3475
TR922	Refinement	TBM	2500-TR922_2-tra1.pdb	0.7581	0.754	-0.0041	0.8105	0.766796	1.07	1.5	0.65	1.23921
TR928	Refinement	TBM	2500-TR928_3-tra4.pdb	0.4274	0.3981	-0.0293	0.4245	0.396442	3.56	1.9	1.48	1.81598
TR942	Refinement	TBM	1000-TR942_3-tra3.pdb	0.3333	0.323	-0.0103	0.332	0.320707	2.32	1.54	1.34	1.53567
TR944	Refinement	TBM	2200-TR944_3-tra4.pdb	0.5603	0.5296	-0.0307	0.5583	0.529974	1.84	1.25	1.16	1.49024
TR947	Refinement	TBM	200-TR947_5-tra4.pdb	0.5157	0.5086	-0.0071	0.5329	0.511695	0.86	1.19	0.84	1.21372
TR948	Refinement	TBM	3100-TR948_3-tra2.pdb	0.5956	0.6007	0.0051	0.6242	0.594791	1.59	0.89	0.85	1.09707
TR868	Refinement	FM/TBM	4000-TR868_4-tra1.pdb	0.6143	0.5929	-0.0214	0.6548	0.604776	0.81	0.94	0.5	0.95
TR890	Refinement	FM/TBM	300-TR890_3-tra2.pdb	0.3245	0.3072	-0.0173	0.3271	0.305804	2.01	1.78	1.4	1.6928
TR896	Refinement	FM/TBM	3500-TR896_3-tra1.pdb	0.468	0.4506	-0.0174	0.4826	0.443089	2.14	1.5	1	1.31006
TR898	Refinement	FM/TBM	3800-TR898_4-tra3.pdb	0.2524	0.2453	-0.0071	0.2618	0.24816	0.66	0.68	0.5	0.827988

TR901	Refinement	FM/TBM	0-TR901_4-tra1.pdb	0.3061	0.3072	0.0011	0.3161	0.30088	2.03	1.48	0.98	1.37665
TR909	Refinement	FM/TBM	2600-TR909_5-tra1.pdb	0.4257	0.4309	0.0052	0.4452	0.430218	3.26	1.67	1.2	1.62829
TR945	Refinement	FM/TBM	500-TR945_5-tra1.pdb	0.412	0.4	-0.012	0.412	0.396554	2.33	1.35	1.11	1.35579
TR594	Refinement	FM	500-TRTR594_3-tra3.pdb	0.3427	0.3427	0	0.3652	0.344157	2.91	1.83	1.2	1.72317
TR862	Refinement	FM	2500-TR862_3-tra2.pdb	0.4032	0.3817	-0.0215	0.422	0.394813	2.26	1.06	0.78	1.24848
TR866	Refinement	FM	1000-TR866_3-tra4.pdb	0.6082	0.601	-0.0072	0.6346	0.602386	1.56	1.13	1.07	1.45
TR869	Refinement	FM	2300-TR869_3-tra2.pdb	0.2885	0.2861	-0.0024	0.2909	0.276501	1.98	1.4	0.94	1.36421
TR870	Refinement	FM	1600-TR870_3-tra3.pdb	0.25	0.2661	0.0161	0.2729	0.257492	3.61	1.79	1.07	1.55659
TR905	Refinement	FM	900-TR905_3-tra1.pdb	0.3244	0.2924	-0.032	0.3223	0.298757	2.36	1.5	1.16	1.52951
Cumulative Scores				23.4559	23.0221	-0.4338	24.0583	22.7597788	121.93	42.95	79.44	108.260763

Table 2. 1 Performance summary for the ModFOLD6 in terms of the selection of models generated by the local quality assessment guided MD-based protocol (higher GDT-HA scores are better, lower Molprobity scores are better).

2.5. Conclusions

ReFOLD was developed by our group to improve the local and global quality of the predicted 3D models with lower computational costs compared to other MD-based protocols. Despite overall improvements in model quality for some targets in CASP12, significant structural deviations from the native basin were observed in refined modes for the TBM targets using the original ReFOLD protocol (Shuid et al., 2017).

The predicted per-residue accuracy scores produced by ModFOLD6 indicate the likely C-alpha distances from the native structure, and this essential information was used to apply a new restraint strategy to improve upon the original MD-based protocol of ReFOLD. To selectively refine protein structures, the thresholds based on the per-residue accuracy score were applied during MD simulations. Using the per-residue accuracy score to guide the MD-based protocol has prevented the refined models from undesired structural deviations and this has been a step towards a more consistent refinement. The results presented in this chapter demonstrate that a more consistent refinement strategy has been achieved by applying the thresholds based on the per-residue accuracy score, compared to that used in the original ReFOLD method.

The local quality assessment guided MD-based protocol was shown to perform better than the original ReFOLD according to both the GDT-HA and Molprobity scores. Although the models refined by the original MD-based protocol of ReFOLD have higher *maximum* GDT-HA scores compared to the local quality assessment guided MD-based protocol, the cumulative *mean* GDT-HA scores of the models refined by the local quality assessment guided MD-based protocol is much higher than the models refined by the original MD-based protocol of ReFOLD. The higher cumulative *mean* GDT-HA score shows that the majority of the refined models are improved more than the models refined by ReFOLD, and the improvement is much clearer in TBM targets, which are often more difficult to refine (as they are often already of high quality), compared to the comparatively easier FM targets (where there is often more room for improvement).

It is also interesting that the local quality assessment guided MD-based protocol was found to be particularly successful according to the Molprobity scores, compared with ReFOLD and the starting models. This notable improvement highlights the importance of considering all atoms when measuring the quality of the refined models, as large improvements may be missed if C-alpha distances were the sole criteria used for benchmarking methods.

There have been many factors limiting the success of the refinement process, but two factors seem to be the most important determinants of the local quality assessment guided MD-based protocol. The first factor is the accuracy of the starting models, as it is hard to improve highly accurate models further compared to poorly predicted models – there is less room for improvement. The accuracy of the predicted per-residue accuracy score is also another factor affecting the improvement of refined models because more accurate per-residue accuracy scores improve the chance of selecting better restraints for refinement.

The selection of the best refinement models is also a difficult task, but this is essential for the process in practical, real-world cases where native structures are unavailable. The ModFOLD6 method was used to select improved models among the 3D models generated by the MD-based protocols prior to the knowledge of experimentally determined 3D structures. Unfortunately, the performance of ModFOLD6 was not completely satisfying and it failed in 38 out of 55 targets in terms of selecting the best model. Nevertheless, the cumulative GDT-HA score of models selected by ModFOLD6 is higher than the cumulative *mean* GDT-HA score. The failure might result from the high similarity between models for the same protein. The detection of very small differences between highly similar models is inherently hard and not what ModFOLD6 was trained to do, rather its main strength lies in the selection of good models from among a wide variety of tertiary structure prediction servers. In future, bespoke versions of ModFOLD could be developed which are specifically trained to detect smaller differences between refinement models.
In the next chapter, the performance of the local quality assessment guided restraint strategy in CASP13 is given and evaluated by utilising the CASP13 official results. The application of a gradual restraint strategy based on the local quality estimation is proposed and its performance is also compared with the fixed restraint strategy which was used in this chapter.

Chapter 3 The Application of the Local Quality Assessment Guided MD-Based Protocol in CASP13 and the Gradual Restraint Strategy

3.1 Background

The local quality assessment guided MD-based protocol developed in Chapter 2 was upgraded using the guidance of the improved per-residue accuracy score produced by ModFOLD version 7 (Cheng et al., 2019; Maghrabi & McGuffin, 2019). The upgraded version of the local quality assessment guided restraint strategy played an important part in our CASP13 refinement pipeline. The refinement pipeline was used to refine the best-predicted server model selected by ModFOLD7, in the regular prediction category (T0), and the refinement target models, in the refinement category (TR).

3.1.1 ModFOLD7

The ModFOLD server was upgraded to the seventh version following CASP12, mainly by increasing the accuracy of the local scoring components. ModFOLD7 consists of multiple pure-single and quasi-single methods to assess the quality of the predicted 3D models (Cheng et al., 2019; Maghrabi & McGuffin, 2019). ModFOLD7 includes an important upgrade to the generation of the per-residue accuracy score, and it accommodates ten local scoring methods (Cheng et al., 2019; Maghrabi & McGuffin, 2019): the Contact Distance Agreement (CDA) score (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017), the Secondary Structure Agreement (SSA) score (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017), ProQ2 (Uziela & Wallner, 2016), ProQ2D (Uziela, Hurtado, et al., 2016), ProQ3D (Uziela, Hurtado, et al., 2016), and VoroMQA (Olechnovič & Venclovas, 2017) were used as pure-single model methods (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017). In addition to the pure-single model methods, the following quasi-single model methods were also integrated: the Disorder "B-factor" Agreement (DBA) score (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017), MF5s (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017), ModFOLDclustQ_single (MFcQs) (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017) and ResQ (Yang et al., 2016), which used the reference model sets generated by IntFOLD5 to produce the per-residue accuracy scores (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017).

ModFOLD7 (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017) was also trained using two separate functions: The superposition based on S-score (Levitt & Gerstein, 1998) which was also produced in the previous version and the residue contact-based lDDT score (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013). The usage of the lDDT score made a significant improvement in the accuracy of the per-residue error produced by ModFOLD7.

The lDDT score assesses global and local accuracies of the models in terms of stereochemical quality (Mariani et al., 2013). Its calculation relies on the atom-atom distance between the 3D models considering all atoms, therefore it is independent of structural superpositions (Mariani et al., 2013).

Multilayer perceptron (MLPs) were also utilised to combine the 10 local scoring methods to produce consensus local quality assessment score (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013).

The performance of the ModFOLD6 and ModFOLD7 is continuously independently evaluated as part of the Continuous Automated Model Evaluation (CAMEO) project (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013). From the results shown in Figure 3.1, it is clear that ModFOLD7 shows improved performance in comparison to ModFOLD6 (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013). ModFOLD6 was using six local scoring methods and ModFOLD7 was upgraded by integrating the new CDA score generated using DeepMetaPSICOV (Kandathil et al., 2019a), ProQ3D (Uziela, Hurtado, et al., 2016), and VoroMQA (Olechnovič & Venclovas, 2017), and ResQ (Yang et al., 2016) pure-single and quasi-single local scoring methods. While ModFOLD6 was trained the superposition based on S-score (Levitt & Gerstein, 1998), ModFOLD7 was trained the S-score and the residue contact-based lDDT score. The integration of the four additional local scoring methods and training with the residue contact-based lDDT score boosted the performance of the quality assessment pipeline in terms of selecting the most native-like decoys and producing more accurate per-residue accuracy scores (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013).

ModFOLD7 was also assessed in CASP13 in Model Quality Estimation category and ranked among the top groups (Cheng et al., 2019; Maghrabi & McGuffin, 2019, 2017; Mariani et al., 2013).

Method	ROC		$ROC^{normalized}$	
	$AUC_{0,1}$	$AUC^{*}_{0,0.2}$	$AUC_{0,1}$	$AUC^{*}_{0,0.2}$
ModFOLD7	0.91	0.69	0.89	0.68
ModFOLD6	0.87	0.59	0.84	0.57

Table 3. 1 The performance comparison of ModFOLD7 and ModFOLD6 local model quality in the Continuous Automated Model Evaluation (CAMEO).
AUC = Area Under the ROC Curve. ROC= Receiver Operating Characteristic. AUC 0-0.1 = Area Under the ROC curve with False Positive Rate ≤ 0.1. AUC 0-0.2 = Area Under the ROC curve with False Positive Rate ≤ 0.2. The table is sorted by the AUC score. Scores closer to 1 indicate higher performance. Data are from https://www.cameo3d.org/quality-estimation

3.1.2 Iterative 3DRefine (i3Drefine)

3Drefine was developed to optimise hydrogen bonds and contacts by applying energy minimisation using a physics and knowledge-based force field (Bhattacharya & Cheng, 2013b). Applying the fully automated iterative refinement protocol of 3Drefine to protein structures includes two steps. The first step is the optimisation of hydrogen bonds and their connections. The optimisation followed by energy minimisation based on physics and knowledge-based force fields with the help of the MESHI molecular modelling packages in the second step (Bhattacharya & Cheng, 2013a, 2013b). The novel algorithm and energy functions used by MESHI rely on five key packages as molecular elements, geometry, energy, optimisers and utilities (Bhattacharya & Cheng, 2013a, 2013b; Kalisman et al., 2005).

The fully automated refinement program had been upgraded to include i3Drefine to refine structures with a strong composite physics and knowledge-based force field as a fast, free, and user-friendly web server utilities (Bhattacharya & Cheng, 2013a, 2013b; Kalisman et al., 2005). 3Drefine was firstly tested in CASP8 refinement and has been found to be a reliable and constructive approach in the following CASP experiments. i3Drefine is a leading server-based

iterative refinement program and showed a better performance compared to many non-server-based approaches in CASP10. Refining a model by i3Drefine is relatively fast and can take less than five minutes for small proteins utilities (Bhattacharya & Cheng, 2013a, 2013b; Kalisman et al., 2005). i3Drefine was also used to refine the 3D models to increase the accuracy of the initial structures in our CASP13 pipeline.

3.2 Aims and Objectives

The development of ModFOLD7 enabled us to upgrade the local quality assessment guided MD-based strategy developed in Chapter 2, both for guiding the MD simulation and selecting the best-refined models. The local quality assessment guided MD-based protocol was also tested as the distinctive part of our refinement pipeline in CASP13. The performance of the upgraded refinement pipeline in CASP13 refinement category was analysed according to GDT-TS (Zhang & Skolnick, 2005), GDT-HA (Zhang & Skolnick, 2005), Molprobity (Davis et al., 2004) and lDDT (Mariani et al., 2013) scores in this chapter.

Unlike the previous CASP experiments, relatively larger multi-domain and oligomeric structures were assigned as the regular CASP13 targets (Adiyaman & McGuffin, 2019). Applying one threshold based on the per-residue accuracy score by considering the distribution of the per-residue accuracy score was not found to be applicable for multi-domain structures after CASP13. It was proposed that if the per-residue accuracy score was low, much stronger restraint should be applied to not deviate the residues from the native basin. On the other hand, if the per-residue accuracy score was high, the residues should be refined further compared to others. Therefore, here we described the first use of a *gradual restraint strategy*, based on the per-residue accuracy score produced by ModFOLD7, instead of applying one fixed threshold in this chapter. Meanwhile, ModFOLD7 was developed by our group and performed better than ModFOLD6 in terms of producing per-residue accuracy scores. Therefore, ModFOLD7 was preferred for the generation of the predicted per-residue accuracy and MoldFOLD7 was also the best available tool to provide an automated approach with our refinement pipeline. The performance of the local quality

assessment guided fixed restraint strategy (from Chapter 2) was compared in terms of performance versus the new gradual restraint-based strategy, using the CASP13 target data.

3.2.1 Participation in the CASP Commons COVID-19 Initiative

At the time of writing this chapter, the COVID-19 pandemic has affected all aspects of our lives. As a part of the CASP Commons COVID-19 initiative, our group played an important part in the prediction of SARS-CoV-2 targets using our IntFOLD and ModFOLD methods. We also applied our ReFOLD methods to increase the accuracy of the top predicted 3D models using the gradual restraint strategy which was developed for this chapter. Therefore, here we also report on our initial results for the predictions that we made for the CASP Commons COVID-19 initiative.

3.3 Materials and Methods

3.3.1 Data Collection

The performance of our upgraded the local quality assessment guided MD-based protocol was independently evaluated during CASP13 experiment as part of the refinement category. The local quality assessment guided fixed and gradual restraint strategies were also compared using the regular CASP13 target models which were downloaded from (http://predictioncenter.org/download_area/CASP13). The 3D server models were scored using ModFOLD7 and the resulting local quality assessment data was used to guide the fixed and gradual restraint strategies.

3.3.2 Computational Design

The performance of our full refinement pipeline in CASP13 refinement category was firstly analysed using the observed structures (Figure 3.2). The comparison of the local quality assessment guided, and gradual restraints based on the per-residue accuracy score was also

investigated by refining the top ranked CASP13 server models for the regular (T0) targets, which were scored using ModFOLD7.

For the refinement category, our pipeline in CASP13 consisted of three protocols, which were similar to those used in CASP12 (Figure 3.2) (Adiyaman & McGuffin, 2019; Shuid et al., 2017). The first protocol used a rapid iterative strategy (i3Drefine) (Bhattacharya & Cheng, 2013a) with ten cycles and the second employed the newly upgraded version of the local quality assessment guided MD-based protocol, described in Chapter 2, to refine each starting model. Thus the major difference for the CASP13 refinement pipeline, compared with CASP12, was the modification of the second protocol, which included the introduction of molecular dynamics simulations that were guided by the per-residue accuracy scores produced by ModFOLD7 (Maghrabi & McGuffin, 2019).

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. The local quality assessment guided restraint strategy was applied by putting a threshold based on the per-residue accuracy scores (either 2, 3 or 5 Å) during the molecular dynamic simulation (Mirjalili & Feig, 2013; Shuid et al., 2017). For each starting model, the threshold was determined by considering the distribution of the per-residue accuracy scores.

Refined models generated from the second protocol were then assessed and ranked using the ModFOLD7_rank global score (which was optimised for selecting the best top model) (Maghrabi & McGuffin, 2019). The third protocol was the further refinement of the top-ranked model from the second protocol using i3Drefine. Finally, all 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 (Figure 3.2).

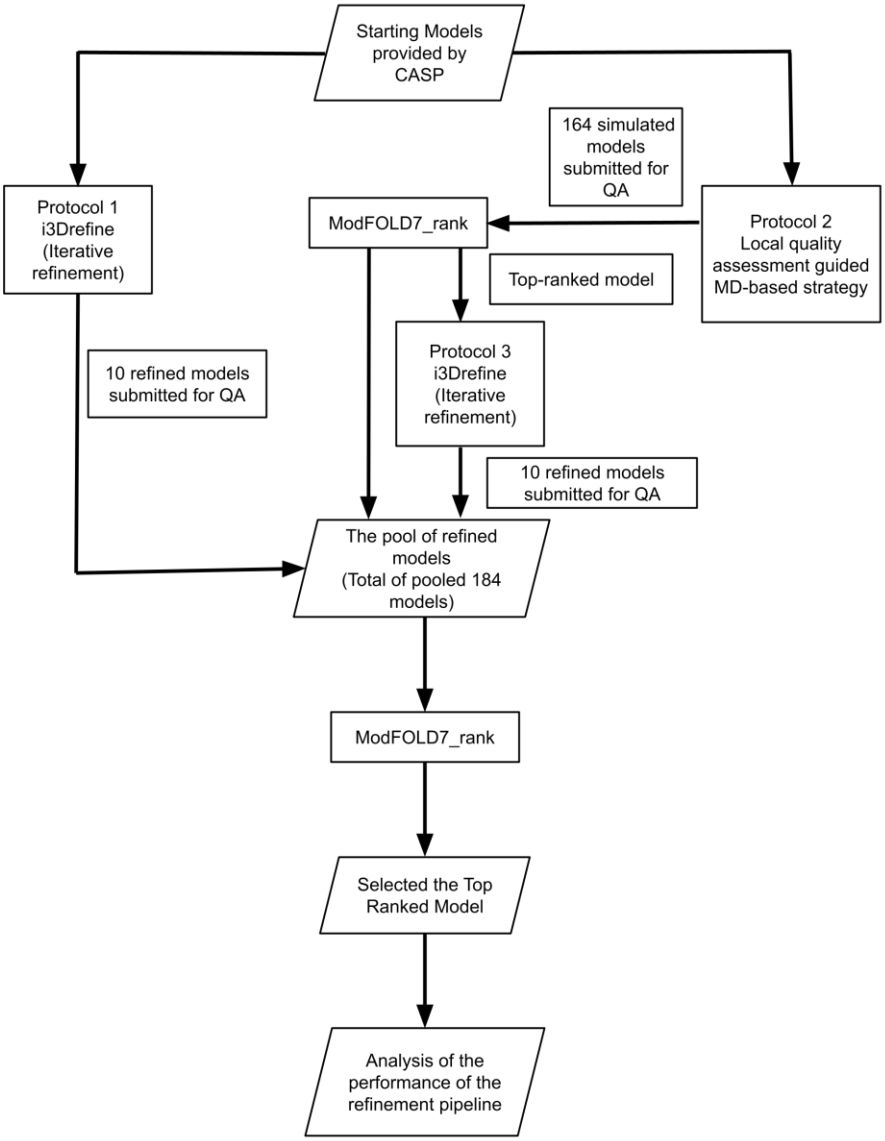

Figure 3. 1 Flowchart of our CASP13 refinement pipeline.
The refinement of the starting model using i3Drefine (Protocol 1), the local quality assessment guided MD-based protocol (Protocol 2), and the second round of i3Drefine iterative refinement strategy (Protocol 3), and. All refined models were ranked by the MoldFOLD7 server using the ModFOLD7_rank option (optimised for selecting the best top model).

The local quality assessment guided fixed and gradual restraints were also compared in terms of improving the quality of the best-predicted server models (Figure 3.3). The application of the restraint strategies starts with obtaining the global and local accuracy scores by submitting the 3D server models to ModFOLD7. After the identification of the poorly and well-predicted regions in the initial structure (the top selected model), a fixed restraint threshold was determined by considering the distribution of the per-residue accuracy errors (Figure 3.3 A2).

For the gradual restraint strategy, it was assumed that the regions identified as highly accurate should be restrained by applying a stronger harmonic positional restraint so as not to let them deviate from the native basin. A weaker restraint was also applied to the poorly predicted regions to allow for increases in the accuracy of these regions towards the experimental accuracy (Figure 3.3 B2) Thus, the gradual restraint ranges from weak (0.05 kCal/mol/Å^2) to strong (1 kCal/mol/Å^2) harmonic positional restraints on all atoms including C-alphas according to the distribution of the per- residue accuracy scores produced by ModFOLD7 (Table 3.1 and Figure 3.3) (Adiyaman & McGuffin, 2019; Maghrabi & McGuffin, 2019; Mirjalili & Feig, 2013; Read et al., 2019; Shuid et al., 2017). Different ranges of the force constant and per-residue accuracy scores were also applied, such as from 0.05 kCal/mol/Å^2 to 10 kCal/mol/Å^2. Nevertheless, the application of gradual restraint defined in Table 3.2 was found to be more effective in terms of the simulation execution, computational cost, and improving the initial structure (Adiyaman & McGuffin, 2019; Maghrabi & McGuffin, 2019; Mirjalili & Feig, 2013; Read et al., 2019; Shuid et al., 2017).

The per-residue accuracy score (Å)	The force constant (kcal/mol/Å^2)
0-2	1
2-4	0.5
4-6	0.1
6-8	0.05
8 and above	0

Table 3. 2 The application of the gradual restraint strategy based on the per-residue accuracy score produced by ModFOLD7.

The MD simulations for both strategies were conducted using NAMD 2.10 (Phillips et al., 2005) with the same parameters optimised for the original MD-based protocol of ReFOLD for the refinement of protein 3D models, as detailed in Chapter 2, to provide a fair comparison of the restraint strategies (Mirjalili et al., 2014; Mirjalili & Feig, 2013; Shuid et al., 2017).

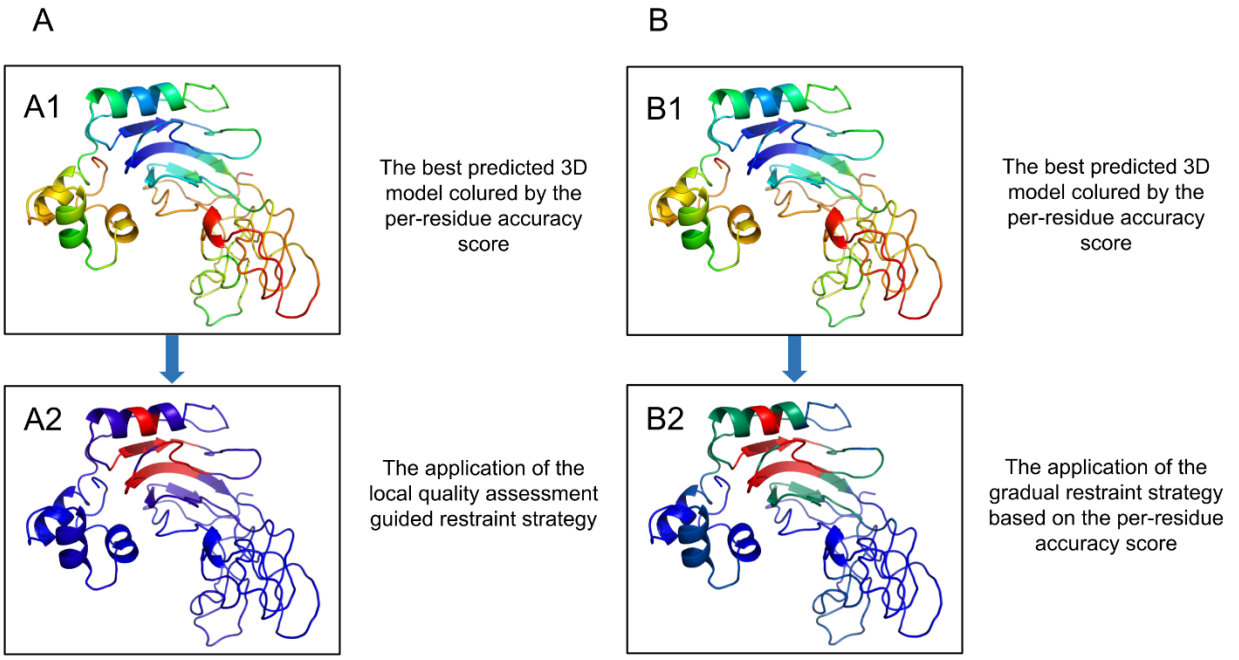

Figure 3. 2 The comparison of the local quality assessment guided fixed (A) and gradual (B) restraint strategies based on the initial per-residue accuracy scores
(A) The application of the fixed restraint strategy. 1) The best-predicted model (the CASP13 regular target T0953s2) is coloured using the per-residue accuracy score. 2) The initial structure is coloured using the occupancy column, where blue regions indicate unrestrained regions and red regions indicate restrained regions during the MD simulation. 3) The superposition of the initial structure (cyan), the best model generated by the local quality assessment guided MD-based protocol (magenta), and native structure (green). The initial structure versus the best model, a GDT_HA improvement from 0.1321 to 0.1452. (B) The application of the gradual restraint strategy. 1) The best-predicted model (the CASP13 regular target T0953s2); is coloured using the per-residue accuracy score. 2) The initial structure is coloured using the occupancy column, where red and green regions applied strong restraints and blue and light blue regions applied weaker restraints depending on the per-residue accuracy score during the MD simulation.

Following CASP13, once the native structure was released, the observed scores (mainly GDT-HA and Molprobity, but also GDT-TS and lDDT) were also used to compare the fixed and gradual restraint strategies. One-tailed unpaired Wilcoxon tests were also used to determine if differences in performance were statistically significant.

3.3.2.1 Use of Gradual Restraints and ModFOLD Version 8 for CASP Commons COVID-19

Prior to CASP14, as part of the CASP Commons COVID-19 initiative, the upgraded gradual restraint strategy was utilised in the same pipeline to increase the accuracy of models for the SARS-CoV-2 targets, but in this case the per-residue accuracy scores used for determining the restraints

were provided by ModFOLD version 8 (Figure 3.4). ModFOLD8 was also employed for the selection of the best-refined 3D models using the global scores from the ModFOLD8_rank option.

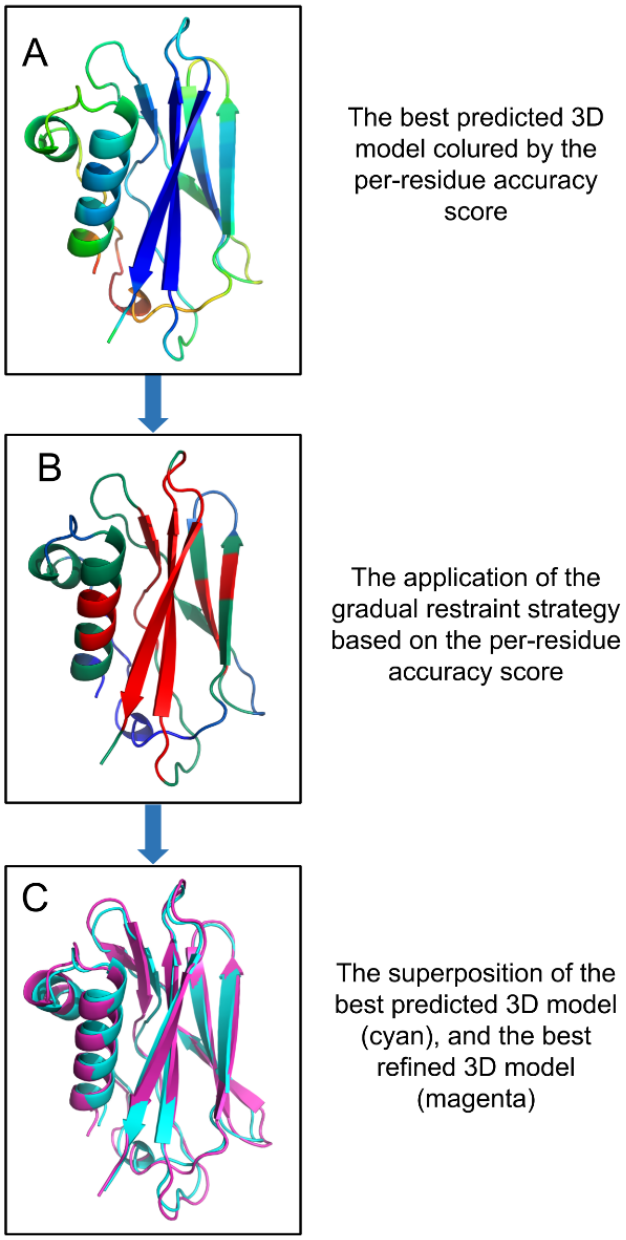

Figure 3. 3 The application of the gradual restraint strategy based on the per-residue accuracy score produced by ModFOLD8 for SARS-CoV-2 targets.
A) The best-predicted model (the SARS-CoV-2 target C1908 (ORF8)); is coloured using the per-residue accuracy score. B) The initial structure is coloured using the occupancy column, where red and green regions applied strong restraint and blue and light blue regions applied less strong restraint depending on the per-residue accuracy score during the MD simulation. C) The superposition of the initial structure (cyan) with the best model generated by the gradual restraint strategy (magenta).

3.4 Results and Discussion

In the refinement category of CASP, the assessors often select the best-predicted models as a specialised set of refinement targets, which are then used to evaluate the performance of refinement approaches. Further to this, our refinement pipeline in CASP13 was also used to refine our models submitted in the regular prediction categories. Here we report on the official CASP13 results along with a detailed assessment and discussion of our pipeline (Read et al., 2019). In addition, we carry out a comparison of the fixed and gradual restraint strategies using the CASP13 regular targets as a benchmark set. Finally, we report the on the initial results of the refinement of the SARS-CoV-2 proteins using the gradual restraint strategy for the CASP commons COVID-19 initiative.

3.4.1 The Performance of the Refinement Pipeline in CASP13

3.4.1.1 Performance in the Regular Category

Even though the refinement pipelines aim at improving the quality of the refinement targets (TR targets), we also used it to increase the quality of the best-predicted model selected by ModFOLD7 in the regular prediction category (T0 targets). This is an important competitive advantage of our manual group (called McGuffin) during the CASP13 experiment – we were able to both predict 3D models using the IntFOLD5 server (McGuffin et al., 2019) and identify the best-predict server models using ModFOLD7 (Maghrabi & McGuffin, 2019), which produced local and global scores for every models. The local quality assessment score produced by ModFOLD7 were then utilised to further improve the quality of the best-predicted server model by employing the local quality assessment guided MD-based protocol in our prediction pipeline. Overall, in the regular target prediction category, our group was ranked 6th out of 146 prediction groups on the both the TBM (Table 3.2) and TBM+TBM/FM domains (Template-based modelling/Free modelling domain) (Table 3.3), 9th out of 146 prediction groups for the FM+TBM/FM domains (Table 3.4), and 13th for FM (free modelling domain), according to the official assessor's formulae (Table 3.5).

3.4.1.2 Performance in the Refinement Category

In the refinement category, our refinement pipeline ranked 9th out of 37 methods according to the overall results based on the GDT-TS scores (Table 3.6). The performance of the refinement protocol was also analysed for each individual refinement target by way of a comparison of the refined model versus the starting models provided by the CASP assessors (Table 3.7). These data also provide a more detailed analysis of ModFOLD7 in terms of its ability to select the best-refined 3D model. The cumulative GDT-TS and GDT-HA scores of starting models provided by CASP were slightly higher than the best-refined 3D model selected by ModFOLD7 (∑GDT-TSstarting of 2083.05 versus ∑GDT-TSbest-refined of 2076.33 and ∑GDT-HAstarting of 1537.19 versus ∑GDT-HAbest-refined of 1523.31) (Table 3.7). It is evident that the refinement pipeline did not improve upon every starting model in the refinement category, and its performance was better in terms of GDT-TS scores compared to the GDT-HA score. This may be a consequence of the optimisation of ModFOLD7_rank, which was developed to select the 3D models with optimal GDT-TS scores. Nevertheless, ModFOLD7 successfully selected 14 improved models compared to the starting models according to the GDT-TS score, and 12 improved models according to the GDT-HA scores from among the 29 CASP13 refinement targets (Table 3.7). Although ModFOLD7 did not manage to select improved models for all targets, it can be said that it managed to select improved models for around half of the targets according to the GDT-TS scores. It must be restated that the ModFOLD server has never been specifically designed for the selection of refinement models which are much more similar to each other in terms of structures compared to the variety of alternative models produced in the conventional prediction pipeline.

Despite the partial success in the refinement category measured according to the GDT-TS and GDT-HA scores, it should be noted that the 3D models submitted by our group have higher cumulative lDDT scores and lower cumulative Molprobity scores compared to the starting models (∑lDDTstarting of 18.04 versus ∑lDDTbest-refined of 18.06 and ∑Molprobitystarting of 69.78 versus ∑Molprobitybest-refined of 68.33) (Table 3.7). Thus, it is clear that the refinement pipeline was indeed *successful* in increasing the overall quality of the starting models, according to both the lDDT and Molprobity scores. It is also noteworthy that both lDDT and Molprobity scoring

measurements take all atoms into consideration to produce the scores, unlike the GDT-HA and GDT-TS scores, which only consider the backbone.

Ranking	Group Code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore(>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore(>0.0)
1	322	Zhang	61	52.5838	1	0.862	1	52.7222	1	0.8643	1
2	222	Seok-refine	61	43.605	4	0.7148	4	47.4017	2	0.7771	2
3	261	Zhang-Server	61	46.1863	2	0.7572	2	46.6378	3	0.7646	3
4	43	A7D	61	26.9878	17	0.4424	20	46.1911	4	0.7572	4
5	145	QUARK	61	44.8207	3	0.7348	3	44.9771	5	0.7373	5
6	**460**	**McGuffin**	**61**	**36.0543**	**10**	**0.5911**	**12**	**43.8806**	**6**	**0.7194**	**6**
7	324	RaptorX-DeepModeller	61	40.9172	5	0.6708	5	42.7321	7	0.7005	8
8	55	VoroMQA-select	61	40.2589	6	0.66	6	42.7009	8	0.7	9
9	221	RaptorX-TBM	61	38.9471	7	0.6385	7	41.5778	9	0.6816	10
10	135	SBROD	59	33.3417	12	0.6329	8	41.3783	10	0.7013	7
11	156	Seok-server	61	37.2846	8	0.6112	10	40.1255	11	0.6578	13
12	89	MULTICOM	61	36.297	9	0.595	11	40.0499	12	0.6566	14
13	86	BAKER	60	31.571	15	0.5595	14	39.6372	13	0.6606	12
14	68	Seok	61	32.19	13	0.5277	15	38.5935	14	0.6327	15
15	344	Kiharalab	61	25.3614	20	0.4158	22	37.3525	15	0.6123	16
16	354	wfAll-Cheng	61	34.8482	11	0.5713	13	36.592	16	0.5999	17
17	390	Bhattacharya	61	31.9754	14	0.5242	16	36.1383	17	0.5924	19
18	368	BAKER-ROSETTASERVER	61	26.5875	18	0.4359	21	35.9077	18	0.5887	20
19	197	MESHI	61	22.4941	22	0.3688	27	35.8494	19	0.5877	21
20	196	Grudinin	61	30.531	16	0.5005	17	35.4646	20	0.5814	23

Table 3. 3 Official CASP13 results for TBM domains according to the CASP assessor 's formula (GDT_HA + (SG+IDDT+CAD)/3 + ASE) for the top 20 groups.

The table is sorted by SUM Zscore (>-2.0). GDT High Accuracy (GDT_HA) (Zhang & Skolnick, 2005), Sphere Grinder Score (SG) (Antczak et al., 2015) , local Distance Difference Test (lDDT) (Mariani et al., 2013), Contact Area Difference Score (CAD) (Olechnovič et al., 2013), and ASE (Accuracy Self Estimate) (Moult et al., 2009; Read et al., 2019)Data are from http://www.predictioncenter.org/casp13/

Ranking	Group code	Group name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	322	Zhang	73	62.7432	1	0.8595	1	63.2734	1	0.8668	1
2	43	A7D	73	42.4062	10	0.5809	12	62.4677	2	0.8557	2
3	222	Seok-refine	73	51.1219	4	0.7003	4	56.0224	3	0.7674	3
4	261	Zhang-Server	73	55.1437	2	0.7554	2	55.9167	4	0.766	4
5	145	QUARK	73	54.0746	3	0.7407	3	54.6177	5	0.7482	5
6	**460**	**McGuffin**	**73**	**44.8183**	**8**	**0.6139**	**9**	**53.5764**	**6**	**0.7339**	**6**
7	55	VoroMQA-select	73	49.633	5	0.6799	5	52.6498	7	0.7212	7
8	324	RaptorX-DeepModeller	73	48.1794	6	0.66	6	51.9211	8	0.7112	8
9	89	MULTICOM	73	47.7389	7	0.654	7	51.8728	9	0.7106	9
10	135	SBROD	71	40.5003	12	0.6268	8	50.2808	10	0.7082	10
11	221	RaptorX-TBM	73	44.1652	9	0.605	10	48.4121	11	0.6632	13
12	86	BAKER	71	36.1806	17	0.5659	14	48.3697	12	0.6813	11
13	68	Seok	73	39.2428	14	0.5376	16	46.7455	13	0.6403	14
14	344	Kiharalab	73	33.5382	18	0.4594	22	46.1782	14	0.6326	15
15	390	Bhattacharya	73	39.5471	13	0.5417	15	45.4523	15	0.6226	17
16	354	wfAll-Cheng	73	42.3676	11	0.5804	13	44.6591	16	0.6118	19
17	368	BAKER-ROSETTASERVER	72	32.8749	19	0.4844	20	44.4382	17	0.6172	18
18	156	Seok-server	73	36.5969	15	0.5013	17	44.3329	18	0.6073	20
19	214	wfRosetta-ModF7	71	30.9113	21	0.4917	19	44.2282	19	0.6229	16
20	197	MESHI	73	28.7632	22	0.394	25	43.8485	20	0.6007	21

Table 3. 4 Official CASP13 results for TBM + TBM/FM domains according to the CASP assessor 's formula (GDT_HA + (SG+IDDT+CAD)/3 + ASE) for the top 20 groups.

The table is sorted by SUM Zscore (>-2.0). GDT High Accuracy (GDT_HA) (Zhang & Skolnick, 2005), Sphere Grinder Score (SG) (Antczak et al., 2015) , local Distance Difference Test (lDDT) (Mariani et al., 2013), Contact Area Difference Score (CAD) (Olechnovič et al., 2013), and Accuracy Self Estimate (ASE) (Moult et al., 2009; Read et al., 2019)Data are from http://www.predictioncenter.org/casp13/

Ranking	Group code	Group name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	43	A7D	43	70.3397	1	1.6358	1	70.3397	1	1.6358	1
2	322	Zhang	43	53.6861	2	1.2485	2	54.0428	2	1.2568	2
3	89	MULTICOM	43	49.905	3	1.1606	3	50.265	3	1.169	3
4	145	QUARK	43	46.1771	4	1.0739	4	46.5848	4	1.0834	4
5	261	Zhang-Server	43	43.1437	5	1.0033	5	43.6174	5	1.0144	5
6	224	Destini	43	39.5602	6	0.92	6	40.9142	6	0.9515	6
7	354	wfAll-Cheng	43	37.2083	7	0.8653	7	39.5994	7	0.9209	7
8	196	Grudinin	43	36.0588	8	0.8386	8	38.0391	8	0.8846	9
9	**460**	**McGuffin**	**43**	**35.6027**	**9**	**0.828**	**9**	**37.6625**	**10**	**0.8759**	**11**
10	117	Jones-UCL	43	34.9813	10	0.8135	10	36.0997	13	0.8395	14
11	135	SBROD	43	34.9431	11	0.8126	11	37.8881	9	0.8811	10
12	197	MESHI	43	34.4519	12	0.8012	12	37.0062	11	0.8606	12
13	324	RaptorX-DeepModeller	43	33.6144	13	0.7817	14	36.0591	14	0.8386	15
14	498	RaptorX-Contact	43	32.4165	14	0.7539	15	36.2038	12	0.8419	13
15	55	VoroMQA-select	43	32.333	15	0.7519	16	35.2287	16	0.8193	17
16	208	KIAS-Gdansk	43	30.932	16	0.7193	17	34.5881	17	0.8044	18
17	418	Seder3nc	43	30.126	17	0.7006	18	35.4041	15	0.8234	16
18	274	MUFold	43	27.0439	18	0.6289	21	30.8619	20	0.7177	22
19	457	Wallner	43	27.0115	19	0.6282	22	29.9949	24	0.6976	27
20	44	ProQ2	43	25.7608	20	0.5991	23	30.7989	21	0.7163	24

Table 3. 5 Official CASP13 results for FM + TBM/FM domains according to the CASP assessor 's formula (GDT_TS + QCS) for the top 20 groups.

The table is sorted by SUM Zscore (>-2.0). Global Distance Test Total Score (GDT_TS) (Zhang & Skolnick, 2005), Quality Control Score (QCS) (Cong et al., 2011) http://www.predictioncenter.org/casp13/

Ranking	Group code	Group name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	43	A7D	31	54.9935	1	1.774	1	54.9935	1	1.774	1
2	322	Zhang	31	42.7836	2	1.3801	2	42.8133	2	1.3811	2
3	145	QUARK	31	36.2191	3	1.1684	3	36.2827	4	1.1704	4
4	89	MULTICOM	31	36.1782	4	1.167	4	36.5381	3	1.1786	3
5	261	Zhang-Server	31	33.2515	5	1.0726	5	33.4101	5	1.0777	5
6	224	Destini	31	31.9069	6	1.0293	6	32.7036	6	1.055	6
7	498	RaptorX-Contact	31	29.6487	7	0.9564	7	30.707	7	0.9905	8
8	197	MESHI	31	27.7536	8	0.8953	9	29.0389	9	0.9367	10
9	354	wfAll-Cheng	31	27.4311	9	0.8849	10	29.8221	8	0.962	9
10	196	Grudinin	31	27.4111	10	0.8842	11	28.5938	10	0.9224	11
11	324	RaptorX-DeepModeller	31	26.9495	11	0.8693	12	27.6441	11	0.8917	12
12	117	Jones-UCL	31	25.9528	12	0.8372	13	26.7971	13	0.8644	14
13	**460**	**McGuffin**	**31**	**25.5098**	**13**	**0.8229**	**14**	**26.5964**	**14**	**0.8579**	**15**
14	135	SBROD	31	25.204	14	0.813	15	27.5029	12	0.8872	13
15	208	KIAS-Gdansk	31	23.7414	15	0.7659	16	26.2444	15	0.8466	16
16	55	VoroMQA-select	31	22.7581	16	0.7341	17	25.5219	17	0.8233	18
17	418	Seder3nc	31	22.1343	17	0.714	18	25.5574	16	0.8244	17
18	457	Wallner	31	20.1724	18	0.6507	19	22.3701	22	0.7216	23
19	192	Elofsson	31	20.15	19	0.65	20	23.6015	18	0.7613	19
20	44	ProQ2	31	19.7657	20	0.6376	21	23.0521	19	0.7436	20

Table 3. 6 Official CASP13 results for FM domains according to the CASP assessor 's formula (GDT_TS + QCS) for the top 20 groups.
The table is sorted by SUM Zscore (>-2.0). Global Distance Test Total Score (GDT_TS) (Zhang & Skolnick, 2005), Quality Control Score (QCS) (Cong et al., 2011) http://www.predictioncenter.org/casp13/

Ranking	Group code	Group name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	356	FEIGLAB	29	30.467	1	1.0506	1	31.434	1	1.0839	1
2	86	BAKER	29	21.8224	2	0.7525	2	24.4866	2	0.8444	2
3	425	BAKER-AUTOREFINE	29	20.1455	3	0.6947	3	22.9743	3	0.7922	3
4	156	Seok-server	29	17.907	4	0.6175	4	18.3973	4	0.6344	4
5	390	Bhattacharya	29	14.1785	5	0.4889	5	14.2819	5	0.4925	5
6	117	Jones-UCL	29	9.6293	9	0.332	11	13.7515	6	0.4742	7
7	102	Bhattacharya-Server	29	13.1079	6	0.452	7	13.4647	7	0.4643	8
8	344	Kiharalab	29	12.8538	8	0.4432	9	13.1466	8	0.4533	9
9	**460**	**McGuffin**	**29**	**13.0312**	**7**	**0.4494**	**8**	**13.1346**	**9**	**0.4529**	**10**
10	174	Zhang-Refinement	27	8.5455	12	0.4646	6	12.9981	10	0.4814	6
11	68	Seok	29	8.5765	11	0.2957	13	12.3752	11	0.4267	12
12	312	MUFold_server	27	1.5804	16	0.2067	15	12.0316	12	0.4456	11
13	190	DC_refine	29	8.2978	13	0.2861	14	11.4749	13	0.3957	14
14	217	Boniecki_pred	28	7.6	14	0.3429	10	11.4674	14	0.4095	13
15	433	AIR	29	9.3973	10	0.324	12	10.9292	15	0.3769	15
16	4	YASARA	28	2.0243	15	0.1437	16	8.5517	16	0.3054	16
17	270	Huang	29	-2.5068	17	-0.0864	17	7.0247	17	0.2422	18
18	208	KIAS-Gdansk	29	-10.4628	19	-0.3608	21	6.4424	18	0.2222	19
19	112	AWSEM	27	-9.8841	18	-0.2179	20	5.8188	19	0.2155	20
20	457	Wallner	19	-22.2482	22	-0.1183	19	5.2959	20	0.2787	17

Table 3. 7 Official CASP13 results for all refinement targets according to the GDT-TS based scores for the top 20 groups.
The table is sorted by SUM Zscore (>-2.0).Global Distance Test Total Score (GDT_TS) (Zhang & Skolnick, 2005), Data are from http://www.predictioncenter.org/casp13/

CASP ID		GDT-TS score			GDT_HA score			Molprobity score			lDDT score		
Target ID	Model	starting model	submitted model	Diff	starting model	submitted model	Diff.	starting model	submitted model	Diff.	starting model	submitted model	Diff.
R0949	R0949TS460_1	64.53	62.4	-2.13	49.03	46.12	-2.91	2.99	2.7	0.29	0.53	0.52	-0.01
R0957s2	R0957s2TS460_1-D1	60.97	61.13	0.16	39.35	39.68	0.33	3.48	2.96	0.52	0.57	0.57	0
R0959	R0959TS460_1	64.55	64.68	0.13	44.71	44.84	0.13	0.72	1.41	0.69	0.62	0.62	0
R0962	R0962TS460_1	80.51	80.08	-0.43	62.99	62.01	-0.98	2.73	2.23	-0.5	0.7	0.69	-0.01
R0968s1	R0968s1TS460_1	66.74	67.16	0.42	45.13	44.7	-0.43	0.82	1.79	0.97	0.61	0.6	-0.01
R0968s2	R0968s2TS460_1	71.3	71.74	0.44	50.43	50.87	0.44	1.04	2.81	1.77	0.6	0.61	0.01
R0974s1	R0974s1TS460_1	84.78	86.59	1.81	65.58	69.2	3.62	0.77	2.1	1.33	0.69	0.72	0.03
R0976-D1	R0976-D1TS460_1	86.25	85	-1.25	68.96	66.25	-2.71	3.68	2.96	0.72	0.74	0.73	-0.01
R0976-D2	R0976-D2TS460_1	83.06	82.06	-1	64.92	61.7	-3.22	3.71	2.98	-0.7	0.73	0.73	0
R0977-D2	R0977-D2TS460_1	75	75.73	0.73	54.54	54.29	-0.25	3.28	2.6	0.68	0.65	0.65	0
R0979	R0979TS460_1	70.65	70.92	0.27	55.43	55.98	0.55	0.71	0.71	0	0.84	0.84	0
R0981	R0981-D3TS460_1	52.46	51.85	-0.61	31.77	31.28	-0.49	3.93	3.59	0.34	0.43	0.42	-0.01
R0981	R0981-D4TS460_1	62.39	64.19	1.8	45.05	46.17	1.12	1.33	2.64	1.31	0.51	0.52	0.01
R0981-D5	R0981-D5TS460_1	60.83	60.63	-0.2	42.32	41.93	-0.39	3.2	3.02	0.18	0.48	0.48	0
R0982-D2	R0982-D2TS460_1	68.75	67.42	-1.33	49.81	47.35	-2.46	3.52	2.15	1.37	0.52	0.5	-0.02
R0986s1	R0986s1TS460_1	80.16	78.8	-1.36	59.24	57.88	-1.36	0.55	1.35	0.8	0.69	0.69	0
R0986s2	R0986s2TS460_1	70.48	68.71	-1.77	49.35	47.74	-1.61	1.66	1.26	-0.4	0.61	0.59	-0.02
R0989-D1	R0989-D1TS460_1	50.75	48.69	-2.06	34.33	32.09	-2.24	3.41	2.23	1.18	0.49	0.44	-0.05
R0992	R0992TS460_1	81.78	80.84	-0.94	65.42	63.32	-2.1	0.86	1.6	0.74	0.68	0.67	-0.01
R0993s2	R0993s2TS460_1	71.94	71.17	-0.77	50.77	50	-0.77	3	2.66	-0.3	0.53	0.6	0.07
R0996-D4	R0996-D4TS460_1	70.3	67.86	-2.44	52.63	50	-2.63	3.43	2.46	0.97	0.6	0.6	0
R0996-D5	R0996-D5TS460_1	73.55	73.97	0.42	55.99	55.37	-0.62	3.14	2.23	0.91	0.64	0.65	0.01
R0996-D7	R0996-D7TS460_1	71.07	71.61	0.54	54.65	55.72	1.07	2.83	2.5	0.33	0.63	0.63	0

R0997	R0997TS460_1	64.32	65.68	1.36	41.76	43.38	1.62	3.49	3.51	0.02	0.55	0.55	0
R0999-D3	R0999-D3TS460_1	75.14	73.75	-1.39	54.44	52.23	-2.21	3.2	2.51	0.69	0.66	0.66	0
R1001	R1001TS460_1	73.02	73.2	0.18	52.7	53.05	0.35	0.7	1.07	0.37	0.69	0.69	0
R1002	R1002-D2TS460_1	88.14	88.98	0.84	72.88	74.16	1.28	3.38	3.22	-0.6	0.69	0.7	0.01
R1004-D2	R1004-D2TS460_1	78.57	81.17	2.6	60.39	64.61	4.22	3.34	2.65	0.69	0.65	0.68	0.03
R1016	R1016TS460_1	81.06	80.32	-0.74	62.62	61.39	-1.23	0.88	2.43	1.55	0.71	0.71	0
	The cumulative scores	2083.05	2076.33	-6.72	1537.19	1523.31	13.88	69.78	68.33	1.45	18.04	18.06	0.02

Table 3. 8 The performance of our refinement pipeline for all refinement targets according to the GDT-TS, GDT-HA, Molprobity and lDDT scores versus the starting model.

The 3D models were generated by our refinement pipeline, and the best-refined 3D model selected by ModFOLD7 was submitted during CASP13. Higher GDT-HA, GDT-TS, lDDT and lower Molprobity scores are better. Data are from http://www.predictioncenter.org/casp13/

CASP Target Category	Submitted vs Starting
GDT_TS Score	0.906
GDT-HA Score	0.9682
Molprobity Score	0.2548
lDDT Score	0.4614

Table 3. 9 Calculated pairwise p-values for the score of submitted models versus the score of starting models on the CASP13 refinement targets according to the GDT-TS, GDT-HA, Molprobity and lDDT scores

H_o: The scores of submitted models are equal or lower in quality than the score of starting models. H_1: The scores of submitted models are higher quality models than the score of starting models. P-values ≤ 0.05 indicate significant statistical differences (in boldface, higher GDT-TS, GDT-HA, lDDT scores and lower Molprobity scores are better).

3.4.2 The Comparison of the Different Restraint Strategies

The fixed and gradual local quality assessment guided restraint strategies were compared in terms of performance using the 32 regular targets (with publicly available structures), according to the GDT-HA, Molprobity and lDDT scores. Of the regular targets, 12 out of 32 targets were designated as TBM targets, 11 as TBM/FM and 9 as FM. As we saw in Chapter 2, the target categories of the initial structures were also found here to be a significant factor affecting the performance of the restraint strategies, in terms of the determination of the restraint and unrestrained regions. Therefore, performance of the restraint strategies was also analysed according to the target categories.

Using the local quality assessment guided fixed restraints prevented more of the 3D models from deviating from the native basin and increased the quality of the initial structures compared to the original MD-based protocol of ReFOLD, as shown in Chapter 2. The application of the local quality assessment guided restraint strategy is based on the determination of a threshold by taking into consideration the distribution of the per-residue error score produced by the ModFOLD server. In CASP13, it was noticed that applying the local quality assessment guided fixed restraint strategy may not be sufficient for certain FM/TBM targets. For FM/TBM targets, one or more domains may be designated as TBM targets, while others may be designated as FM. In this situation it is hard to determine a one-size-fits-all threshold. The per-residue error scores produced by the ModFOLD server are predictions of the likely distances for each residue (in Ångströms) from the native structure. In other words, if the per-residue error score is low in one domain or region, then the residues are much closer to the native state compared to the higher per-residue error scores in another domain/region. Therefore, the gradual restraint was proposed as the restraint should be applied according to the degree of need for each residue in each domain.

The fixed and gradual restraints strategies performed quite similar to each other for the FM/TBM targets according to the cumulative minimum, mean, and maximum GDT-HA scores (∑GDT-HAmin of 3.2011, ∑GDT-HAmean of 3.3788, and ∑GDT-HAmax of 3.5394 versus ∑GDT-HAmin of 3.1952, ∑GDT-HAmean of 3.3686, and ∑GDT-HAmax of 3.5594 versus ∑GDT-HAstarting of 3.4118) (Appendix 19). Both protocols also managed to improve the quality of all starting models (Figure 3.4-3.5 and Appendix 19-21).

In Figure 3.5, it can also be seen that despite having similar performances according to the observed scores, the population of the improved models generated by the gradual restraint strategy (~25.89%) is much higher than the fixed restraint strategy (~15.90%) compared to the starting models for the FM/TBM targets (Appendix 19). Therefore, for the population of models generated by the gradual restraint strategy there is a higher probability of randomly selecting an improved 3D model in comparison with the population of models generated by the fixed restraint strategy, due to the higher population of the improved models (Appendix 19).

Both restraint protocols showed similar characteristics in their ability to increase the accuracy of the FM and TBM targets according to the cumulative minimum, mean, and maximum GDT-HA scores as in Figure 3.6-3.7, and Appendix 22-28. It is also worthy of note that the gradual restraint strategy performed much better than the fixed restraint protocol in terms of the population improved 3D models for FM targets (~45.0% versus ~40.45%) (Figure 3.6, and Appendix 25-28) and TBM targets (~34.14% versus ~32.063%) (Figure 3.7, and Appendix 22-24). The gradual restraint performed better for FM targets, and roughly half of the generated models are improved compared to the starting models. This result implies that the gradual restraint strategy may boost the performance of our overall prediction pipeline for the prediction of FM targets (Figure 3.6, and Appendix 25-28). Furthermore, using the gradual restraints also increased the population of the improved models for the TBM targets, which are notably more challenging targets for refinement approaches (Figure 3.7, and Appendix 22-24).

The GDT-HA score based on the C-alphas superposition of the native structure with the predicted structures, and it has been used as the major scoring method in the refinement pipeline by CASP. However, all atoms are not taken into account for the calculation of the GDT-HA score. Therefore, the Molprobity score was used to compare the restraint strategies by considering all atoms. Both protocols managed to increase the accuracy of all targets compared to the starting models, and the gradual restraint strategy showed a better performance than the local quality assessment guided MD-based protocol as in Appendix 29-32.

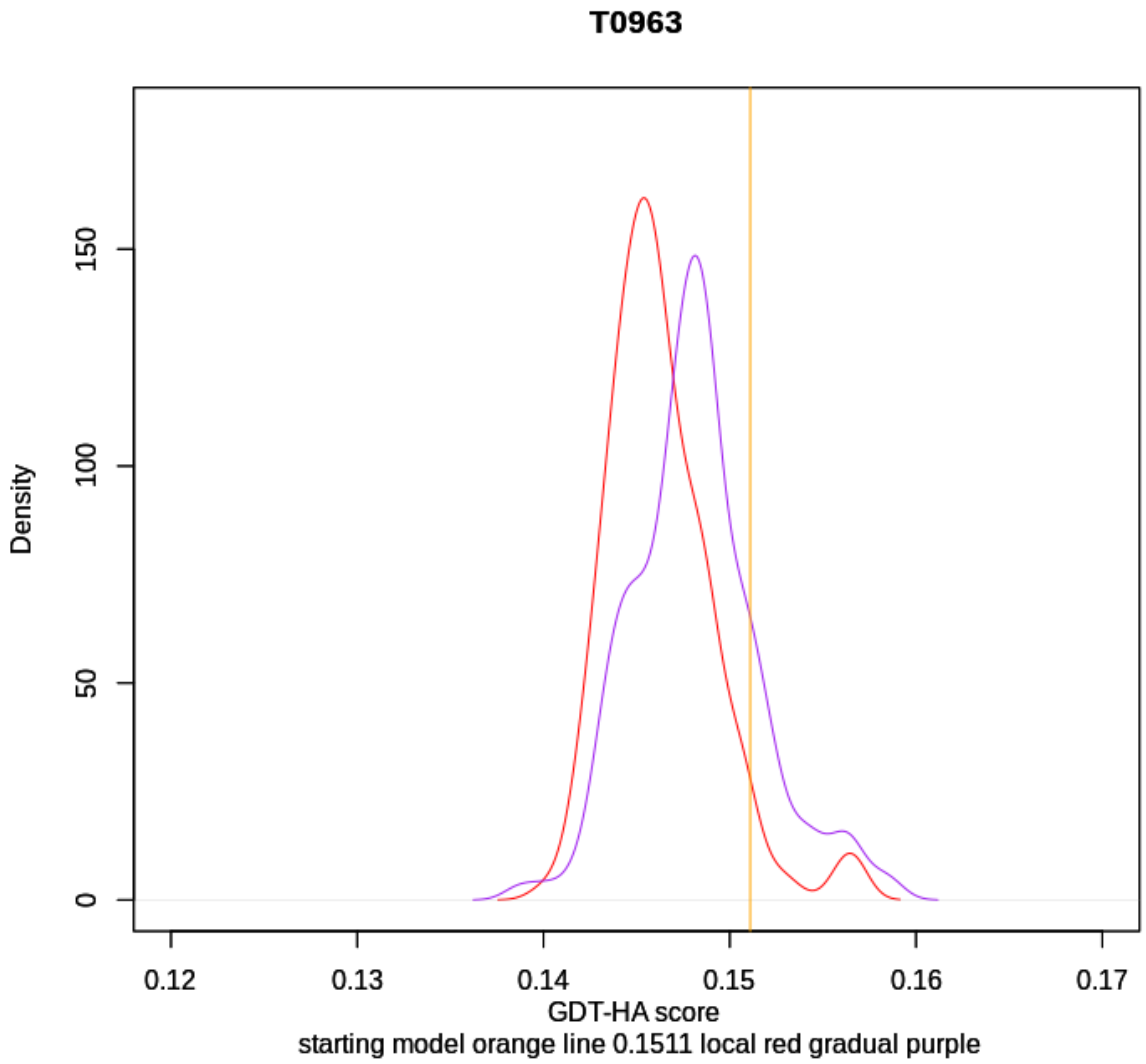

Figure 3. 4 A comparison of the gradual restraint strategy and fixed restraint strategy on an FM/TBM targets according to the GDT-HA score.

Performance of methods on T0963 (an FM/TBM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.1511, and higher GDT HA scores are better)

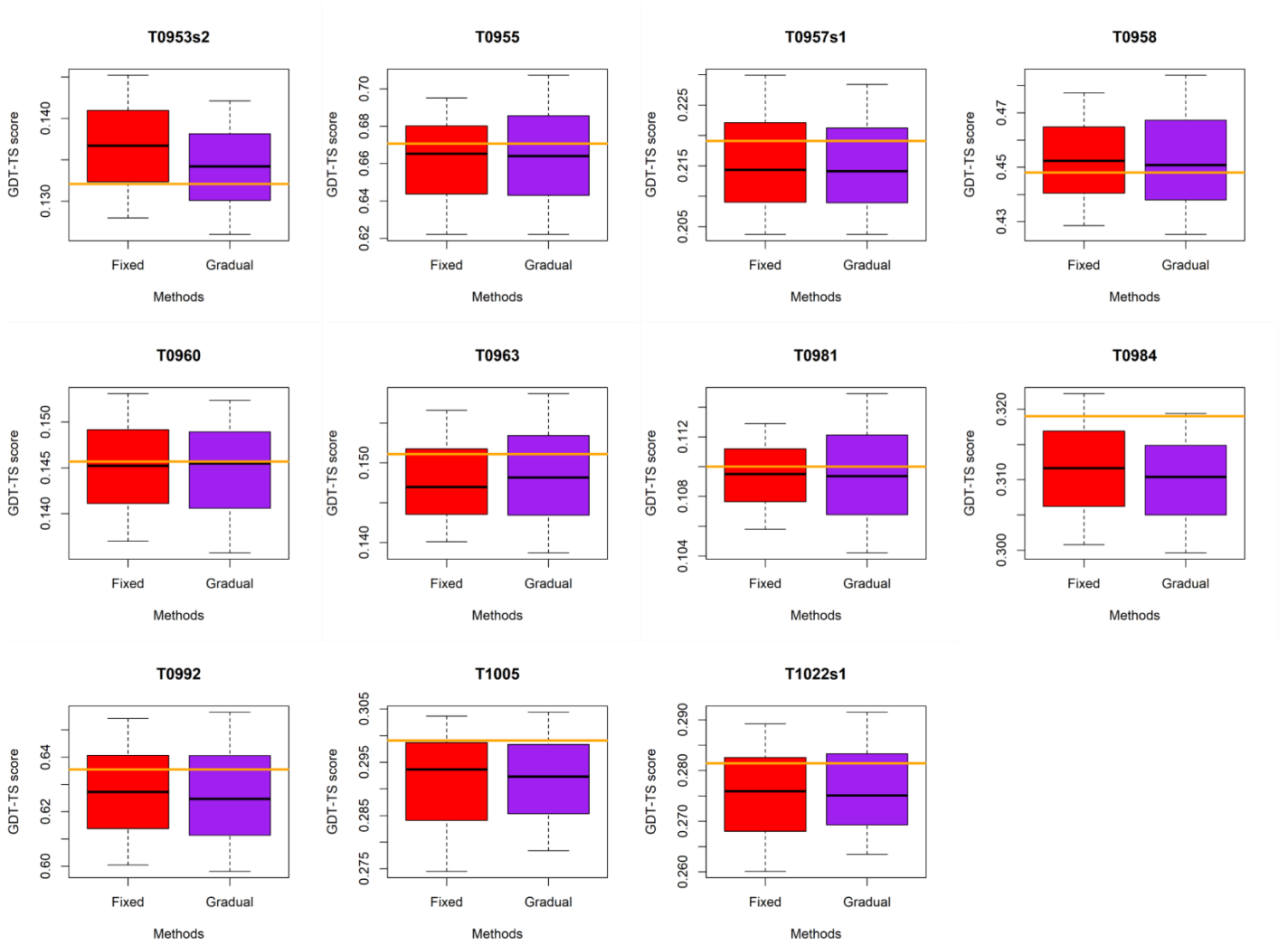

Figure 3. 5 A comparison of the fixed restraint strategy and the gradual restraint strategy on the CASP13 FM/TBM targets according to the GDT-HA score.

The red bars represent the scores of models generated using the fixed restraint strategy, purple bars represent models generated using the gradual restraint strategy, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

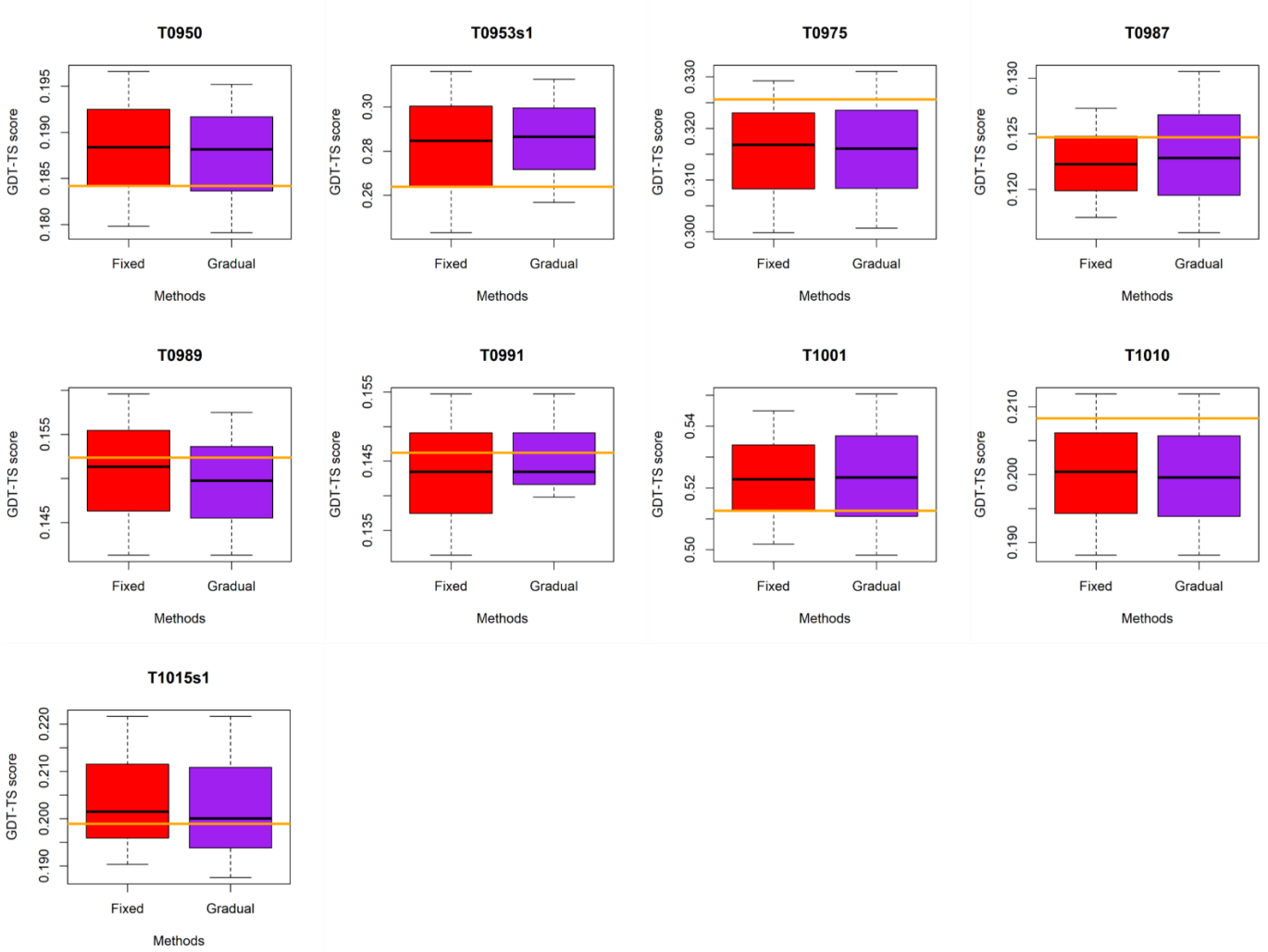

Figure 3. 6 A comparison of the fixed restraint strategy and the gradual restraint strategy on the CASP13 FM targets according to the GDT-HA score.

The red bars represent the scores of models generated using the fixed restraint strategy, purple bars represent models generated using the gradual restraint strategy, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

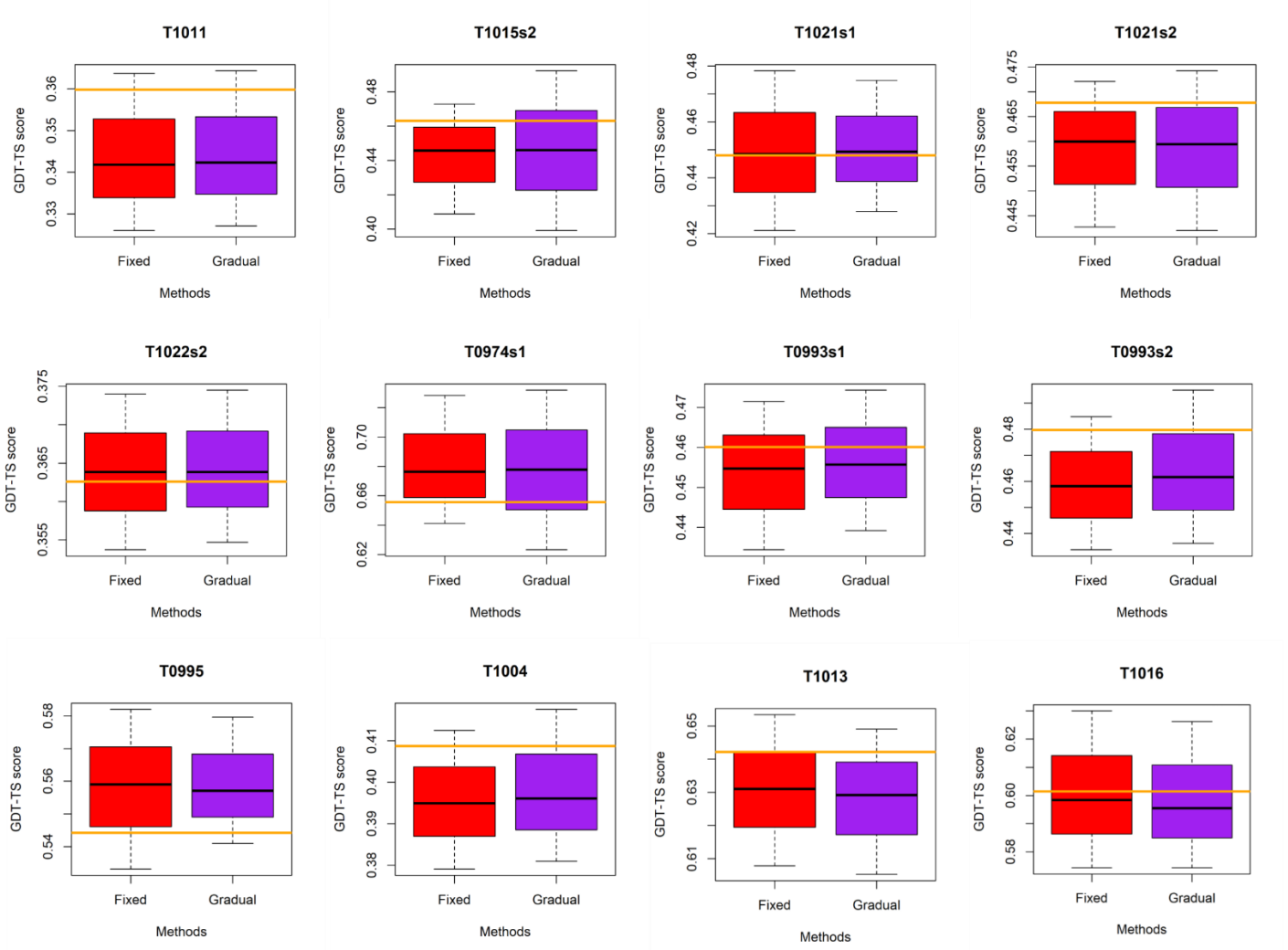

Figure 3. 7 A comparison of the fixed restraint strategy and the gradual restraint strategy on the CASP13 TBM targets according to the GDT-HA score.

The red bars represent the scores of models generated using the fixed restraint strategy, purple bars represent models generated using the gradual restraint strategy, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

3.4.3 The Refinement of SARS-COV-2 Protein Models for CASP Commons COVID-19 2020, and Using the Gradual Restraint Strategy

The COVID-19 pandemic is an unprecedented global challenge. In 2020, the final year of this PhD project, the worldwide scientific community has been working relentlessly to better understand SARS-CoV-2, its molecular mechanisms and find viable vaccines and treatments. The CASP Community has focused on the prediction of the structures of SARS-CoV-2 proteins and domains where no experimental data or obvious templates exist. Therefore, each of the ten CASP Commons targets were classified as FM targets. The aim of the initiative is to provide more complete knowledge of the structures so we can better understand their functions for the development of possible treatments, drug targeting methods and vaccines. Our group contributed to the effort; we provided 3D models using our IntFOLD server, and we quality assessed the predicted models from all groups using ModFOLD8. For each of the 10 SARS-CoV-2 target proteins, the best-predicted 3D model identified by ModFOLD8 was refined using the gradual restraint strategy in an attempt to further improve the quality. The top five refined models for each of the ten targets were then selected by ModFOLD8 and submitted by our "McGuffin" group.

The CASP assessors also used a pair-wise comparison based on LDDT and GDT-TS scores for the evaluation of the 3D models due to non-availability of the native structure. The global consensus LDDT and GDT-TS scores were also calculated using the similarity between each pair of models and considering the cumulative similarity of a model to all other models. The CASP assessors noted that their higher evaluation scores do not necessarily always mean better models, but they are a good indication of quality and they show high similarity to others.

According to the CASP assessment of model accuracy, our refinement pipeline performed well at modelling the ten SARS-CoV-2 targets, using both the global consensus LDDT and GDT-TS scores. The refinement protocol also managed to provide half of the top 10 models for C1901, C1902, C1903, C1904, and C1905, and almost a quarter of the top 20 models for C1906, C1908, and C1909 according to the initial CASP official estimates of model accuracy (Tables 3.10-3.17). These initial results indicate that our pipeline, which includes the gradual restraint strategy, is competitive with the many different approaches that participated in the CASP Commons COVID-19 2020 initiative.

The refinement of the ten SARS-CoV-2 targets highlights the importance of applying protein structure prediction pipelines to tackle real-world problems. Based on this it is likely that our McGuffin group pipeline, which integrates the gradual restraint strategy, has provided some of the most accurate models of the unknown SARS-CoV-2 proteins to date (Tables 3.10-3.17).

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1		MULTICOM	0.285	0.14
2	C1901TS044_1	FEIGLAB	0.287	0.14
3	**C1901TS213_1**	**McGuffin**	**0.291**	**0.139**
4	**C1901TS213_3**	**McGuffin**	**0.29**	**0.139**
5	**C1901TS213_5**	**McGuffin**	**0.29**	**0.139**
6	**C1901TS213_2**	**McGuffin**	**0.29**	**0.139**
7	**C1901TS213_4**	**McGuffin**	**0.289**	**0.139**
8	C1901TS228_1	DellaCorteLab	0.288	0.136
9	C1901TS273_1	Takeda-Shitaka-Lab	0.284	0.128
10	C1901TS215_5	PerezLab_Gators	0.252	0.111
11	C1901TS215_3	PerezLab_Gators	0.251	0.111
12	C1901TS215_4	PerezLab_Gators	0.247	0.111
13	C1901TS215_1	PerezLab_Gators	0.248	0.11
14	C1901TS215_2	PerezLab_Gators	0.249	0.11
15	C1901TS438_5	Destini	0.248	0.108
16	C1901TS438_2	Destini	0.241	0.107
17	C1901TS438_1	Destini	0.246	0.107
18	C1901TS438_4	Destini	0.245	0.107
19	C1901TS152_2	MULTICOM	0.276	0.105
20	C1901TS413_3	TFold-server	0.273	0.105

Table 3. 10 Official CASP results for C1901 (638 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1902TS413_4	TFold-server	0.344	0.187
2	**C1902TS213_1**	**McGuffin**	**0.347**	**0.182**
3	**C1902TS213_2**	**McGuffin**	**0.348**	**0.182**
4	**C1902TS213_4**	**McGuffin**	**0.347**	**0.182**
5	**C1902TS213_3**	**McGuffin**	**0.348**	**0.181**
6	**C1902TS213_5**	**McGuffin**	**0.346**	**0.181**
7	C1902TS152_4	MULTICOM	0.346	0.18
8	C1902TS152_2	MULTICOM	0.346	0.173
9	C1902TS413_1	TFold-server	0.348	0.173
10	C1902TS273_2	Takeda-Shitaka-Lab	0.33	0.169
11	C1902TS413_3	TFold-server	0.331	0.168
12	C1902TS152_3	MULTICOM	0.332	0.168
13	C1902TS438_3	Destini	0.369	0.168
14	C1902TS438_4	Destini	0.352	0.167
15	C1902TS413_2	TFold-server	0.336	0.167
16	C1902TS152_5	MULTICOM	0.352	0.167
17	C1902TS413_5	TFold-server	0.345	0.167
18	C1902TS438_1	Destini	0.36	0.167
19	C1902TS438_2	Destini	0.37	0.167
20	C1902TS438_5	Destini	0.362	0.165

Table 3. 11 Official CASP results for C1902 (500 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1903TS044_1	FEIGLAB	0.345	0.212
2	C1903TS301_1	FEIGLAB-S	0.345	0.212
3	**C1903TS213_5**	**McGuffin**	**0.348**	**0.21**
4	**C1903TS213_4**	**McGuffin**	**0.349**	**0.21**
5	**C1903TS213_2**	**McGuffin**	**0.347**	**0.21**
6	**C1903TS213_1**	**McGuffin**	**0.347**	**0.21**
7	C1903TS438_1	Destini	0.353	0.209
8	C1903TS438_4	Destini	0.352	0.209
9	C1903TS438_5	Destini	0.353	0.209
10	C1903TS438_3	Destini	0.35	0.209
11	**C1903TS213_3**	**McGuffin**	**0.347**	**0.209**
12	C1903TS438_2	Destini	0.351	0.207
13	C1903TS228_1	DellaCorteLab	0.347	0.207
14	C1903TS152_2	MULTICOM	0.363	0.203
15	C1903TS413_5	TFold-server	0.356	0.203
16	C1903TS152_1	MULTICOM	0.358	0.203
17	C1903TS247_1	AWSEM-Suite-Commons	0.339	0.201
18	C1903TS247_2	AWSEM-Suite-Commons	0.339	0.2
19	C1903TS413_3	TFold-server	0.356	0.2
20	C1903TS152_3	MULTICOM	0.342	0.199

Table 3. 12 Official CASP results for C1903 (290 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1904TS401_1	FEIGLAB-R	0.347	0.183
2	C1904TS273_1	Takeda-Shitaka-Lab	0.31	0.151
3	C1904TS044_1	FEIGLAB	0.312	0.151
4	**C1904TS213_2**	**McGuffin**	**0.314**	**0.15**
5	**C1904TS213_4**	**McGuffin**	**0.313**	**0.15**
6	**C1904TS213_5**	**McGuffin**	**0.314**	**0.15**
7	C1904TS152_2	MULTICOM	0.309	0.15
8	**C1904TS213_3**	**McGuffin**	**0.313**	**0.15**
9	**C1904TS213_1**	**McGuffin**	**0.314**	**0.149**
10	C1904TS228_1	DellaCorteLab	0.312	0.146
11	C1904TS215_5	PerezLab_Gators	0.271	0.122
12	C1904TS215_2	PerezLab_Gators	0.266	0.121
13	C1904TS215_4	PerezLab_Gators	0.264	0.12
14	C1904TS413_5	TFold-server	0.298	0.119
15	C1904TS152_3	MULTICOM	0.298	0.119
16	C1904TS215_3	PerezLab_Gators	0.265	0.117
17	C1904TS438_5	Destini	0.301	0.117
18	C1904TS215_1	PerezLab_Gators	0.264	0.117
19	C1904TS438_1	Destini	0.303	0.117
20	C1904TS438_3	Destini	0.303	0.117

Table 3. 13 Official CASP results for C1904 (686 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1905TS413_1	TFold-server	0.275	0.201
2	**C1905TS213_4**	**McGuffin**	**0.285**	**0.2**
3	**C1905TS213_5**	**McGuffin**	**0.287**	**0.2**
4	**C1905TS213_3**	**McGuffin**	**0.287**	**0.2**
5	**C1905TS213_2**	**McGuffin**	**0.286**	**0.199**
6	**C1905TS213_1**	**McGuffin**	**0.287**	**0.199**
7	C1905TS401_1	FEIGLAB-R	0.298	0.198
8	C1905TS413_3	TFold-server	0.274	0.196
9	C1905TS152_1	MULTICOM	0.266	0.19
10	C1905TS413_5	TFold-server	0.264	0.19
11	C1905TS152_2	MULTICOM	0.259	0.188
12	C1905TS413_4	TFold-server	0.257	0.188
13	C1905TS413_2	TFold-server	0.267	0.187
14	C1905TS438_5	Destini	0.307	0.178
15	C1905TS438_3	Destini	0.305	0.178
16	C1905TS301_1	FEIGLAB-S	0.311	0.177
17	C1905TS044_1	FEIGLAB	0.311	0.177
18	C1905TS152_3	MULTICOM	0.279	0.176
19	C1905TS196_1	ntsu	0.283	0.175
20	C1905TS102_2	D-Haven	0.283	0.175

Table 3. 14 Official CASP results for C1905 (275 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1906TS301_1	FEIGLAB-S	0.438	0.327
2	C1906TS044_1	FEIGLAB	0.438	0.327
3	C1906TS152_3	MULTICOM	0.434	0.327
4	C1906TS438_3	Destini	0.43	0.324
5	**C1906TS213_4**	**McGuffin**	**0.438**	**0.324**
6	C1906TS152_2	MULTICOM	0.428	0.323
7	**C1906TS213_3**	**McGuffin**	**0.439**	**0.323**
8	C1906TS438_2	Destini	0.43	0.323
9	**C1906TS213_1**	**McGuffin**	**0.437**	**0.323**
10	C1906TS438_1	Destini	0.433	0.323
11	**C1906TS213_2**	**McGuffin**	**0.439**	**0.323**
12	C1906TS438_5	Destini	0.43	0.323
13	C1906TS152_1	MULTICOM	0.427	0.322
14	C1906TS413_1	TFold-server	0.418	0.322
15	C1906TS413_2	TFold-server	0.413	0.321
16	C1906TS413_3	TFold-server	0.424	0.321
17	C1906TS438_4	Destini	0.428	0.32
18	**C1906TS213_5**	**McGuffin**	**0.438**	**0.32**
19	C1906TS413_4	TFold-server	0.421	0.319
20	C1906TS299_5	FALCON-DeepFolder	0.442	0.317

Table 3. 15 Official CASP results for C1906 (222 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1908TS152_2	MULTICOM	0.314	0.315
2	C1908TS413_3	TFold-server	0.313	0.314
3	C1908TS152_1	MULTICOM	0.316	0.313
4	C1908TS273_2	Takeda-Shitaka-Lab	0.312	0.311
5	C1908TS413_2	TFold-server	0.313	0.311
6	C1908TS413_5	TFold-server	0.312	0.309
7	C1908TS438_1	Destini	0.315	0.309
8	**C1908TS213_5**	**McGuffin**	**0.326**	**0.309**
9	C1908TS273_3	Takeda-Shitaka-Lab	0.311	0.309
10	**C1908TS213_2**	**McGuffin**	**0.327**	**0.309**
11	**C1908TS213_4**	**McGuffin**	**0.327**	**0.309**
12	C1908TS299_5	FALCON-DeepFolder	0.309	0.308
13	C1908TS278_5	FALCON	0.309	0.308
14	**C1908TS213_3**	**McGuffin**	**0.326**	**0.308**
15	C1908TS438_3	Destini	0.314	0.308
16	**C1908TS213_1**	**McGuffin**	**0.326**	**0.307**
17	C1908TS413_1	TFold-server	0.304	0.307
18	C1908TS438_2	Destini	0.316	0.307
19	C1908TS438_4	Destini	0.312	0.306
20	C1908TS438_5	Destini	0.313	0.306

Table 3. 16 Official CASP results for C1908 (121 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

Ranking	Model Name	Predictor	LDDT_cons	GDT_TS_cons
1	C1909TS438_5	Destini	0.453	0.534
2	C1909TS123_1	IntFOLD6	0.439	0.533
3	C1909TS278_1	FALCON	0.447	0.533
4	C1909TS299_1	FALCON-DeepFolder	0.447	0.533
5	**C1909TS213_1**	**McGuffin**	**0.449**	**0.53**
6	C1909TS123_3	IntFOLD6	0.441	0.53
7	C1909TS152_4	MULTICOM	0.444	0.53
8	C1909TS158_2	FALCON-TBM	0.438	0.529
9	C1909TS369_1	Yang	0.42	0.527
10	**C1909TS213_4**	**McGuffin**	**0.45**	**0.526**
11	C1909TS309_1	Zhang-TBM	0.435	0.526
12	**C1909TS213_5**	**McGuffin**	**0.447**	**0.525**
13	**C1909TS213_3**	**McGuffin**	**0.448**	**0.525**
14	C1909TS309_5	Zhang-TBM	0.435	0.525
15	C1909TS369_5	Yang	0.429	0.525
16	**C1909TS213_2**	**McGuffin**	**0.451**	**0.525**
17	C1909TS369_4	Yang	0.416	0.524
18	C1909TS299_4	FALCON-DeepFolder	0.445	0.523
19	C1909TS152_2	MULTICOM	0.444	0.523
20	C1909TS278_4	FALCON	0.445	0.523

Table 3. 17 Official CASP results for C1909 (38 residues) according to the consensus GDT-TS and LDDT scores for the top 20 models.
The table is sorted by GDT_TS_cons score. Data are from https://predictioncenter.org/caspcommons/models_consensus2.cgi

3.5 Conclusions

The original ReFOLD (Shuid et al., 2017) method was developed by the McGuffin group to increase the accuracy of the predicted 3D models through the integration of a rapid MD-based protocol, inspired by that of Feig and Mirjalili (Feig & Mirjalili, 2016; Mirjalili et al., 2014; Mirjalili & Feig, 2013). However, significant deviations from the native structure were observed for the refinement of many of the predicted 3D models, particularly for TBM targets, due to lack of reliable guidance during the MD simulations. The per-residue accuracy score produced by ModFOLD6 (Maghrabi & McGuffin, 2017) was proposed to be used as a guide for the original MD-based protocol of ReFOLD, in order to avoid such structural deviations. The local quality assessment guided fixed restraint strategy was therefore devised (Chapter 2), which managed to prevent the MD models from structural drifts by applying a single restraint threshold that was based on the distribution of the per-residue accuracy score produced by ModFOLD6 .

The fixed restraint MD-based protocol was further upgraded using ModFOLD7 (Maghrabi & McGuffin, 2019) to guide the MD simulations and select the best-predicted 3D model. This upgraded version of ReFOLD (ReFOLD2) was used to improve the quality of 3D models for both the CASP13 regular and refinement targets and it played a key role in the success of the McGuffin group in the competition.

Our CASP13 prediction pipeline consisted of three main stages; the prediction of the 3D models by the IntFOLD server (McGuffin et al., 2019), the local and global assessment of the predicted 3D server models by ModFOLD7 (Maghrabi & McGuffin, 2019), and the refinement of the best-predicted server models and refinement targets using the upgraded version of ReFOLD including the fixed restraint MD-based protocol. The refinement pipeline also performed well in terms of improving the quality of the TBM and FM domains and it ranked in the top 10 approaches in the regular prediction and refinement CASP13 categories (Read et al., 2019). Although ModFOLD7 was not specifically developed for the selection of the improved models in the refinement pipeline, the performance of MoldFOLD7 with regard to the selection of the refined models was evaluated.

It is promising that ModFOLD7 managed to identify improved models in comparison with the initial structure for roughly half of the refinement targets, according to the CASP13 official results.

The CASP13 regular targets included multi-domain proteins that were relatively larger compared to the previous CASP experiments (Read et al., 2019). Therefore, the determination of a fixed restraint threshold for the application of the local quality assessment guided MD-based protocol is less applicable. For the multi-domain structures especially for FM/TBM targets the one-size-fits-all approach used by the fixed restraints is less appropriate. For instance, a domain predicted by TBM is probably more accurate compared to FM due to usage of available structures, thereby applying a fixed threshold according to the distribution of the per-residue accuracy scores may not be suitable. For this reason, we proposed a gradual restraint strategy based on the per-residue accuracy score, which considered the degree of refinement required for each residue during the MD simulations.

The fixed and gradual restraint strategies showed good performance with both approaches successfully increasing the accuracy of the initial structures according to the GDT-HA scores. The application of the gradual restraints improved more models overall compared to the fixed restraint strategy, particularly for the FM/TBM targets, with ~25.89% of models improved versus ~15.90% respectively. For all targets, the overall percentage of the improved models was also higher using the gradual restraint strategy, with ~34.36% of models improved versus ~28.86% using fixed restraints.

Beyond the CASP experiments, our protein structure prediction methods are used by researchers worldwide to predict structures that will help to solve real-world biological problems. The IntFOLD server (McGuffin et al., 2019) was developed to predict 3D models, the ModFOLD server (Maghrabi & McGuffin, 2019) is used to evaluate of the predicted 3D models by providing local and global scores and ReFOLD aims at improving the quality of the best-predicted 3D models which includes automation of a rapid MD based protocol. The automation of more accurate MD-based protocols will play an important part in the improvement of our servers and will make more accurate models available for biological research. From the results in this chapter, it is evident that the application of the gradual restraint is found to be more effective than using a fixed restraint

strategy. Therefore, we plan to upgrade the ReFOLD server with the integration of the gradual restraint strategy in the short-term.

The value of *in silico* modelling of protein tertiary structures was emphasised by its application in CASP Commons to shed light on the difficult unknown structures of the SARS-CoV-2 virus. This community wide effort aims to take a more comprehensive look at the structures to comprehend their functions and interactions within the cells. Our group participated in the prediction of the SARS-CoV-2 targets using the IntFOLD server, assessment of the predicted 3D models by ModFOLD8, and the refinement of the best-predicted server model selected by ModFOLD8 utilising the gradual restraint strategy described in this chapter. Our manual prediction group (McGuffin) provided a significant number of the top 10 models for the SARS-CoV-2 targets according to the official CASP estimates of model accuracy. The gradual restraint strategy was also used in our manual prediction pipelines for the CASP14 experiment (at the time of writing the prediction part of the experiment is over but the results are not yet available).

The fixed and gradual restraint strategies based on the local quality estimation were applied for the refinement of the whole protein structures. The next chapter will focus on the refinement of the predicted protein-ligand binding site rather than the whole structure in order to improve the quality of the specific regions.

Chapter 4 The Refinement of Predicted Protein-Ligand Binding Sites

4.1 Background

Complete understanding of the biological functions of proteins is directly related to our knowledge of their interacting partners. Protein-ligand interactions play a critical role in the functionality of the protein structures, as well as other larger interacting partners. Determining protein-ligand binding sites allows us to progress towards elucidating molecular mechanisms and improves our understanding of protein interactions with drugs (Rhizobium, 2013; Roche et al., 2012a, 2013a; Roche, Buenavista, et al., 2011; Roche, Tetchner, et al., 2011). The determination of the protein-ligand interactions via *in vitro* methods can be costly and time-consuming (McGuffin, 2008a; Moult et al., 2007, 2016; Roche et al., 2014; Roche & McGuffin, 2016b; Schmidt et al., 2011). In addition, in some situations it may not be practical or possible to bridge the protein sequence-structure knowledge gap using experimental methods. The use of *in silico* methods can help us to predict the function of the proteins and their interactions with ligands and may be a useful alternative approach to aid the discovery of treatment pathways for human and animal diseases.

In silico methods for the prediction of the protein-ligand binding sites can be divided into two main categories: sequence-based and structure-based methods (Rhizobium, 2013; Roche et al., 2012a, 2013a; Roche, Buenavista, et al., 2011; Roche, Tetchner, et al., 2011). Sequence-based methods are based on the sequence similarity of homologous proteins and data from evolutionary conservation by means of different sequence alignments, including pairwise and multiple sequence alignment (MSA) (Chen et al., 2014; Sankararaman et al., 2009, 2010; Wass & Sternberg, 2008; Wierschin et al., 2015; Ye et al., 2008; Yu et al., 2013, 2015). Most of the sequence-based methods rely on the interpretation of the MSA data for each residue function to identify conserved residues (Chen et al., 2014; López et al., 2007; Roche et al., 2015; Roche & McGuffin, 2016a, 2016b; Sankararaman et al., 2009, 2010; Talavera et al., 2009; Wass & Sternberg, 2008; Ye et al., 2008; Yu et al., 2013).

Structure-based methods additionally utilise the 3D information from predicted or experimentally determined 3D structures to predict protein-ligand binding sites. Structure-based methods can also be sub categorised as geometric-based methods and energetic-based approaches (Roche et al.,

2015; Roche & McGuffin, 2016a). While the principle of geometric-based methods is the identification of binding site pockets, energetic-based approaches identify the pockets by considering interaction energies (Brylinski & Skolnick, 2008; Cao & Li, 2014; Erdin et al., 2010; Fuller et al., 2015; Heo et al., 2014; Hernandez et al., 2009; Huang & Schroeder, 2006; Izidoro et al., 2015; Madabushi et al., 2002; Roche et al., 2013b; Roche, Tetchner, et al., 2011; Roche et al., 2012a; Yang, Roy, Zhang, et al., 2013; Zhu et al., 2014).

4.1.1 The FunFOLD Server

The McGuffin group developed the FunFOLD server to predict the protein-ligand binding sites as a structural template-based method. FunFOLD version 3, which is the current version of the FunFOLD server (Roche et al., 2012a, 2013b; Roche, Tetchner, et al., 2011) integrates the quality estimate scores from the FunFOLDQA method (see further description below) (Roche et al., 2012a). FunFOLD makes use of the top selected 3D model predicted for the target sequence by the latest version of the IntFOLD server (McGuffin et al., 2015, 2019) along with the list of identified structural templates. The main assumption of the FunFOLD method is that proteins with similar folds will often have similarly located binding sites. Thus, the fold templates identified within the modelling process are used in the prediction of the protein-ligand interaction sites. In the first stage, FunFOLD (Roche, Tetchner, et al., 2011) utilises the 3D model and the template lists (PDB IDs) generated by the IntFOLD (McGuffin et al., 2019) server as inputs, superposing the target 3D model with each of the structural templates that contain biologically relevant ligands. The TM-align (Zhang & Skolnick, 2005) method is used for the structural super positioning and the BioLip database (Yang, Roy, & Zhang, 2013) is used to identify ligands in the templates that are considered to be biologically relevant. The BioLip database combines computational and manual examinations of biologically relevant ligand entries including binding residues in the database, ligand-binding affinity (Yang, Roy, & Zhang, 2013), EC numbers and GO terms . For the superpositions only templates with TM-scores higher than 0.4 are considered. Next, the potential binding sites are detected by determining the contact distance between the target 3D structure and possible ligands including Van der Waals bonds less than 0.5 Å (Roche et al., 2012a,

2013b, 2015; Roche, Tetchner, et al., 2011; Roche & McGuffin, 2016a; Yang, Roy, & Zhang, 2013; Zhang & Skolnick, 2005).

There are some caveats when using structure-based approaches, such as FunFOLD, for the prediction of the protein-ligand interactions (Roche et al., 2015; Roche & McGuffin, 2016a). The first obvious limitation is if no predicted 3D models or experimental structures are available, and so in this case purely sequence-based methods might still be used, if sequence MSA profiles can provide sufficient information. Another issue is if none of the identified templates contain biologically relevant ligands even if templates exist that have the same folds with the target structure (Roche et al., 2015; Roche & McGuffin, 2016a). The prediction of the best quality structure may not always be possible by the prediction servers (such as IntFOLD), so this may also affect the accuracy of the predicted binding sites. Despite having these limitations, structure-based methods have made progress over the years and continue to play a role in function prediction from structure (Roche et al., 2015; Roche & McGuffin, 2016a).

4.1.2 Scoring Protein Ligand Binding Site Predictions

The Binding-site Distance Test (BDT) score and the Matthews Correlation Coefficient (MCC) scores have been used to assess the performance of predicted protein-ligand binding sites in the CASP (Gallo Cassarino et al., 2014; Schmidt et al., 2011) and CAMEO (Haas et al., 2013, 2018) experiments. The CASP category for the prediction of ligand binding site was introduced in CASP8 to predict binding site residues with the possible biologically relevant ligands (López et al., 2009). In CASP10, 13 out of 97 targets were found to be with biologically relevant ligands and the number of targets with the biologically relevant ligands was a limiting factor for the evaluation of the binding site prediction methods. Due to the lack of targets, the CASP category was subsequently moved to become part of the CAMEO experiment (López et al., 2009). Typically, CAMEO released 10-20 targets every week between January 2012 and April 2016 (Gallo Cassarino et al., 2014; Haas et al., 2013, 2018, 2019; López et al., 2009; Wu et al., 2018). Unfortunately, CAMEO-LB category was discontinued since 2016. However, CAMEO is

planning to start a new category which covers the modelling of protein complexes including the ligand-binding site prediction (Haas et al., 2018, 2019).

4.1.3 Observed Quality Scores

The MCC score is based on a statistical comparison of the list of observed binding residues with the list of predicted residues considering solely the residue numbers within the sequence, which are either assigned as true positives, false positives, true negatives, and false negatives (Matthews, 1975; Roche et al., 2010). The MCC score ranges from -1 to 1, and the scores around 0 indicate random selections. Higher MCC scores also represent better predictions. Using the MCC score may not be adequate for the predictions made by structure-based methods as the 3D structural information is not taken into consideration for its calculation (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a).To consider the 3D observed structures, the BDT score developed by the McGuffin group was utilised for the investigation of improvements in ligand-binding site predictions (Roche et al., 2010).

The BDT score was developed by the McGuffin group for the assessment of the predicted protein-ligand binding site predictions by considering the actually structural distance between predicted binding residues and observed binding residues according to the information in the observed 3D structure (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a). Unlike the MCC score, the BDT score is also calculated utilising the observed 3D structure coordinates. The range of the BDT score is from 0 to 1, and the score close to 1 indicates a more accurate binding site prediction (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a) . The BDT score was used in CASP9 (Schmidt et al., 2011) and CASP10 (Gallo Cassarino et al., 2014) along with the MCC score and is also used as a standard measurement in CAMEO. It is only possible to produce both the MCC and BDT scores when the observed structure is available with the bound ligand, hence these are observed measures of binding site model quality (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a).

4.1.4 Predicted Quality Scores

FunFOLDQA (Roche et al., 2012a) was also developed to assess the quality of the binding sites predicted by FunFOLD (Roche, Tetchner, et al., 2011) *prior* to the availability of experimental structures, by combining numerous sequence and structure based metrics using an artificial neural network and training it to learn the observed quality of the models according to the Binding-site Distance Test (BDT) (Roche et al., 2010) and Matthews Correlation Coefficient (MCC) (Matthews, 1975) scores. For the neural network architecture of FunFOLDQA (Roche et al., 2012a)**,** three layers and five features and eleven neurons were utilised (Roche et al., 2012a). The FunFOLD server provides the estimates of binding site accuracy according along with models of the predicted protein-ligand interactions (Roche & McGuffin, 2016a).

4.2 Aims and Objectives

The accuracy of predicted protein-ligand binding sites plays an important part in the wider adoption of 3D models of protein structures. FunFOLD3 provides a detailed prediction of the ligands and binding residues, which can be used to infer the function of modelled structures. Nevertheless, the accuracy of the modelled binding sites themselves may not always be adequate in order to accurately elucidate protein functions. Therefore, improving the quality of the modelled binding site regions might be a useful step towards a more complete atomic-level understanding of protein interactions.

The aim of the work presented in this chapter is to increase the accuracy of the FunFOLD3 predicted binding sites in protein 3D models by utilising the MD-based refinement protocol, which was initially developed in Chapter 2. The MD-based protocol will be used to improve the quality of the modelled binding site residues in order to fine tune the predicted interactions between residues and ligands. The predicted binding residues and their neighbouring residues will be the main focus in this refinement pipeline. It is postulated that by refining the local model quality of the binding residues and their neighbouring residues in the 3D models, which are used as inputs to FunFOLD3, we can likewise improve the observed quality of the FunFOLD3 binding site

predictions. Hence, the predicted binding site residues and their neighbouring residues in 3D models were highlighted for focused refinement, while the rest of the protein structure was restrained during the MD-simulations. The performance of the binding site refinement protocol was tested on CASP12 and CASP13 targets and we used the BDT and MCC scores to analyse the improvement in the quality of the predicted binding sites (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a; Zhang & Skolnick, 2005). The GDT-HA scores to measure the global effect on model quality (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a; Zhang & Skolnick, 2005).

Refined 3D models with targeted improvements in the quality of the binding sites may help us to more accurately determine the nature of the protein-ligand interactions and therefore shed light on protein functionality. Following the performance evaluation presented here, this novel binding site refinement pipeline will be integrated with future versions the FunFOLD and IntFOLD servers, providing improved binding site predictions in our freely accessible servers.

4.3 Materials and Methods

4.3.1 Data Collection

The 64 CASP12 and 82 CASP13 regular targets were used to test the new refinement protocol of the predicted protein-ligand binding sites. The best-predicted server models for the targets and their sequence and template lists (PDB IDs) were used to predict the protein-ligand binding sites using FunFOLD3. The best-predicted input server models were identified by the ModFOLD server and their template lists were generated by the IntFOLD server during CASP12 and CASP13 experiments. The amino acid sequences and the native structures were obtained from the CASP website (http://predictioncenter.org/download_area/). The TM-score and BDT score tools were also utilised to produce the GDT-HA, BDT, and MCC scores in order to evaluate the performance of the binding site-focused MD-based protocol (Matthews, 1975; Roche et al., 2010, 2015; Roche & McGuffin, 2016a; Zhang & Skolnick, 2005).

4.3.2 Computational Design

Our benchmarking of the refinement of the predicted binding sites in 3D models consists of three main stages: 1. the identification of the predicted and observed binding residues using FunFOLD3, 2. the refinement of the predicted binding sites in 3D models using the binding site focused MD-based protocol, and 3. the analysis of the resulting 3D models generated by the MD-based protocol using the different scoring measurements (Figure 4.1).

In the first stage, the FunFOLD3 standalone method (executable JAR file from https://www.reading.ac.uk/bioinf/downloads/FunFOLD3Package.tar.gz) was used to predict the protein-ligand interactions for each of the CASP12 and CASP13 regular targets during the time frames of the prediction experiments. Recent versions of Java, PyMOL (Delano L.W., 2002), the TM-align tool (Zhang & Skolnick, 2005) and the most updated the CIF chemical components files and the Biolip databases (Yang, Roy, & Zhang, 2013) were required to run FunFOLD3 method locally. The best-predicted server models ranked by the ModFOLD server (Maghrabi & McGuffin, 2017), the template list files (containing PDB IDs) generated by the IntFOLD server (McGuffin et al., 2019) , and the target sequences in FASTA format were all used as inputs for the FunFOLD3 standalone method. The FunFOLD3 method provides the lists of predicted binding residues and possible ligands which are likely to interact with the initial structure, along with 3D models of the protein-ligand complexes as its output. The proLigContacts standalone tool was also run on the native structures with ligands to identify observed binding residues (https://www.reading.ac.uk/bioinf/downloads/proLigContacts2.jar).

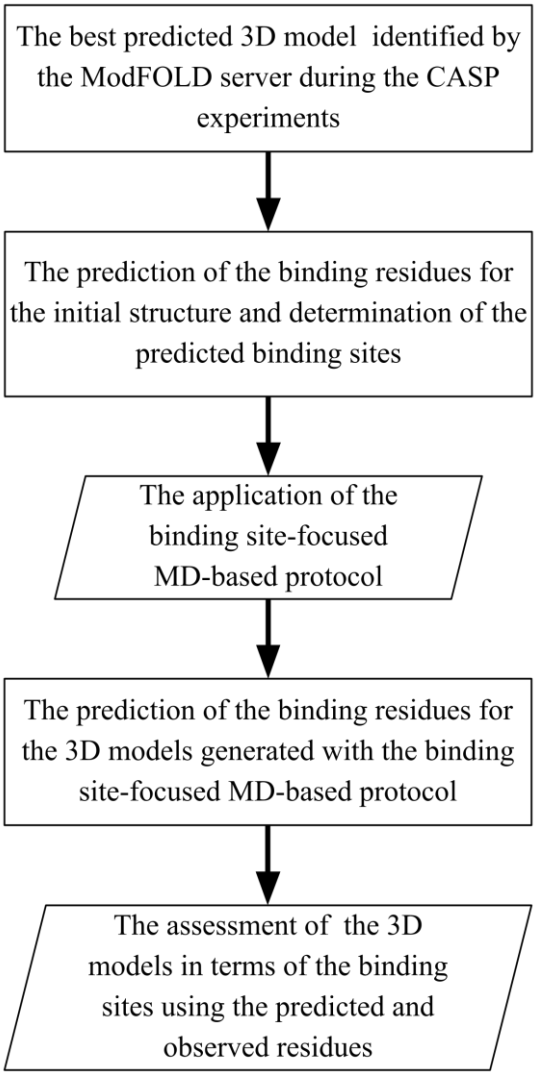

Figure 4. 1 Flowchart summarising the application of the binding site-focused MD-based protocol.

For each target, the best-predicted server model was firstly identified using the latest ModFOLD servers during the CASP12 and CASP13 experiments. The top models were then used as the initial input structures for FunFOLD3, which was then run to predict the binding site. The initial predicted binding site (the binding residues plus their neighbouring residues) was then located in the model. Following the determination of the binding site, the binding site-focused MD-based protocol was then applied to refine the site. Subsequently, FunFOLD3 was re-run, this time using the top refined 3D models as inputs, and new sets of binding residues were then predicted for each model. The predicted and the observed residues were then compared using the BDT and MCC scores and global model quality was scored using GDT-HA scores in the final stage.

After the prediction of the binding site using the initial 3D model, their neighbouring residues within 5 Å were determined, using a PyMOL script, in order to define the full binding site to be targeted for refinement. In order to focus the refinement process on the binding site in the initial

structure, a weak harmonic positional restraint (0.05 kCal/mol/Å^2) was applied to the rest of the protein structure, thereby allowing more movement in the binding site during the MD simulation (Mirjalili et al., 2014; Mirjalili & Feig, 2013). The MD simulations were carried out using NAMD 2.10 (Phillips et al., 2005) and the parameters which were optimised for the original MD-based protocol of ReFOLD, as described in Chapter 2 (Mirjalili et al., 2014; Mirjalili & Feig, 2013; Shuid et al., 2017). 164 refined 3D models were generated for each target using the binding-site focused MD-based protocol. Subsequently, FunFOLD3 was re-run in order to predict the binding residues for each refined model.

Following the prediction of the binding residues for each refined 3D model, the predicted binding residues and ligands were compared to the observed binding residues for the targets which had experimentally solved structures containing biologically relevant ligands. The biological relevance of ligands within solved structures was determined by investigating the available structure data, the ligand databases and literature.

The BDT tool (https://www.reading.ac.uk/bioinf/downloads/BDT.jar) was used to produce the BDT and MCC scores (Matthews, 1975; Roche et al., 2010) in order to assess the accuracy of the binding site residue predictions, based on each refined 3D model, compared with the observed binding residues. The GDT-HA score was also produced using the TM-score tool (Zhang & Skolnick, 2005) to analyse the accuracy of the 3D models compared with the native structures.

4.4 Results and Discussion

The CASP12 and CASP13 targets were analysed in terms of their biologically relevant binding sites using their observed structures, which were released in Protein Data Bank (PDB). After the analysis of the available observed structures, 9 targets were found to have coordinates for biologically relevant bound ligands, so these were analysed with regards to the refinement of the modelled binding sites. These targets were T0909 (5g5n), T0911 (6e9n), T0912 (5mqp), T0953s2 (6f45), T0954 (6cvz), T1009 (6dru), T1011 (6m9t), T1016 (6e4b) and T1018 (6n91). The observed binding site residues versus the initial predicted and the best refined binding site residues according

to the MCC score are compared in Table 4.1. More than one predicted and observed binding sites were determined for T0953s2, T0954, T1009, T1011, T1016, and T1018 and the best initial prediction was chosen for further refinement (Table 4.1).

The predicted binding sites were targeted for refinement using the binding site-focused MD-based protocol and the resulting residue prediction accuracy (BDT, MCC) and model quality scores (GDT-HA) are shown in Table 4.2, and Figure 4.2. The binding sites identified by FunFOLD3 for T0909, T0912, T0954 and T1011 were poorly predicted compared to those of T0911, T0953s2, T1009, T1016 and T1018 according to the BDT and MCC scores (Table 4.2 and Appendix 33-37). These targets and their binding sites will be investigated in this section in order to analyse the performance of the binding site-focused MD-based protocol.

Despite the low initial accuracy of the binding site residues predicted by FunFOLD3 for T0909, T0912, T0954 and T1011, following the targeted refinement of the modelled binding sites, the subsequent binding residue accuracy was increased according to the BDT and MCC scores (Table 4.1, Table 4.2). Although T0954 and T1011 were defined by the assessors as TBM targets, the lower BDT and MCC score might indicate the lack of availability of suitable templates with bound ligand, which are required to generate the binding site predictions (Table 4.2 and Appendix 35-36). T0909 and T0912 were categorised as FM/TBM targets, so the low accuracy of the binding site might be also related to the lack of the templates. Nonetheless, the binding site-focused MD-based protocol managed to generate improved models, for example, the T0909 refined models show better scores versus the initial models according to the BDT, MCC and GDT-HA scores (BDTmax of 0.03, MCCmax of -0.00605 and GDT-HAmax of 0.2913 versus BDTstarting of 0.0253, MCCstarting of -0.0067 and GDT-HAstarting of 0.2703) (Table 4.2, Figure 4.3, and Figure 4.4).

The higher GDT-HA maximum scores for most of the refined models indicate that the binding site-focused MD-based strategy can act to improve the global model quality even though the binding site residue predictions themselves are not highly accurate. However, it should be noted that the MD-based protocol failed to improve the starting model of T0912 according to GDT-HA score despite improving the predicted binding site (BDTmax of 0.0241, MCCmax of -0.0047 and GDT-HAmax of 0.326 versus BDTstarting of 0.0122, MCCstarting of -0.0058 and GDT-

HAstarting of 0.3285) (Table 4.2, Figure 4.5, and Figure 4.6). In light of this comparison, the binding-site focused MD-based protocol showed promising performance in terms of increasing the accuracy of the focused region during the MD simulation in spite of the low initial accuracy of the binding site predictions.

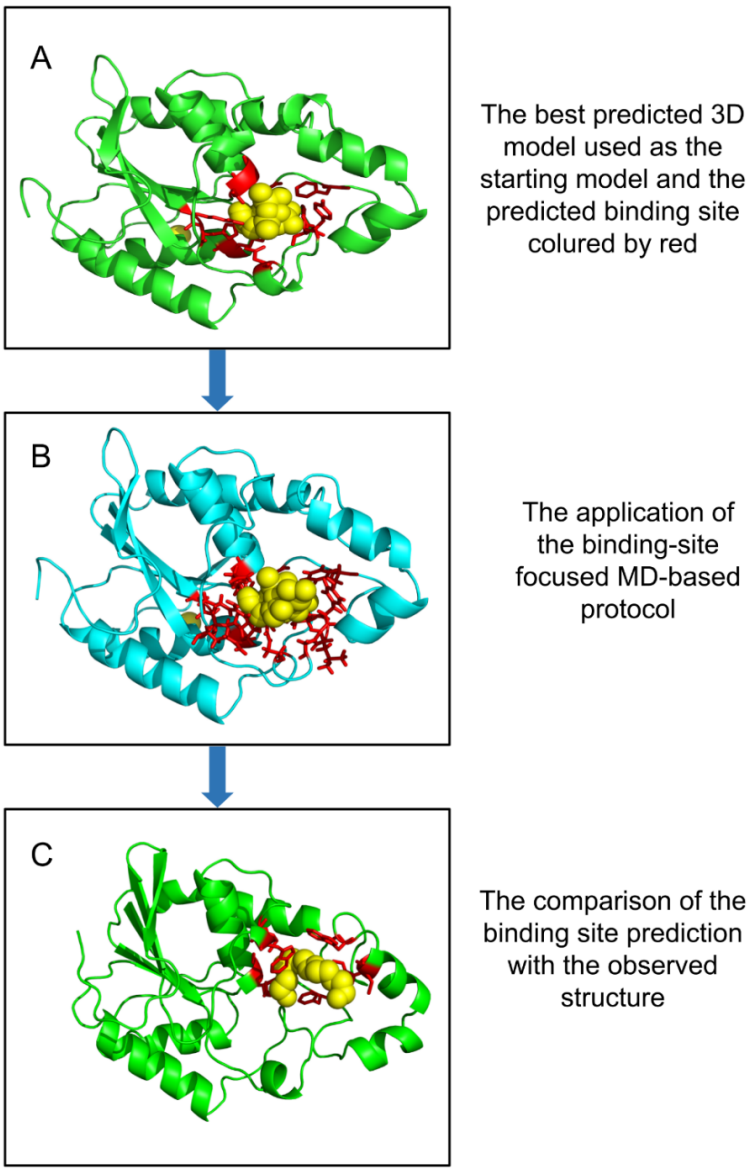

Figure 4. 2 The refinement of a CASP13 target T1016 by the binding site-focused MD-based protocol.

(A) The best-predicted initial server model (green) with the binding site predicted by FunFOLD3 (red sticks) and predicted ligand (yellow spheres). (B) The best-refined model (cyan) with the new predicted binding site (red sticks) and predicted ligand (yellow spheres). (C) The observed structure (green), the observed binding site (red sticks) and observed ligand (yellow spheres). The initial structure versus the best model shows a GDT_HA improvement from 0.6015 to 0.6176, a BDT improvement from 0.227 to 0.235, and an MCC improvement from 0.522 to 0.661.

The binding site predictions for T0911, T0953s2, T1009, T1016, and T1018 have higher BDT and MCC scores compared to other targets (Table 4.2, and Appendix 33-34,37). These targets were defined as TBM targets by the CASP assessors except for T0953s2, which was designated FM/TBM, so it is likely that these targets had better information from more available templates. The initial binding site prediction for T1018 is the best one among the targets studied, and the targeted refinement further improved the quality according to the BDT, MCC scores and GDT-HA scores (BDTmax of 0.528, MCCmax of 0.661 and GDT-HAmax of 0.5989 versus BDTstarting of 0.487, MCCstarting of 0.522 and GDT-HAstarting of 0.5666) (Table 4.2, and Figure 4.7-4.8). Although the accuracy of the predicted binding site for T1009 was increased according to the BDT and MCC scores (BDTmax of 0.489, MCCmax of 0.372 and versus BDTstarting of 0.462, MCCstarting of 0.337), the 3D models generated by the MD-based protocol were not improved according to the GDT-HA score (GDT-HAmax of 0.4359 versus GDT-HAstarting of 0.4352) (Table 4.2, Figure 4.9, and Figure 4.10).

Although the 3D models generated by the binding site-focused MD-based protocol have a slightly lower cumulative *mean* BDT and MCC scores compared to the starting models (∑BDTmean of 1.5774, and ∑MCCmean of 1.35033 versus ∑BDTstarting of 1.7064, and ∑MCCstarting of 1.4614) (Table 4.2), the cumulative *maximum* BDT and MCC scores are higher when the refined models are used to generate the FunFOLD3 predictions compared with using the initial models (∑BDTmax of 1.9838, and ∑MCCmax of 1.96005) (Table 4.2). It is evident that the binding site-focused refinement protocol increased accuracy of the predicted binding sites for all targets, according to the *maximum* BDT and MCC scores (Table 4.1, Table 4.2, and Appendix 33-37).

The GDT-HA score measurement was also used to evaluate the effect of the binding site-focused MD-based protocol on the overall or global quality of the resulting models. Despite using a focused refinement on the predicted binding site during the MD simulations, the strategy also had the beneficial effect of increasing the maximum global accuracy for most targets except for T0912 and T1009. The 3D models generated by the binding site-focused MD-based protocol have higher GDT-HA maximum scores compared to the starting model, despite lower cumulative mean and minimum scores (∑GDT-HAmax of 3.5015, ∑GDT-HAmean of 3.3285, and ∑GDT-HAmin of

3.2088 versus ∑GDT-HA score of starting model 3.4298) (Table 4.2, Figure 4.3-4.10 and Appendix 33-37).

CASP ID	PDB ID	CASP category	The observed binding residues	The initial predicted binding residues	The best-predicted binding residues
T0909	5g5n	FM/TBM	159	146, 147, 169, 170, 208	146, 147, 170, 208
T0911	6e9n	TBM	47,79,137,161,264,271,272,370,371,372,373,374,397,272, 370, 371, 372, 373, 374, 397,272, 370, 371, 372, 373, 374, 397	44, 160, 164, 165, 168, 271, 272, 301, 366, 393	44, 47, 164, 165, 168, 271, 272, 301, 366
T0912	5mqp	FM	155,262,264,268	521, 522, 553	521,553
T0953s2	6f45	FM/TBM	164	120, 121, 122, 123, 124, 125, 126, 156, 157, 158, 159, 164, 165, 166, 167, 170, 174, 198	119, 120, 121, 122, 123, 124, 125, 126, 156, 157, 158, 164, 165, 166, 198
T0954	6cvz	TBM-hard	123, 124, 129, 130, 131	77, 119, 231, 273, 274, 275	77, 274
T1009	6dru	TBM-hard	527,534,286,357,396,404,406,409,257,520 ,559]	173, 257, 286, 325, 393, 395, 396, 470, 484, 487, 520, 557	257, 286, 325, 393, 395, 396, 470, 484, 487, 520
T1011	6m9t	TBM-hard	64,67,115,116,119,123,146,149,150,215 ,216,218,455,489,492,493,496,499	11, 48, 146, 220, 221, 223, 462	48, 146, 220, 221, 222, 223, 455, 482, 492
T1016	6e4b	TBM-easy	81,82,84,106,107,110,151,168,169,132, 133,135,136,139,186	7,8,14,19,20,21,57,81,84,149,150	8, 19, 20, 21, 57, 81, 84, 149, 150
T1018	6n91	TBM-easy	12,14,197,278	14, 56, 59, 98, 170, 197, 278, 279	12, 14, 59, 60, 98, 170, 197, 221, 278

Table 4. 1 Predicted, observed and the best-refined binding residues for the CASP12 and CASP13 targets. The best-refined binding residues is given according to MCC score.

CASP TARGET			BDT score				MCC score				GDT-HA score			
CASP ID	PDB ID	CASP category	starting model	minimum	Mean Score	Maximum Score	starting model	minimum	Mean Score	Maximum Score	starting model	minimum	Mean Score	Maximum Score
T0909	5g5n	FM/TBM	0.0253	0.0212	0.026	0.03	-0.0067	-0.008	-0.00655	-0.00605	0.2703	0.2673	0.278317	0.2913
T0911	6e9n	TBM	0.258	0.128	0.191	0.2938	0.2049	0.0589	0.127	0.3027	0.2972	0.2763	0.289635	0.3009
T0912	5mqp	FM/TBM	0.0122	0.0078	0.0157	0.0241	-0.0058	-0.0106	-0.0066	-0.0047	0.3285	0.298	0.298	0.326
T0953s2	6f45	FM/TBM	0.1345	0.099	0.125	0.168	0.227	-0.016	0.21668	0.25	0.1321	0.127	0.134762	0.1452
T0954	6cvz	TBM-hard	0.0255	0.01	0.0217	0.0279	-0.016	-0.016	-0.0136	-0.0093	0.4379	0.4247	0.435821	0.4532
T1009	6dru	TBM-hard	0.462	0.388	0.456	0.489	0.337	0.238	0.331	0.372	0.4359	0.4008	0.413282	0.4352
T1011	6m9t	TBM-hard	0.0749	0.022	0.102	0.188	0.101	-0.019	0.127	0.2726	0.3598	0.33	0.342918	0.3632
T1016	6e4b	TBM-easy	0.227	0.209	0.226	0.235	0.098	0.079	0.0994	0.1218	0.6015	0.5705	0.598789	0.6176
T1018	6n91	TBM-easy	0.487	0.307	0.414	0.528	0.522	0.303	0.476	0.661	0.5666	0.5142	0.537011	0.5689
The Cumulative score			1.7064	1.192	1.5774	1.9838	1.4614	0.6093	1.35033	1.96005	3.4298	3.2088	3.328535	3.5015

Table 4. 2 The performance of the binding site-focused MD-based protocol according to the BDT, MCC and GDT-HA scores (higher scores better).

The Score	Maximum vs Starting
BDT score	**0.0009766**
MCC score	**0.000976**
GDT-HA Score	**0.006836**

Table 4. 3 Calculated pairwise p-values for the maximum score versus the score of starting models on the CASP13 refinement targets according to the BDT, MCC and GDT-HA scores.

H_o: The maximum score is equal or lower in quality than the score of starting models. H_1: The maximum score is higher quality models than the score of starting models. P-values ≤ 0.05 indicate significant statistical differences (in boldface, higher scores are better).

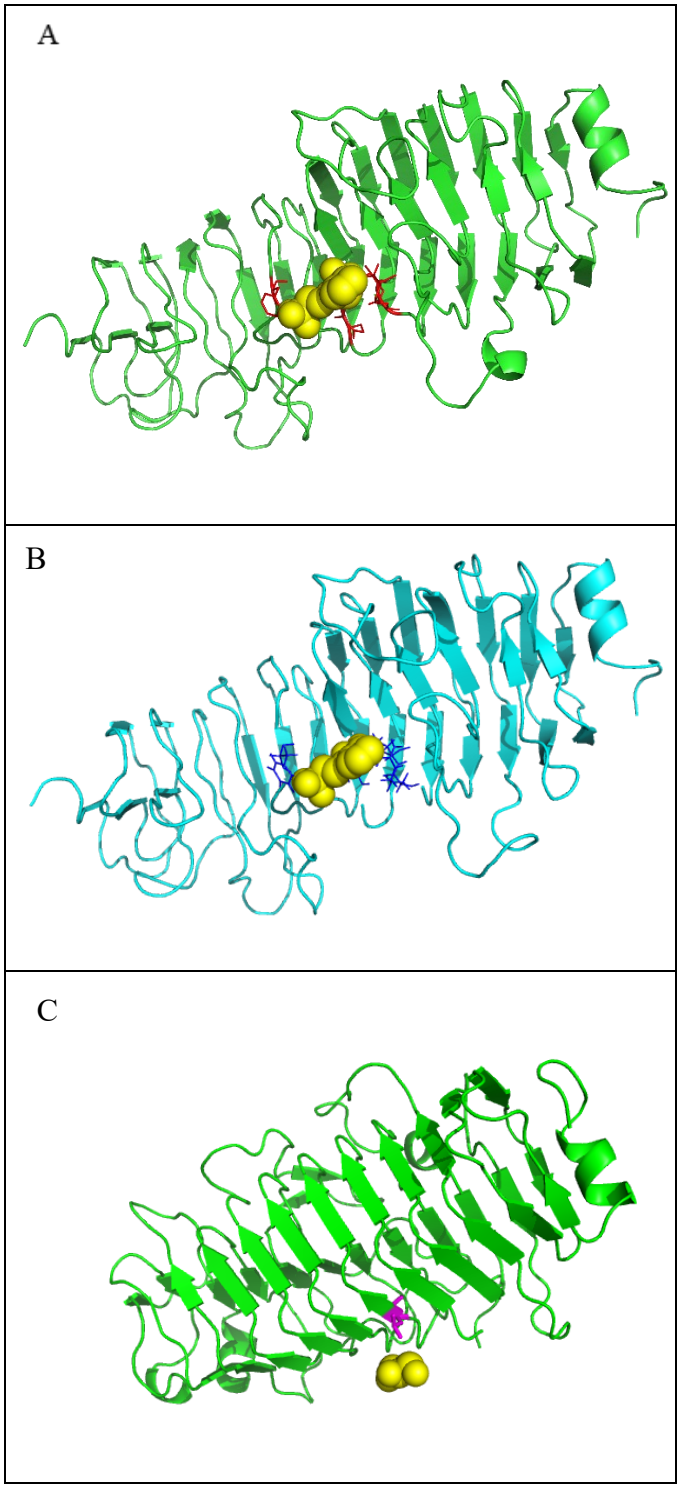

Figure 4. 3 The refinement of a CASP13 target T0909 by the binding site-focused MD-based protocol.

(A) The best-predicted initial server model (green) with the binding site predicted by FunFOLD3 (red sticks) and predicted ligand (yellow spheres). (B) The best-refined model (cyan) with the new predicted binding site (blue sticks) and predicted ligand (yellow spheres). (C) The observed structure (green), the observed binding site (red sticks) and observed ligand (yellow spheres). The initial structure versus the best model shows a GDT_HA improvement from 0.2703 to 0.2913, a BDT improvement from 0.0253 to 0.03, and an MCC improvement from -0.0067 to -0.00605.

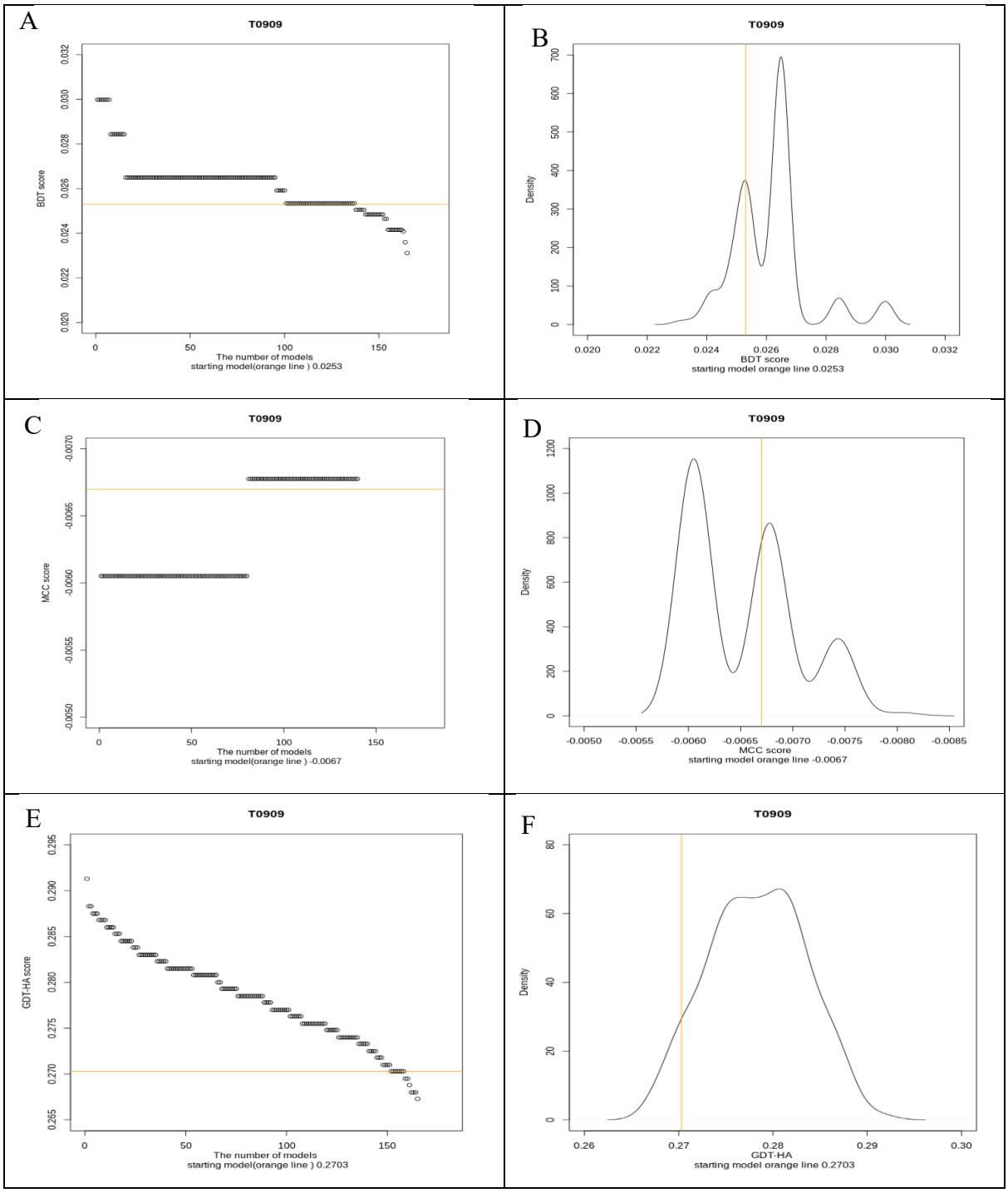

Figure 4. 4 The performance of the binding site-focused MD-based protocol for T0909 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

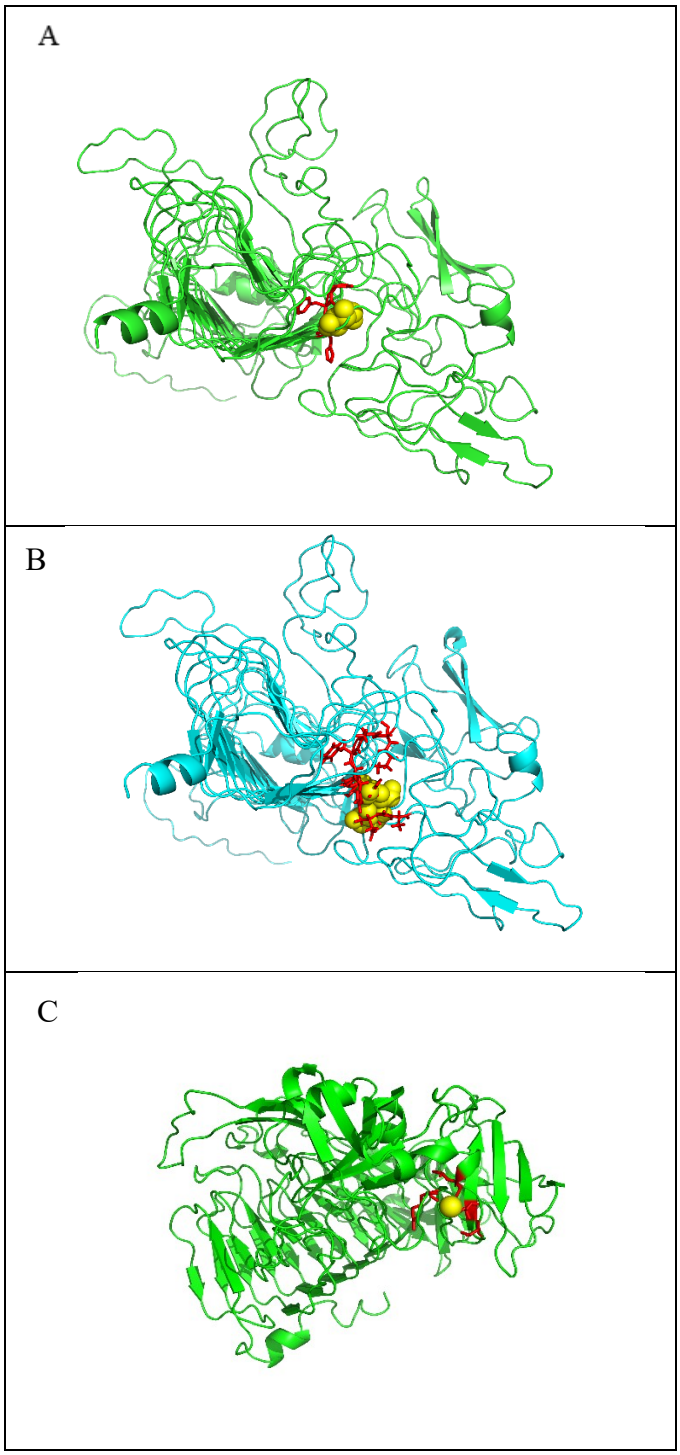

Figure 4. 5 The refinement of a CASP13 target T0912 by the binding site-focused MD-based protocol.

(A)The best-predicted initial server model (green) with the binding site predicted by FunFOLD3 (red sticks) and predicted ligand (yellow spheres). (B) The best-refined model (cyan) with the new predicted binding site (red sticks) and predicted ligand (yellow spheres). (C) The observed structure (green), the observed binding site (red sticks) and observed ligand (yellow spheres). The initial structure versus the best model shows a BDT improvement from 0.0122 to 0.0241, and an MCC improvement from -0.0058 to -0.0047.

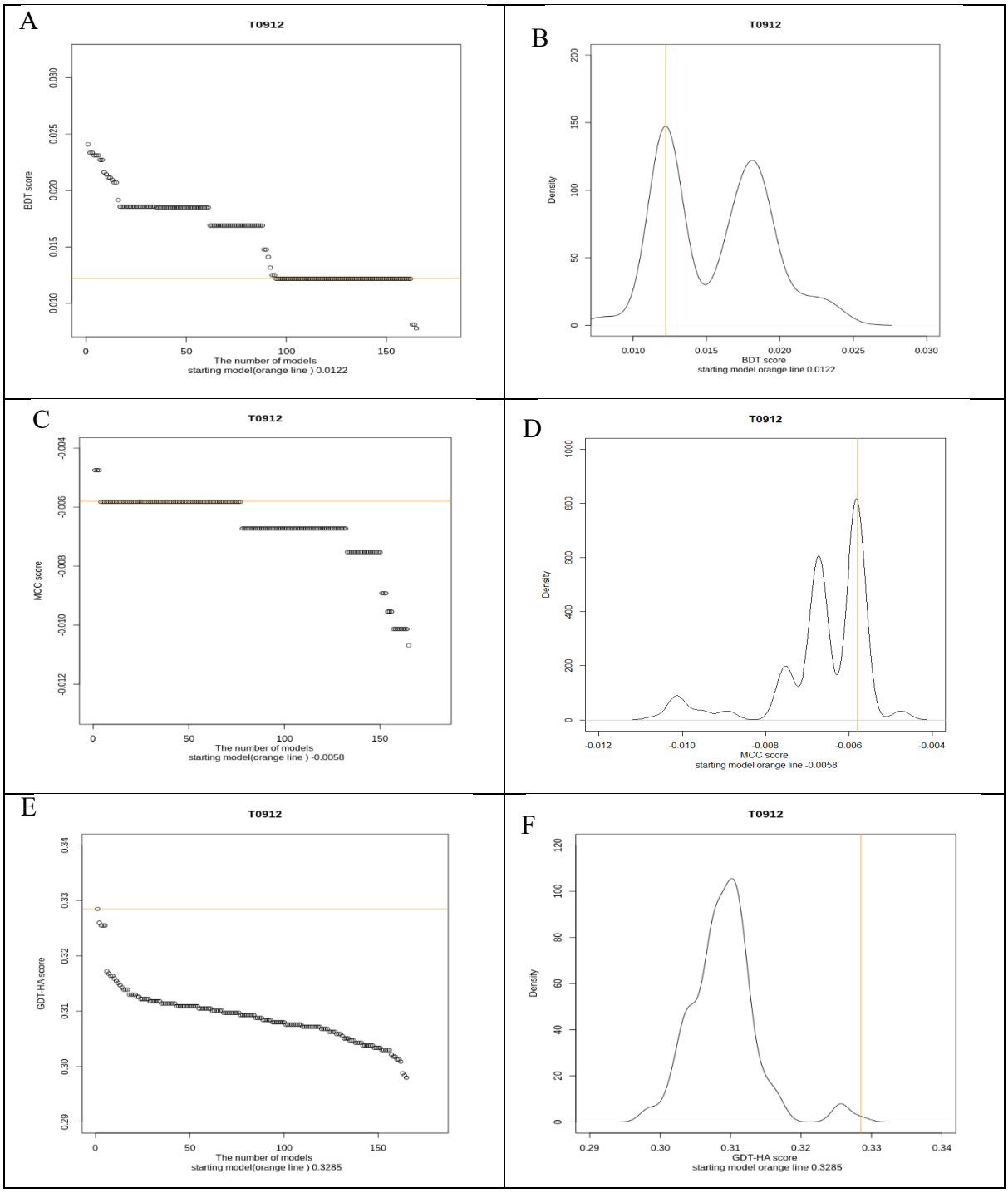

Figure 4. 6 The performance of the binding site-focused MD-based protocol for T0912 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

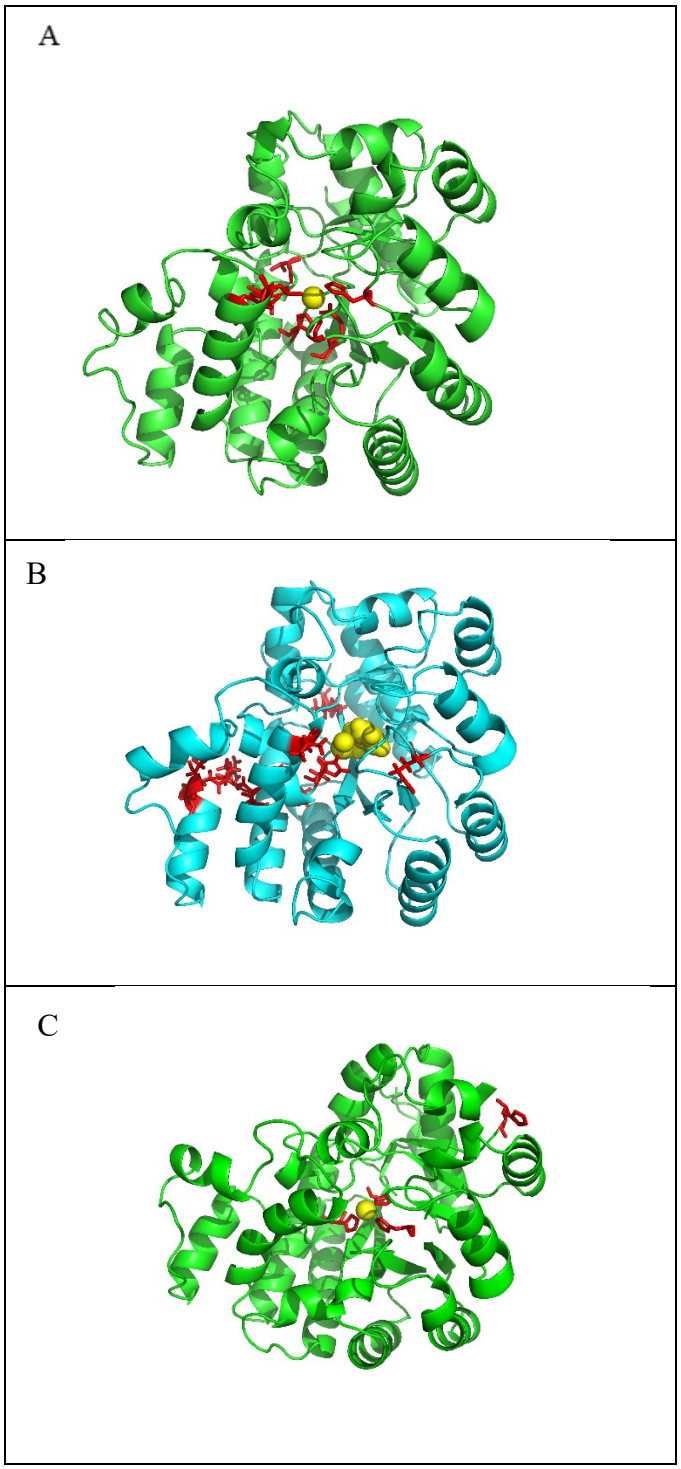

Figure 4. 7 The refinement of a CASP13 target T1018 by the binding site-focused MD-based protocol.

(A) The best-predicted initial server model (green) with the binding site predicted by FunFOLD3 (red sticks) and predicted ligand (yellow spheres). (B) The best-refined model (cyan) with the new predicted binding site (blue sticks) and predicted ligand (yellow spheres). (C) The observed structure (green), the observed binding site (red sticks) and observed ligand (yellow spheres). The initial structure versus the best model shows a GDT_HA improvement from 0.5666 to 0.5689, a BDT improvement from 0.487 to 0.528, and an MCC improvement from 0.522 to 0.661.

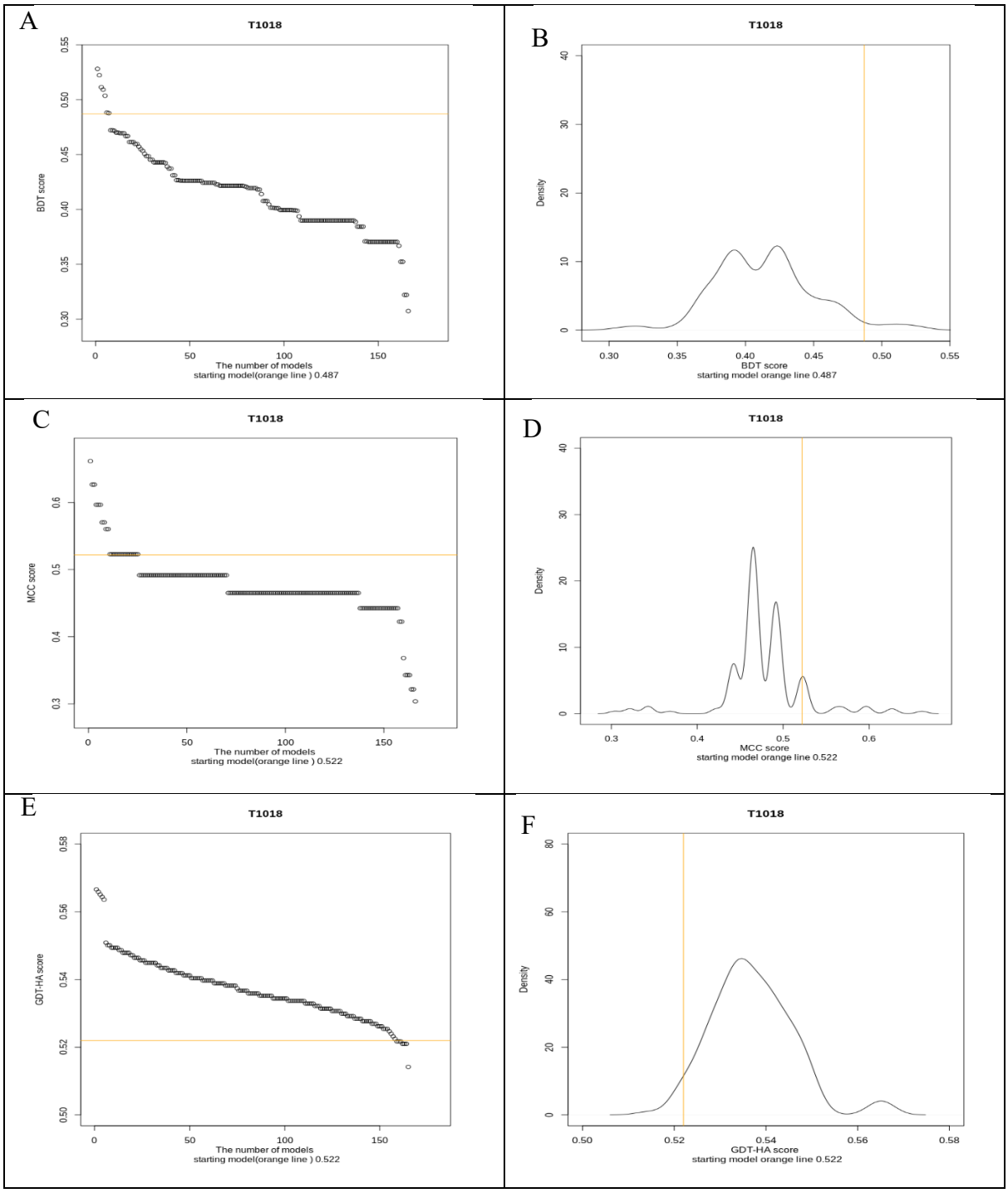

Figure 4. 8 The performance of the binding site-focused MD-based protocol for T1018 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

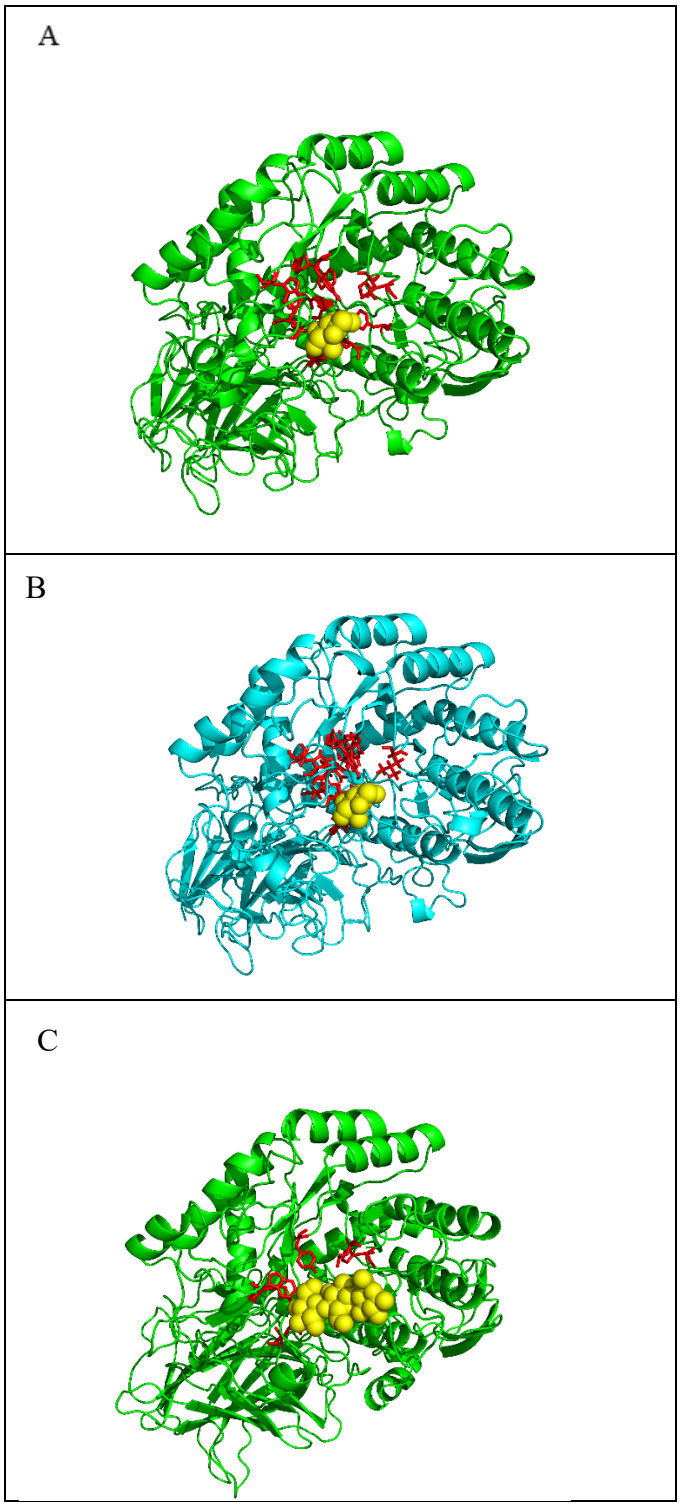

Figure 4. 9 The refinement of a CASP13 target T1009 by the binding site-focused MD-based protocol.

(A) The best-predicted initial server model (green) with the binding site predicted by FunFOLD3 (red sticks) and predicted ligand (yellow spheres). (B) The best-refined model (cyan) with the new predicted binding site (blue sticks) and predicted ligand (yellow spheres). (C) The observed structure (green), the observed binding site (red sticks) and observed ligand (yellow spheres). The initial structure versus the best model a BDT improvement from 0.462 to 0.489, and an MCC improvement from 0.337 to 0.372.

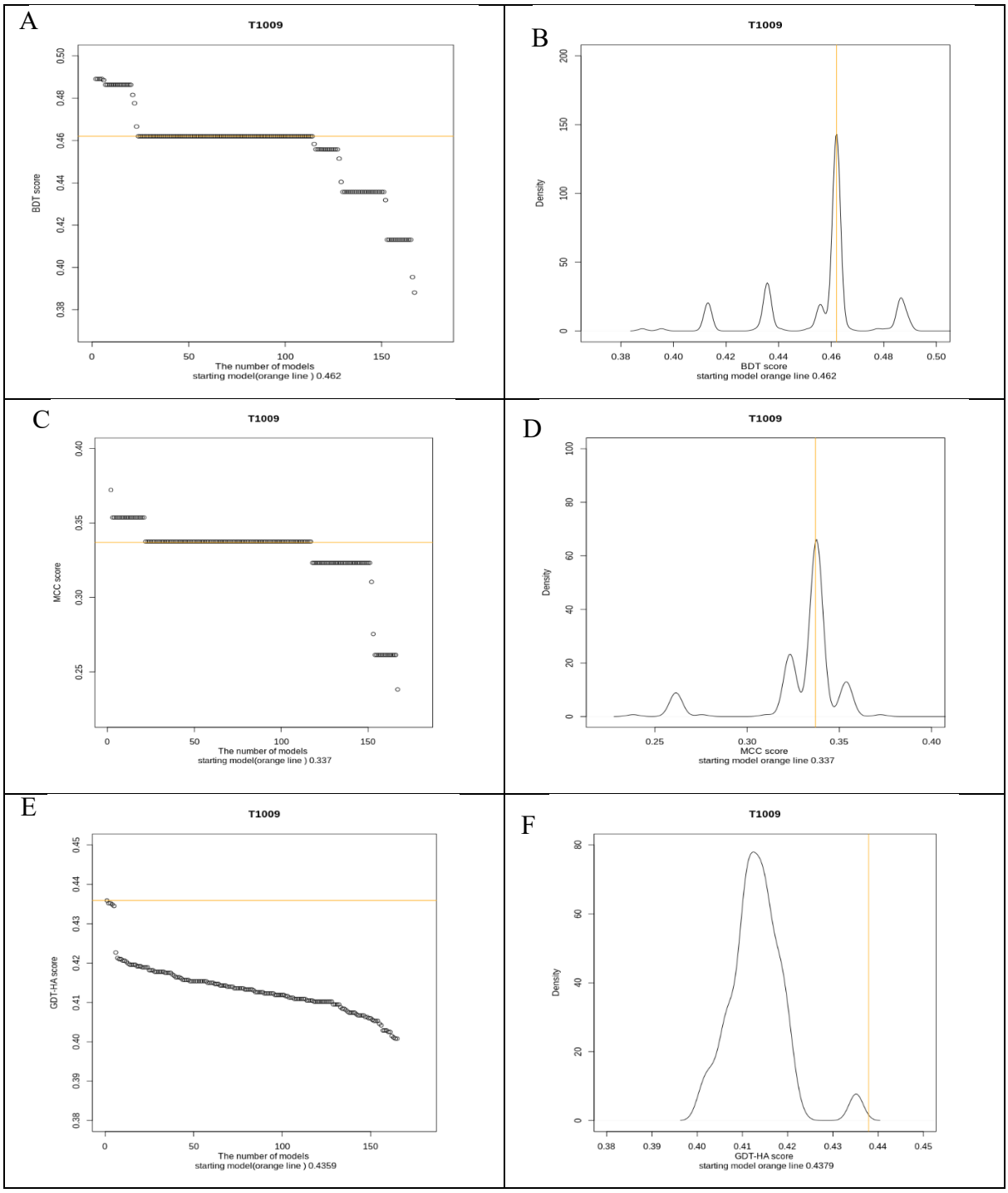

Figure 4. 10 The performance of the binding site-focused MD-based protocol for T1009 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

4.5 Conclusions

The prediction of the protein-ligand binding sites is crucial in terms of understanding the function of the protein structures, particularly in cases where native structures are unavailable. The interactions between proteins and ligands may further our knowledge of the mechanisms of action and their effects. Determining the atomic details of protein-ligand interactions using experimental methods may not always be feasible to bridge the sequence-structure-function knowledge gap. However, *in silico* methods are an alternative way to elucidate the protein functions and model protein-ligand interactions starting with amino acid sequence information.

FunFOLD3 was developed by the McGuffin group to predict protein-ligand binding sites from sequences utilising a template-based modelling approach. The BDT and MCC scores were also developed for the evaluation of the binding site prediction performance based on the comparisons of predicted and observed binding residues. In this chapter, our aim was to increase the accuracy of the binding site predictions made by FunFOLD3 by utilising a binding site-focused MD-based refinement protocol. The performance of this protocol was analysed in terms of: 1. its ability to help improve binding site residue predictions through improvement of the input model binding site (evaluated using the BDT, MCC scores), and, 2. its ability to also improve the global quality of the initial model (evaluated using the GDT-HA scores).

The MCC and BDT scores were used to analyse specifically the improvement of the binding site prediction. The calculation of the MCC score is based on a statistical comparison of the observed and predicted binding residues, and the BDT score is also calculated considering the distance between the predicted and observed binding residues in the native structure. Therefore, GDT-HA score, which relies on the superposition of the entire predicted 3D model with the observed structure, was also used to analyse the improvements to the global quality. It is also worthy of note that although the MD-based protocol was used to refine the predicted binding site, it performed well in terms of improving the quality of the 3D models according to GDT-HA score.

It is promising that the binding site-focused refinement strategy managed to improve all predicted binding sites, whether they were initially highly accurate or less-accurate according to the BDT and MCC scores. It was also observed that the target difficulty is an important factor governing the accuracy of the initial model and predicted binding site. The binding site predictions for TBM targets are much more accurate compared to the FM/TBM targets, with the exception of T0954 and T1011. This is likely to be related to the availability of appropriate template structures with bound ligands. Therefore, it can be said that the binding site predictions made by FunFOLD3 for the TBM targets are likely to be more reliable and provide more helpful information in terms of functionality and the role of the interactions. The binding site-focused MD-based strategy may boost the performance of FunFOLD3 by further improving the quality of the input 3D models used to make predictions.

Although the results presented here are promising, for a more thorough analysis of the performance of the MD-based protocol, it will be tested with further targets when they are available. Despite scanning all CASP12, CASP13 targets, only 9 targets with publicly released structures were found to contain biologically relevant ligands. The FunFOLD3 predictions will continue be made by the McGuffin group for all of the CASP14 targets, and these can be added into the analyses to further test the approach in future. Nevertheless, the analysis shows that the binding-site focused MD-based protocol has potential to increase the accuracy of the predicted binding sites towards the native structure. The approach will be considered for integration with future versions of FunFOLD and the IntFOLD server.

Residue-residue contact prediction methods have shown considerable potential in the contact prediction categories of recent CASP experiments. In the next chapter, as the first attempt, this prior knowledge is utilised to improve the quality of the initial structures in the refinement pipeline.

Chapter 5 The Utilisation of Residue-Residue Contact Predictions to Provide Guidance to the MD-Based Refinement Protocol

5.1 Background

The development of advanced deep learning methods along with the increase in the sizes of sequence families have dramatically increased the potential of the residue-residue contact prediction methods in recent years. The predicted contacts between amino acids have been an important part of *in silico* protein modelling, and the accuracy of the residue-residue contacts plays a critical role in the quality of 3D models and function predictions (Gromiha & Selvaraj, 2004). The idea of building 3D models by utilising predicted residue-residue contact maps emerged 20 years ago (Adhikari et al., 2015; Adhikari & Cheng, 2016; Mirny & Domany, 1996; Vendruscolo & Domany, 2000). Nevertheless, it is only relatively recently that advancements in residue-residue contact predictions, and have reached a reasonable enough accuracy in order to become a mainstay in modelling pipelines (Adhikari & Cheng, 2016; Jones, 2001; Rohl et al., 2004; Vitkup et al., 2001). Residue-residue contact predictions have been primarily used to build 3D models, (Jones 2001; Marks et al. 2011), followed by drug design, (Kliger et al., 2009), and model quality estimation (Miller & Eisenberg, 2008; Z. Wang et al., 2011). Predicting contacts for the construction of *ab initio* protein 3D models for FM targets is still unsolved, but considerable progress has been made in recent years to increase the accuracy of methods (Adhikari & Cheng, 2016). Since CASP13, deep learning methods for prediction of residue distances have arguably become the major step forward in improving the quality of 3D models predicted for FM targets.

Contact prediction methods can be categorised according to the usage of the information in the prediction process. Methods can be (1) coevolution-derived, (2) machine learning, (3) template-based, (4) physicochemical-based, and (5) hybrid methods (Adhikari & Cheng, 2016; Schneider & Brock, 2014; Yachdav et al., 2014). The methods can also be more broadly classified as correlated mutation-based or machine learning-based methods (Adhikari & Cheng, 2016; Björkholm et al., 2009; Di Lena et al., 2012; Schneider & Brock, 2014). The top-performing machine learning-based approaches are based on the usage of the deep learning architectures, and some of the methods also utilise correlated mutation information (Björkholm et al., 2009; Cheng et al., 2009; Cheng & Baldi, 2007; Di Lena et al., 2012; Eickholt & Cheng, 2013; Fariselli et al., 2001; Michel et al., 2014; Shackelford & Karplus, 2007; Skwark et al., 2014; Vullo et al., 2006; Z. Wang et al., 2009; Wu & Zhang, 2008). Many different kinds of information including residue type, secondary structure, and the sequence profiles are used by

the machine learning-based approaches in order to predict the residue-residue contacts (Adhikari & Cheng, 2016; Bacardit et al., 2012; Björkholm et al., 2009; Cheng & Baldi, 2007; Fariselli et al., 2001; Li et al., 2011; Vullo et al., 2006; Z. Wang et al., 2009; Wu & Zhang, 2008). Coevolution-derived methods focus on the mutations, which are linked with other residue mutations, and so contact predictions that are made by coevolution-derived methods usually make use of multiple sequence alignments (MSAs) for identifying correlated mutations (Adhikari & Cheng, 2016; Buslje et al., 2009; Ekeberg et al., 2013; Göbel et al., 1994; Jeong & Kim, 2012; Jones et al., 2015; Kamisetty et al., 2013; Lapedes et al., 1999; Olmea & Valencia, 1997; Schneider & Brock, 2014; Shindyalov et al., 1994; Tetchner et al., 2014; Weigt et al., 2009).

MSA, also known as `sequence profile` is aimed to collect and align multiple protein sequence of the target sequence. MSAs contains sequences provide a substantial data about motifs, conserved positions to modelling of protein structures and function prediction (Zhang et al., 2020). For the protein modelling, MSA is initially used for the secondary structure prediction (Jones, 1999; Wu & Zhang, 2008), then residue-residue contact prediction (Adhikari et al., 2018; Hanson et al., 2018; He et al., 2017; S. Wang et al., 2017), template-based modelling (Söding, 2005; Wu & Zhang, 2008; Zhang et al., 2020; Zheng et al., 2019), and ligand binding site predictions (Gil & Fiser, 2019; Yu et al., 2013). Therefore, developing the MSA tools plays a critical role for the bioinformatics tools. Contact prediction methods use different MSA tools for residue covariation besides the whole chain (De Juan et al., 2013). While PSI-BLAST has been widely used for the sequence profile generation (Altschul et al., 1997), different combinations of HHblits (Remmert et al., 2012), HH-suite (Steinegger & Söding, 2018), Jackhmmer and HMMER suite (Eddy, 1998) are among the popular MSA tools for the residue-residue contact prediction methods (D. W. A. A. Buchan & Jones, 2018; Di Lena et al., 2012; Greener et al., 2019; Kandathil et al., 2019a; Ovchinnikov et al., 2017; Schaarschmidt et al., 2018; Y. Wang et al., 2019; Wu et al., 2011)

Residue-residue contacts are defined as pairs of close residues, usually within 8 Å of each other within the 3D structure. The residue distance is measured between carbon-betas (carbon-alphas in case of glycine) using the x, y, and z coordinates of the atoms in the PDB structure files (Adhikari & Cheng, 2016; Duarte et al., 2010). The threshold based on the distance also determines the number of residues which are in contact in a 3D model. The determination of

which residues are in contact is a key to reconstructing the protein fold (Adhikari & Cheng, 2016; Niggemann & Steipe, 2000). The residue contacts are divided into three main categories - short-range, medium-range, and long-range - depending on the distance between residues in 3D models (Adhikari & Cheng, 2016; Schaarschmidt et al., 2018). The accuracy of the prediction of long-range contacts is crucial for the reconstruction of 3D models, but it is also quite hard to predict them with high accuracy compared to others. Therefore, the long-range contacts are often considered as a separate category in the assessment of contact prediction methods (Adhikari & Cheng, 2016; Eickholt & Cheng, 2013; Kryshtafovych et al., 2011; Monastyrskyy et al., 2014). The distribution and coverage of contacts is also a critical factor in building 3D models, for methods that rely on predicted contacts (Adhikari & Cheng, 2016; Eickholt & Cheng, 2013; Kryshtafovych et al., 2011; Monastyrskyy et al., 2014). For instance, if most of the predicted contacts with high accuracy are only gathered in a particular region, then the prediction of highly accurate structures may require additional information about the rest of the structures. As a result, the proportion of the contacts is also traditionally evaluated with the top L/2 or just the top L/5 predicted contacts (where L is the sequence length) (Adhikari & Cheng, 2016; Eickholt & Cheng, 2013; Kryshtafovych et al., 2011; Monastyrskyy et al., 2014).

5.1.1 The Residue-Residue (RR) Contact Predictions Category in CASP

The contact prediction category was first introduced in CASP2 to assess the ability of methods to predict residue-residue contacts (Aloy et al., 2003; Bohr et al., 1993; Ezkurdia et al., 2009; Graña et al., 2005; Kryshtafovych et al., 2016; Lesk, 1997; Lesk et al., 2001; Orengo et al., 1999; Rangwala & Karypis, 2010; Skolnick et al., 1997; Taylor et al., 2014). It was proposed that the long-range contacts can be utilised for *ab initio* modelling by constraining the contacts in 3D models (Schaarschmidt et al., 2018; Skolnick et al., 1997; Taylor et al., 2014). Therefore, the prediction of 3D models which were assisted by contacts has been included as a separate evaluation category since CASP10 (Schaarschmidt et al., 2018; Skolnick et al., 1997; Taylor et al., 2014). Nevertheless, the low accuracy of residue-residue contact predictions had limited the success of contact-assisted structure prediction approaches (Schaarschmidt et al., 2018). CASP also gives more importance to the accuracy of the long-range contacts compared to short-range and medium-range contacts, as the prediction of the short-range and medium-range

contacts is not as hard as long-range contacts. The top L/5 contacts might be highly accurate, but the coverage of the predicted contacts may be too low for the accurate reconstruction of the protein fold (Adhikari et al., 2015; Adhikari & Cheng, 2016; Eickholt & Cheng, 2013; Marks et al., 2011; Michel et al., 2014).

The contact prediction methods which made use of machine learning and coevolution-derived approaches have shown better performance in the recent CASP blind assessments. FM modelling targets are usually used for the evaluation of the contact prediction methods as such methods are of most use when no templates are available to predict the protein 3D models (Adhikari & Cheng, 2016; Kryshtafovych et al., 2016; Schaarschmidt et al., 2018; Shrestha et al., 2019).

Although residue-residue contacts have been assessed since CASP2, significant progress was not observed until CASP11 (Schaarschmidt et al., 2018). The accuracy of the contact prediction fluctuated at around 20% of precision until CASP11 (Schaarschmidt et al., 2018; Shrestha et al., 2019). An average precision of 27% was reached in CASP11, and this improvement was roughly doubled by an average precision of 47% in CASP12 as a remarkable milestone of the contact prediction category with the application of deep neural network (DNN) (Jones et al., 2015; Kryshtafovych et al., 2016; Schaarschmidt et al., 2018). The application of various DNN methods have been the mainstay of the bioinformatics tools in recent CASP experiments particularly Convolutional Neural Network (CNN) in CASP13. DNN approaches has been successful where no homologs can be found (Greener et al., 2019; Kandathil et al., 2019b). While the traditional neural networks (NN) have usually consisted of a single hidden layer, CNN methods are able to make the use of multiple layers up to 20 layers for the contact predictions (Greener et al., 2019; Kandathil et al., 2019b).

The MetaPSICOV developed by the David Jones group from UCL was one of the methods that showed a noticeable performance improvement for the prediction of residue-residue contacts compared to previous methods (Jones et al., 2015; Kryshtafovych et al., 2016; Schaarschmidt et al., 2018). Although this major leap was achieved by numerous groups and there was not a huge gap in terms of precision performance between the top prediction groups in CASP12 (Jones et al., 2015; Kryshtafovych et al., 2016; Schaarschmidt et al., 2018).

CASP13 saw a further major leap in performance in the contact prediction category, with the top-performing groups reached up to 70% precision using deep neural network algorithms (Shrestha et al., 2019). After the 20 years of contact prediction in CASP experiments, the prediction groups have reached a sufficient accuracy residue-residue contacts for further *in silico* studies (Schaarschmidt et al., 2018; Shrestha et al., 2019). Many of the top-performing Free Modelling groups in CASP13 were also performing well in contact prediction (Schaarschmidt et al., 2018; Shrestha et al., 2019).

5.1.2 DeepMetaPSICOV

MetaPSICOV was developed with the aim of increasing the accuracy of the predicted contacts derived from multiple sequence alignments by using a combination of different machine learning algorithms as a meta-predictor (Jones et al., 2015; Kosciolek & Jones, 2016). The method also combines the direct-co-evolution with statistical approaches and the evaluation of "classic" protein properties using two tandem neural networks (Jones et al., 2015; Kosciolek & Jones, 2016). The first stage is based on the generation of the initial contact maps utilising 672 features in total to predict the likelihood of residue i and j being in contact using three windows (Jones et al., 2015).. In the second stage, the initial contact map was analysed to eliminate outliers and fill the gaps taking advantage of 731 features to correlate the output from the first stage using two windows of 11 alignment columns (Jones et al., 2015). MetaPSICOV utilised also combination of three different coevolution methods including: PSICOV (Jones et al., 2012), mfDCA (Kaján et al., 2014), and GREMLIN (Jones et al., 2015; Kamisetty et al., 2013; Kosciolek & Jones, 2016; Seemayer et al., 2014)

MetaPSICOV (the server was registered as CONSIP2 in CASP11) reached an average precision of 27% as the best contact prediction method in CASP11 (Jones et al., 2015; Kosciolek & Jones, 2016).

MetaPSICOV2 was subsequently upgraded using a wider neural network and input window of 15 residues instead of 9-residue window originally used (D. W. A. A. Buchan & Jones, 2018). This version of MetaPSICOV2 was tested in CASP12 and achieved an average precision of 43% (D. W. A. A. Buchan & Jones, 2018). The next milestone was the development of fully

convolutional neural networks (FCN), which were successfully applied to predict the whole contact maps using the architecture of DeepCov instead of generating initial contact maps with MetaPSICOV (Jones & Kandathil, 2018; Kandathil et al., 2019a).

In CASP13 we saw that the application of deep neural networks had boosted the performance of contact prediction methods (Kandathil et al., 2019a). The DeepMetaPSICOV (DMP) (Kandathil et al., 2019a) method was a further development, which used a deep, fully convolutional residual neural network. DMP includes the combination of input features used in MetaPSICOV and DeepCov (Jones & Kandathil, 2018) as well as the outputs from PSICOV (Jones et al., 2012), CCMpred (Seemayer et al., 2014) and FreeContact (Kaján et al., 2014). These features are converted into 2D maps (Kandathil et al., 2019a) and DMP has the capacity of predicting residue-residue contacts for a wide range of proteins such as membrane proteins, even if the initial MSA are relatively shallow (Kandathil et al., 2019a). DMP was among top 5 contact prediction methods and it reached an average precision of ~60% in CASP13 (Jones & Kandathil, 2018; Kandathil et al., 2019a; Shrestha et al., 2019)

The MetaPSICOV method was previously used for the calculation of the Contact Distance Agreement (CDA) score, which was a novel component of the ModFOLD6 global and local quality scoring method (Jones et al., 2015; Maghrabi & McGuffin, 2017). The CDA score based on the agreement between the residue contacts predicted by MetaPSICOV and the contacts measured according to the Euclidean distance (in Å) between residues in a predicted 3D model, and the generation of the CDA score explained in details in section 5.3.2 (Maghrabi & McGuffin, 2017).

5.2 Aims and Objectives

After the impressive performance of DeepMetaPSICOV in CASP13, it is proposed to utilise the CDA score based on the agreement between the DeepMetaPSICOV predicted contacts and the contacts in the starting model, in order to guide the MD-based refinement protocol (Kandathil et al., 2019a; Shuid et al., 2017). DeepMetaPSICOV method is also one of the top-performing methods available for local installation, so DeepMetaPSICOV was preferred to guide the MD-based protocol. The contact-assisted MD-based protocol may be able to generate 3D models that are closer to the native structure by restraining the highly accurate residue-

residue contacts predicted by DeepMetaPSICOV, during the MD simulation (Kandathil et al., 2019a; Shuid et al., 2017). In other words, the unrestrained refinement of the highly accurate residue-residue contacts may lead to a deterioration in the quality of 3D models, thereby the contacts should be restrained to avoid unnecessary deviations away from the native structure. In Chapters 2 and 3 the fixed local quality assessment score produced by the ModFOLD server was used in order to guide the MD-based refinement protocol. In this chapter, the contact-assisted restraint strategy will be evaluated and compared with our previous fixed restraint strategy based on the local quality estimation (Kandathil et al., 2019a; Shuid et al., 2017).

5.3 Material and Methods

5.3.1 Data Collection

The CASP13 regular targets were used to evaluate the performance of the new contact-assisted MD-based refinement protocol and compare its performance with the original MD-based protocols of ReFOLD and the fixed restraint strategy which is based on the local quality estimation (see Chapter 2) (Kandathil et al., 2019a; Shrestha et al., 2019; Shuid et al., 2017). Contact predictions, which were made by the DMP group (491) using DeepMetaPSICOV (DMP) during CASP13, were obtained from CASP13 download website (http://predictioncenter.org/download_area/CASP13/predictions/) (Kandathil et al., 2019a; Shrestha et al., 2019; Shuid et al., 2017). The TM-score and Molprobity tools were also used to produce GDT-HA (Zhang & Skolnick, 2005) and Molprobity scores (Davis et al., 2004) for the models generated by the three MD-based protocols.

5.3.2 Computational Design

The benchmarking of the contact-assisted MD-based protocol included three main stages: 1) the generation of the CDA score, 2) the application of gradual restraints based on the CDA score during the MD simulation to generate refined 3D models, and 3) the assessment of the refined 3D models using observed model quality scores (Figure 5. 1) (Kandathil et al., 2019a; Shuid et al., 2017).

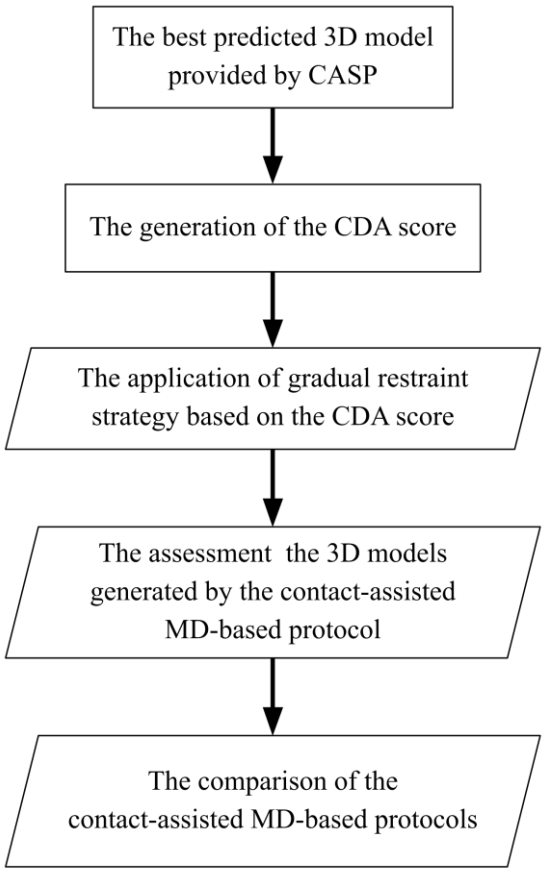

Figure 5. 1 Flowchart showing the workflow for the application of the contact-assisted MD-based protocol.
After obtaining the best predicted initial 3D model, the CDA score was generated, and then a gradual restraint strategy based on the CDA score was applied during the MD simulation. The 3D models generated by the contact-assisted MD-based protocol were scored versus the native structures. The performance of the contact-assisted MD-based protocol was compared with the original MD-based protocols of ReFOLD and the fixed local quality assessment guided MD-based protocol.

The first stage included the generation of the CDA scores using the residue-residue contacts which were made using the DMP method in CASP13. The residue-residue contacts predicted by DMP were used to produce the CDA score, which ranges from 0 to 1, by utilising: the target protein sequence in FASTA format, the best predicted 3D model, and the contact prediction file. The CDA score was originally developed for use in ModFOLD6, but here we used the DeepMetaPSICOV predictions instead of MetaPSICOV, and scored all pairs of residues in each model which were measured to be within 8Å (Kandathil et al., 2019a; Shuid et al., 2017). For instance, if a residue i was in contact with both residue j and k in the predicted 3D model, then the DeepMetaPSICOV scores were obtained for ij and ik contacts (Kandathil et al., 2019a; Shuid et al., 2017). The CDA score was then calculated by taking the mean DeepMetaPSICOV score for the contacts for each residue, using the formula, CDA=(∑ p)/c, where p is the

DeepMetaPSICOV score for each contacting pair and c is the number of contacts for the residue (within 8Å) in the predicted 3D model where the DeepMetaPSICOV score exists (Kandathil et al., 2019a; Maghrabi & McGuffin, 2019, 2017; Shuid et al., 2017).

After the calculation of the CDA score for each residue in the predicted 3D model, a gradual restraint strategy based on the CDA score was applied in the MD simulation stage. It was also postulated that if the CDA score was high, a stronger restraint should be applied to keep the residues in contact in the predicted 3D model (Kandathil et al., 2019a; Mirjalili & Feig, 2013; Shuid et al., 2017). A low CDA score suggests that the residue may be further away from the native structure (because the contacts for that residue in the model do not agree so much with those predicted by DeepMetaPSICOV), and so it should be refined in order to improve the overall quality of the predicted 3D model. Therefore, a gradual restraint strategy, which varied between weak (0.05 kCal/mol/Å^2) and strong (1 kCal/mol/Å^2) harmonic positional restraints on all atoms including C-alphas, was applied by considering the distribution of the CDA score during the MD simulation (Table 5.1 and Figure 5.2) (Kandathil et al., 2019a; Shuid et al., 2017). The range of the force constant was optimized for the application of the gradual restraint based on the local quality estimation as from 0.05 kCal/mol/Å^2 to 1 kCal/mol/Å^2 in Chapter3. Here, we tried to apply different ranges of the CDA score. For instance, strong restraint (1 kCal/mol/Å^2) was applied to the residues with the CDA score from 0.8 to 1.0. The CDA score ranges defined in Table 5.1 were found to be more successful in terms of increasing the population of the improved 3D models and simulation execution. (Adiyaman & McGuffin, 2019; Maghrabi & McGuffin, 2019; Mirjalili & Feig, 2013; Read et al., 2019; Shuid et al., 2017)

The CDA score	The force constant (kcal/mol/Å^2)
0.9-1	1
0.7-0.9	0.5
0.5-0.7	0.1
0.3-0.5	0.05
0-0.3	0

Table 5. 1 The application of the gradual restraint strategy based on the CDA score.

The molecular dynamics simulations were conducted using NAMD 2.10 (Phillips et al., 2005) in GPU mode and the same parameters that were optimised for the original MD protocol for ReFOLD (Chapter 2) were used in order to gauge the effects of the gradual restraint strategy based on the CDA score (Kandathil et al., 2019a; Shuid et al., 2017).

The contact-assisted MD-based protocol was compared with the original MD-based protocols of ReFOLD and the fixed local quality assessment guided MD-based protocol using the CASP13 targets. One-tailed unpaired Wilcoxon tests were performed based on the differences in the observed quality, which was measured using the GDT-HA and Molprobity scores.

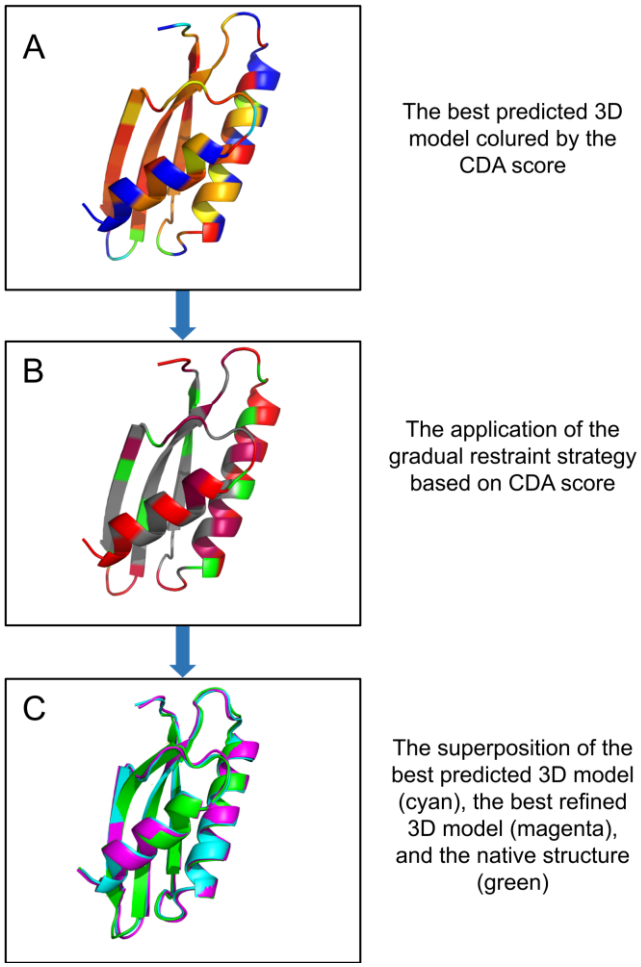

Figure 5. 2 The refinement of a CASP13 target using the contact-assisted MD-based protocol.
The gradual restraint strategy based on the CDA score was applied during the MD simulation. (A) The best predicted model (the CASP13 regular target T1006); is coloured using the CDA scores. (B) The red and pink regions indicate where strong restraints were applied, and the grey and green regions show where weaker restraints were applied during the MD simulation based on the CDA score. (C) Superposition of the initial structure (cyan), the best model generated by the contact-assisted MD-based protocol (magenta), and the native structure (green). The refinement process improved the model with the GDT_HA score increasing from 0.8701 to 0.8831.

5.4 Results and Discussion

Forty-four regular CASP13 targets were used to analyse the performance of the contact-assisted restraint strategy. Of the targets, 20 out of 44 were designated TBM targets, 11 were FM targets and 13 were FM/TBM targets. Subsequently, performance on the same dataset was compared with the fixed local quality assessment guided restraint strategy, which was developed in Chapter 2 and tested in CASP13, and the original MD-based protocol of ReFOLD (Shuid et al., 2017).

In this chapter, state of the art residue-residue contact predictions were utilised in order to guide the MD-based refinement protocol, and the performance of this contact-assisted restraint strategy was analysed (Shuid et al., 2017). Significant progress in our refinement pipeline has been already made by using the local model quality assessment scores to inform restraints in order to guide the MD simulations (Chapter 2 and 3) (Shuid et al., 2017). For the fixed local quality assessment guided restraint strategy, a threshold based on the predicted per-residue accuracy score produced by ModFOLD7 (Maghrabi & McGuffin, 2019) was applied in Chapter 2. Further to this, after witnessing the success of the contact prediction for prediction of 3D models in CASP13, it was postulated that residue-residue contacts from improved methods such as DMP, may also be used to provide guidance to the MD-based protocol to avoid undesired structural deviations (Kandathil et al., 2019a; Shuid et al., 2017). A gradual restraint strategy based on the CDA score was also postulated to show a better performance rather than a fixed threshold, so we also chose to apply gradual restraints in this case, in a similar way to that used for the local model quality estimation as in Chapter 3.

It is also known that contact prediction methods are able to predict highly accurate residue-residue contacts, even in cases where there is low similarity between the target and available structures. In light of such information, it is possible to postulate that using contact prediction methods may boost the performance of the MD-based protocol in particular for FM targets and domains. Therefore, the performance of the three MD-based protocols was analysed according to the prediction methods of initial structures.

For TBM-easy and TBM-hard targets, the three MD-based protocols showed similar performance, thus here the prediction methods were combined under the single TBM category

(Figure 5.3-3.4 and, and Appendix 38-40). TBM-easy targets were categorised by CASP if there is a good homologous proteins in the PDB database for TBM targets. The contact-assisted MD-based protocol and fixed local quality assessment guided MD-based protocol were close in performance to each other and both better than the original MD-based protocol of ReFOLD in terms of the cumulative minimum GDT-HA score (∑GDT-HAmin of 9.3105 and 9.286 versus 8.4731, respectively) (Appendix 38), the cumulative mean GDT-HA score (∑GDT-HAmean of 9.7513442 and 9.754875 versus 9.5123675, respectively) (Figure 5.3-5.4 and, and Appendix 38-40). This means that the contact-assisted and the fixed local quality assessment guided restraint strategies were more successful at preventing structural deviations from the native basin compared with the original MD-based protocol of ReFOLD for TBM targets (Figure 5.3-5.4 and, and Appendix 38-40).

In Figure 5.3 it is apparent that the models generated by the contact-assisted and the fixed local quality assessment guided MD-based protocols are closer to that of the starting model than the models generated by the original MD-based protocol of ReFOLD have a wider range of GDT-HA models due to the application of the weaker restraint on C-alphas (see also Figure 5.3-3.4 and, and Appendix 38-40). The cumulative maximum GDT-HA scores of models generated by the contact-assisted MD-based protocol and the fixed local quality assessment guided MD-based protocol are lower than the cumulative maximum GDT-HA score of the models generated by the original MD-based protocol of ReFOLD (∑GDT-HAmax of 10.2563 and 10.2657 versus 10.5197) (Appendix 38). However, the likelihood of selecting the best-generated model is low even where QA methods are used. Therefore, it is pertinent to aim for methods that increase the mean quality of the population of models so as to increase the odds of selecting an improved model.

Although the contact-assisted MD-based protocol performed similarly to the fixed local quality assessment guided protocol, the percentage of improved models generated by the contact-assisted MD-based protocol (~34.45%) is higher than the fixed local quality assessment guided protocol (~33.14%) and the original MD-based protocol of ReFOLD (~26.55%) for TBM targets (Figure 5.4, and Appendix 38). The higher proportion of the improved models from the contact-assisted MD-based protocol increases the chances of the selection of improved models. It is also important that the contact-assisted restraint strategy was able to generate more improved models for TBM targets, because the refinement of the TBM targets has been

challenging historically. They are also more likely to deteriorate in quality compared to the initial structures. The accuracy of the contact predictions might be higher for TBM targets due to the usage of the available structures, and this may also boost the performance of the contact-assisted MD-based protocol compared to the fixed local quality assessment score produced by ModFOLD7. Therefore, we observe that the contact-assisted restraint strategy has provided the most reliable guidance to the MD-based protocol for the TBM targets.

The three MD-based protocols showed a similar trend for FM/TBM targets in terms of the cumulative minimum, mean and maximum GDT-HA scores (∑GDT-HAmin(contact) =3.6736,∑GDT-HAmin(local)=3.6729), ∑GDT-HAmin(ReFOLD)=3.3332, ∑GDT-HAmean(contact)=3.887622, ∑GDT-HAmean(local)=3.898057, ∑GDT-HAmean(ReFOLD)=3.743523), ∑GDT-HAmax(contact)=4.1217, ∑GDT-HAmax(local)=4.0941), ∑GDT-HAmean(ReFOLD)= 4.2023) (Appendix 41). While approximately 25% of the models generated by the contact-assisted MD-based protocol are improved, almost 22% of the models generated by the original MD-based protocol of ReFOLD and roughly 15% of the models by the fixed local quality assessment guided MD-based protocol are improved compared to the initial structure (Figure 5.5-5.6, and Appendix 41-43). It is clear that the contact-assisted MD-based outperformed the other two protocols according to the population of improved models, and the original MD-based protocol of ReFOLD performed much better than the fixed local quality assessment guided MD-based protocol, contrary to the data for the TBM targets.

The application of the gradual restraint based on the CDA score also made a considerable progress for the FM/TBM targets compared to the other two protocols, particularly the fixed local quality assessment guided MD-based protocol. For the application of the fixed local quality assessment guided restraint strategy, a threshold based on the predicted per-residue accuracy score produced by ModFOLD7 was determined by considering the distribution of the scores (see Chapter 2). However, the determination of the threshold was not quite applicable for FM/TBM as the domains were predicted by different prediction methods as TBM and FM. Therefore, the gradual restraint strategy based on the CDA score produced by using DMP performed much better to generate improved models in comparison with the other two protocols.

The three MD-based protocols managed to increase the accuracy of the starting models predicted by FM. Although the original MD based protocol of ReFOLD did not show improved performance compared with the other two protocols according to the cumulative minimum and mean GDT-HA scores ((∑GDT-HAmin(contact) =2.9282,∑GDT-HAmin(local)= 2.9359), ∑GDT-HAmin(ReFOLD)=2.7558),∑GDT-HAmean(contact)=3.1095539,∑GDT-HAmean (local)= 3.1142502) ,∑GDT-HAmean(ReFOLD)= 3.098978)) (Appendix 44), it showed a better performance on the FM targets compared to the FM/TBM and TBM targets (Appendix 38-47). It is also worthy of note that almost half of the models generated by the three MD-based protocols were improved compared to the initial structure, perhaps as there is much more room for improvement with FM target models (Figure 5.7-5.8 , and Appendix 44-47). Furthermore, the population of the models generated by the contact-assisted MD-based protocol (~51.13%) is higher than the fixed local quality assessment guided MD-based protocol (~47.3%) and the original MD-based protocol of ReFOLD (~48.5%) (Figure 5.7-5.8 , and Appendix 44-47).

The Molprobity score was also used to compare the three MD-based protocols. Molprobity is a native structure independent scoring method and all atoms are considered for its calculation, unlike the GDT-HA score. The three MD-based protocols managed to improve the quality of the initial structures according to the cumulative minimum, mean, and maximum scores (Appendix 48-51). This also means that all 3D models generated by the three protocols are improved according to the Molprobity score. Although the original MD-based protocol of ReFOLD outperformed the other two MD-based protocols, the contact-assisted MD-based protocol performed better than the fixed local quality assessment guided MD-based protocol (Figure 5.9, Appendix 48-51).

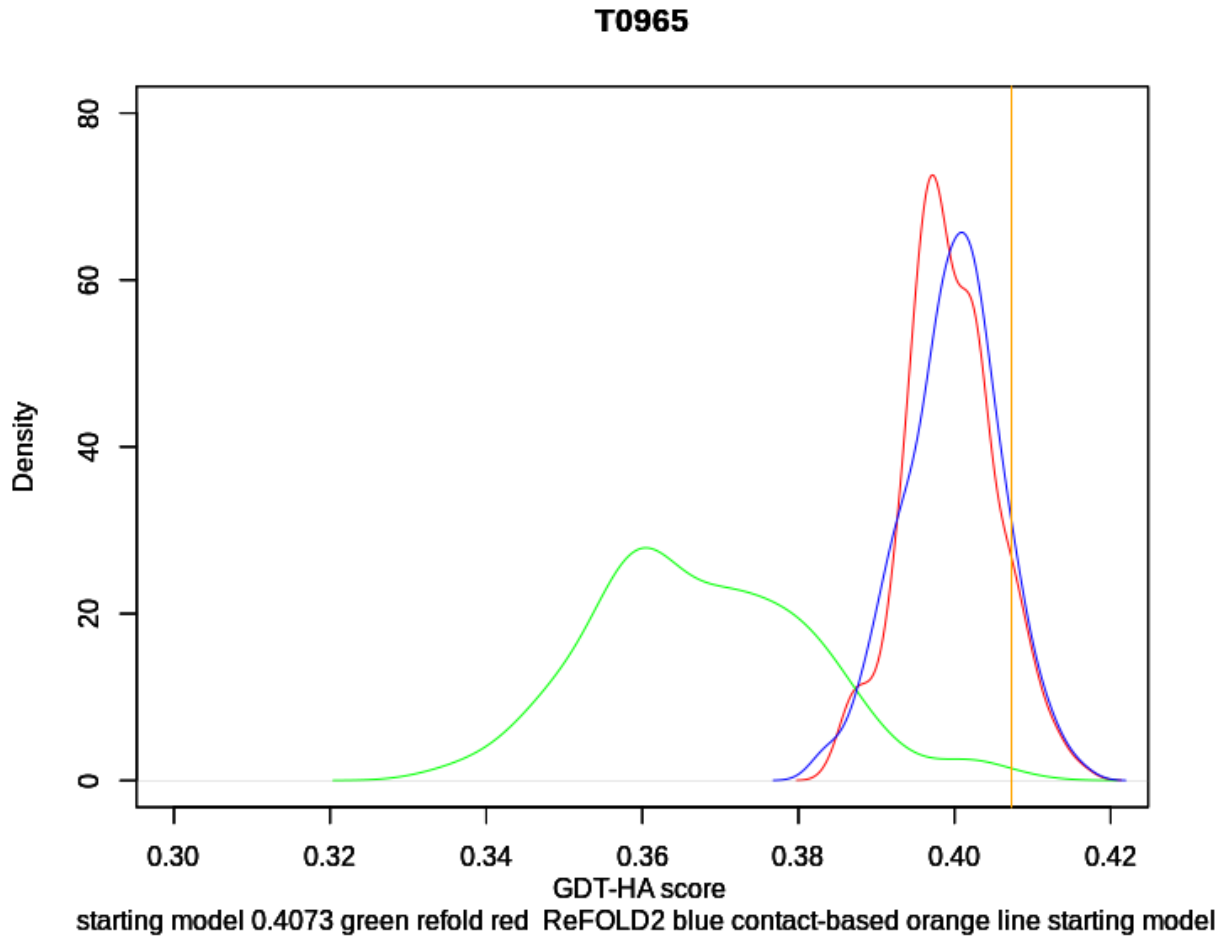

Figure 5. 3 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on a TBM target.

Performance of methods on T0965 (a TBM-hard CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.4073 and higher GDT HA scores are better).

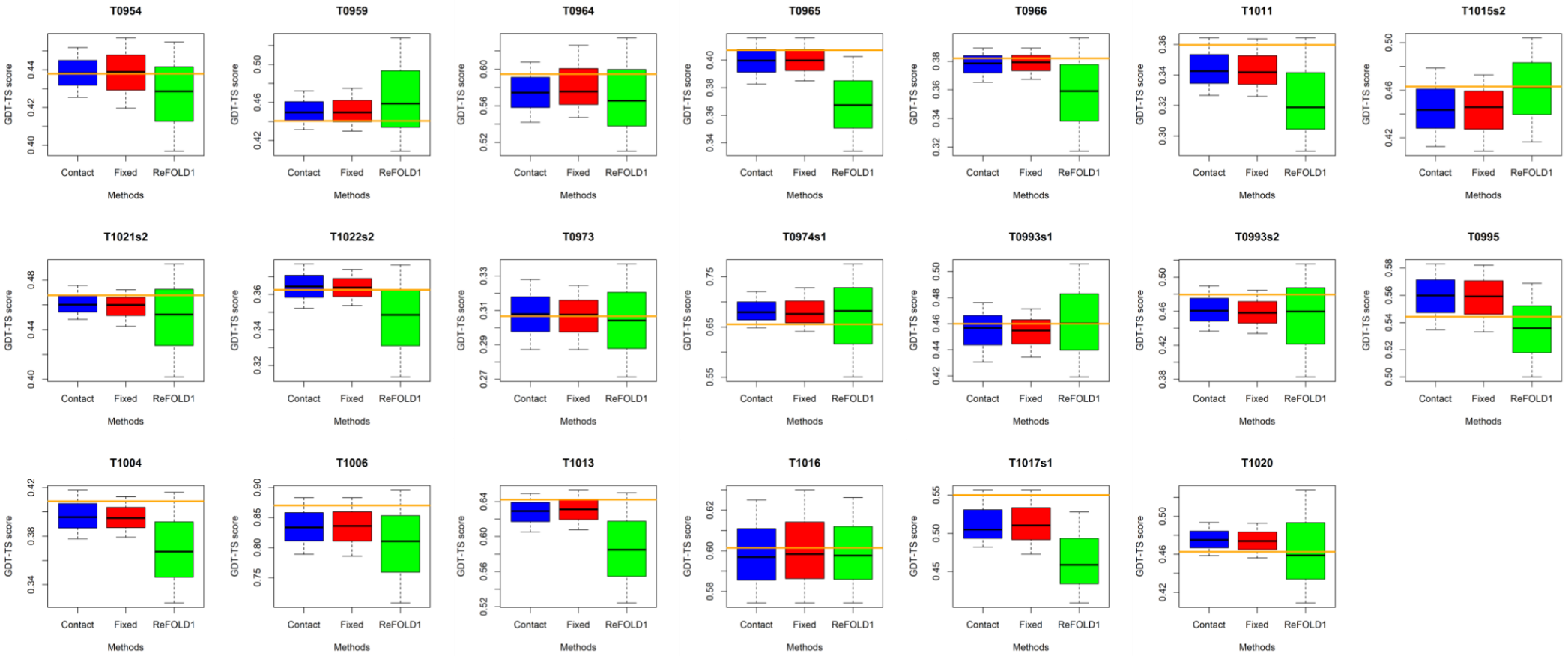

Figure 5. 4 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on the CASP13 TBM targets according to the GDT-HA score.

The blue bars represent the scores of models generated using the contact-assisted MD-based protocol, the red bars represent the scores of models generated using the fixed restraint strategy, the green bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

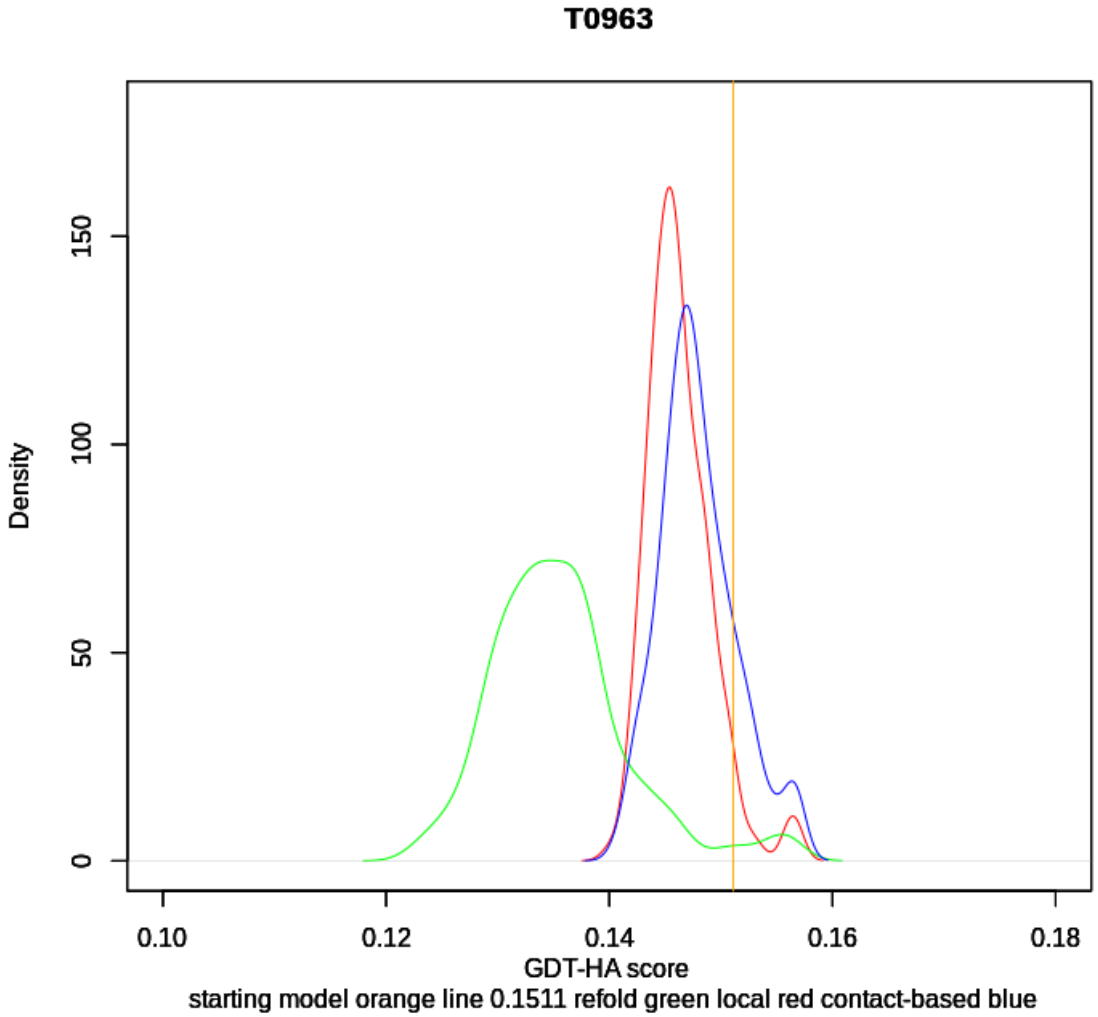

Figure 5. 5 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol an FM/TBM target.

Performance of methods on T0963 (an FM/TBM-hard CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.1511 and higher GDT HA scores are better).

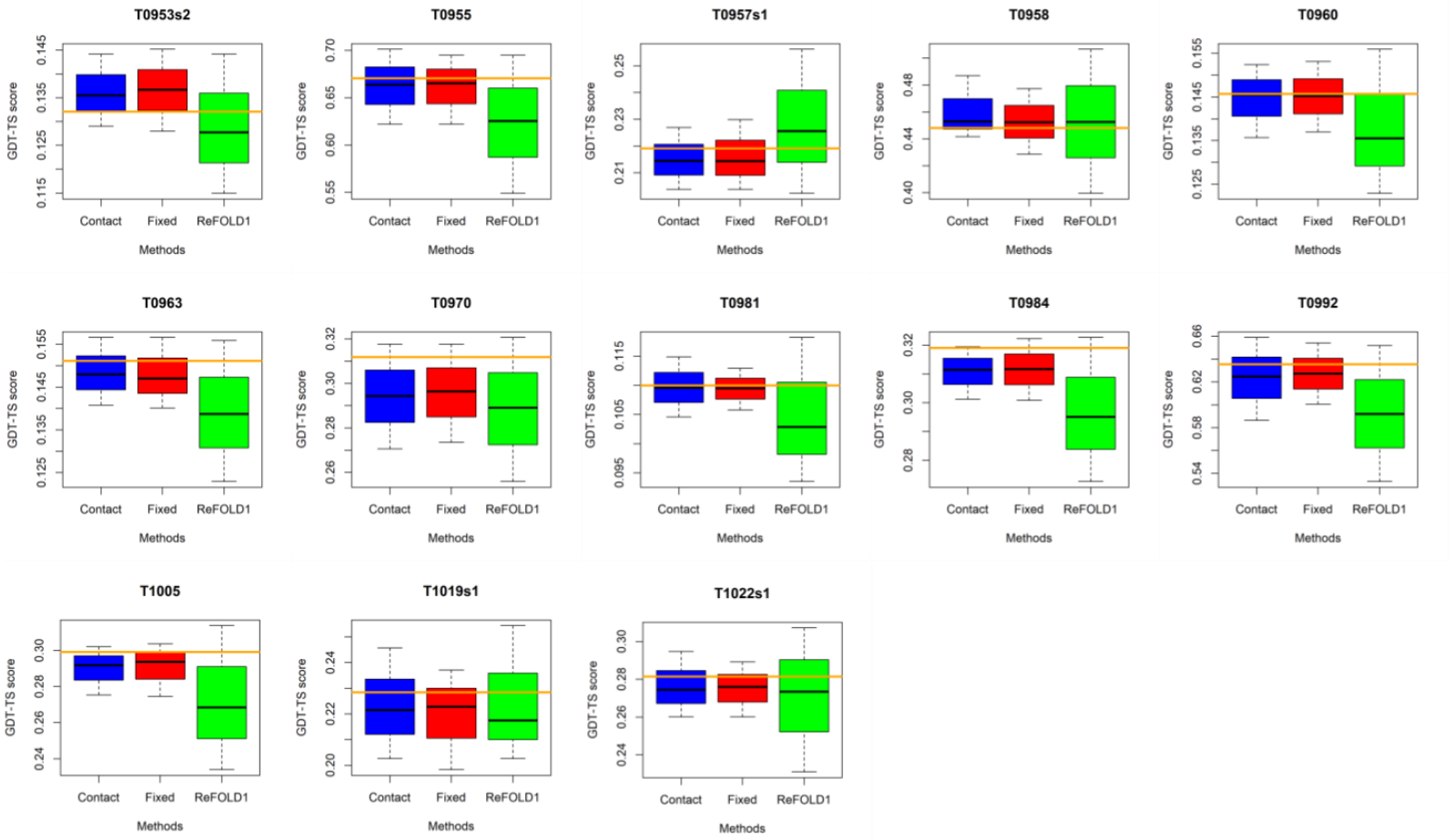

Figure 5. 6 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on the CASP13 FM/TBM targets according to the GDT-HA score.

The blue bars represent the scores of models generated using the contact-assisted MD-based protocol, the red bars represent the scores of models generated using the fixed restraint strategy, the green bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

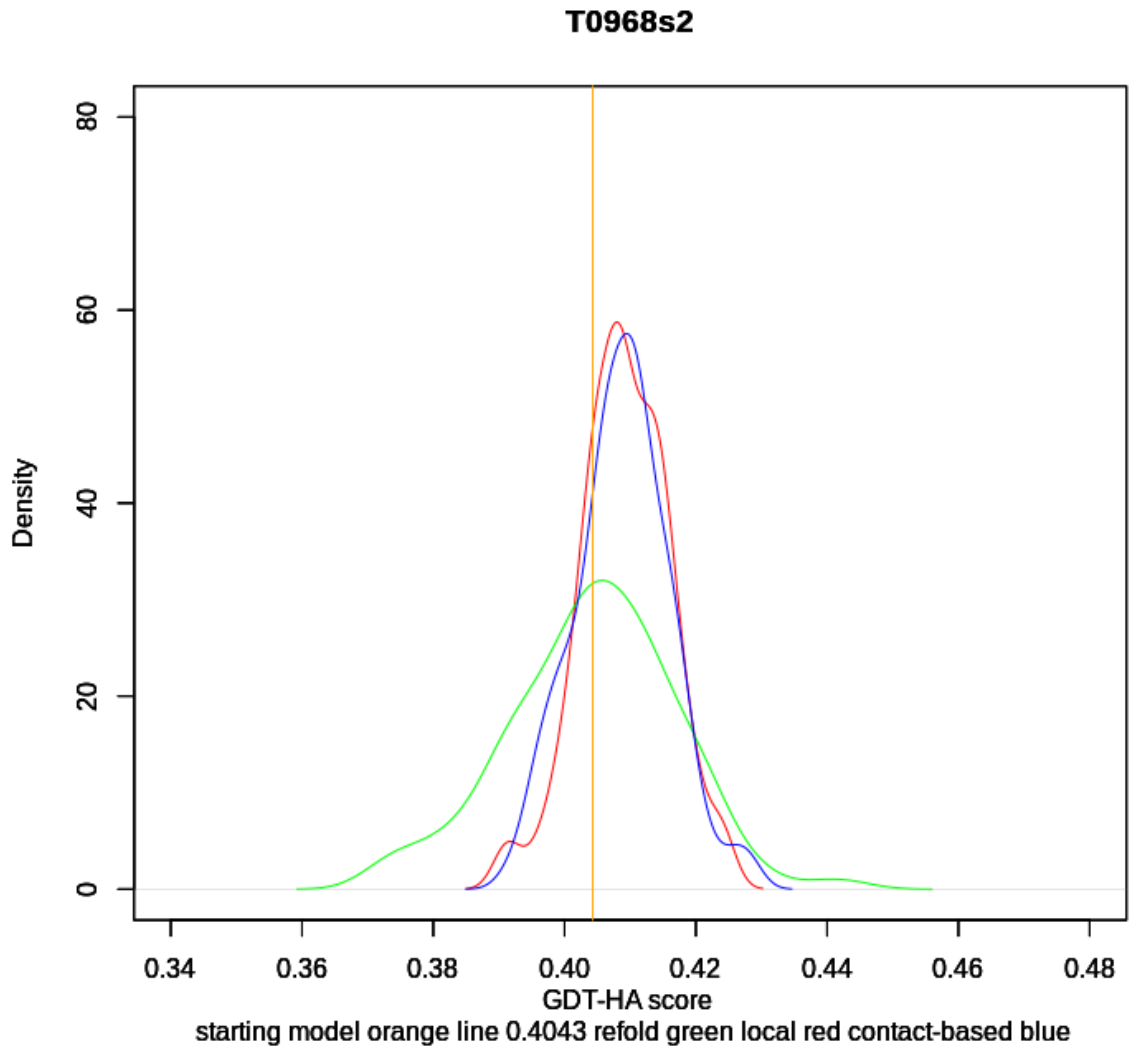

Figure 5. 7 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol an FM target.

Performance of methods on T0968s2 (an FM CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.4043 and higher GDT HA scores are better).

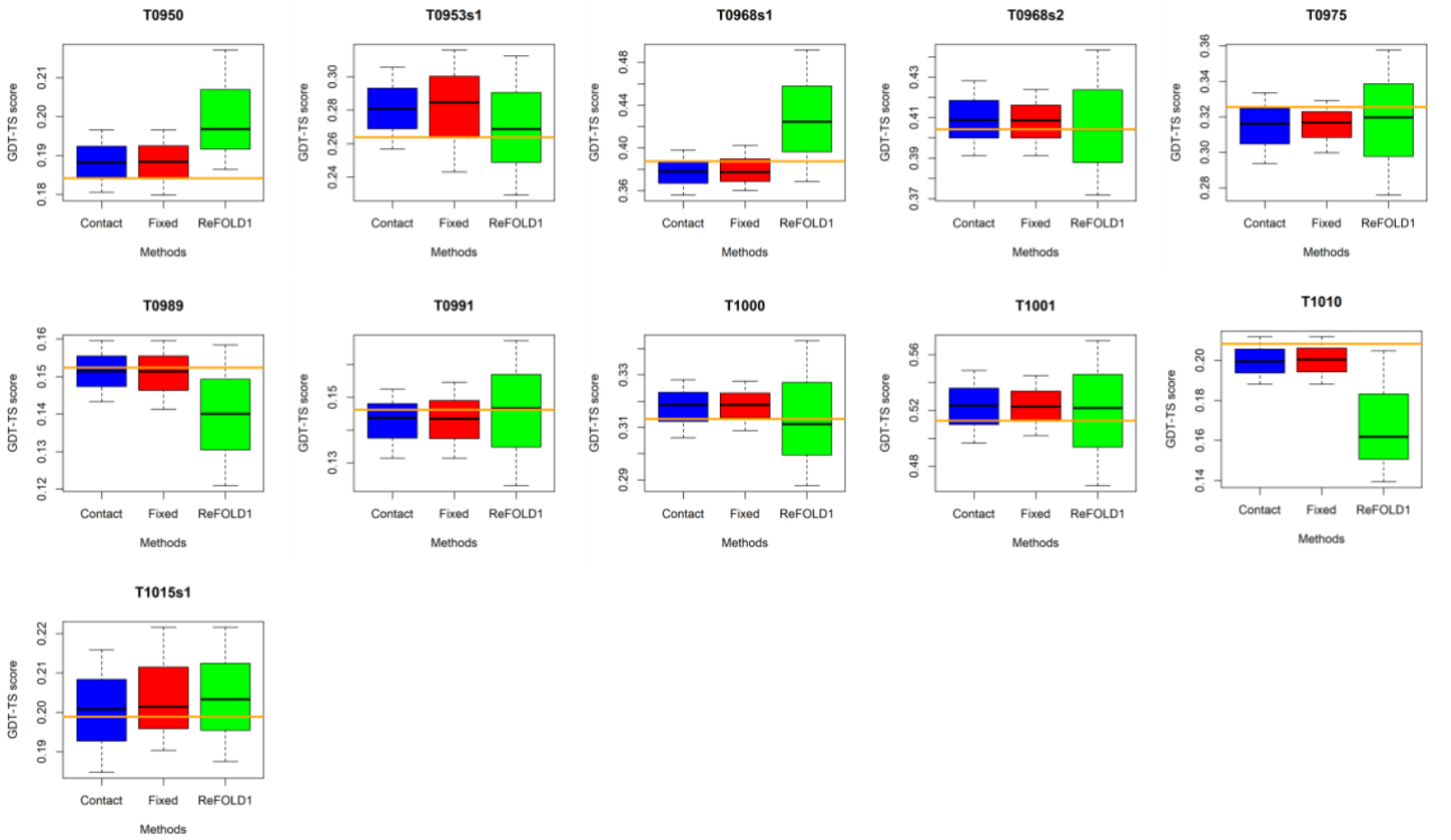

Figure 5. 8 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on the CASP13 FM targets according to the GDT-HA score.

The blue bars represent the scores of models generated using the contact-assisted MD-based protocol, the red bars represent the scores of models generated using the fixed restraint strategy, the green bars represent models generated using the original MD-based protocol of ReFOLD, the black lines represent the median values within each box, and the orange lines represent the starting model for each target (higher GDT-HA scores are better)

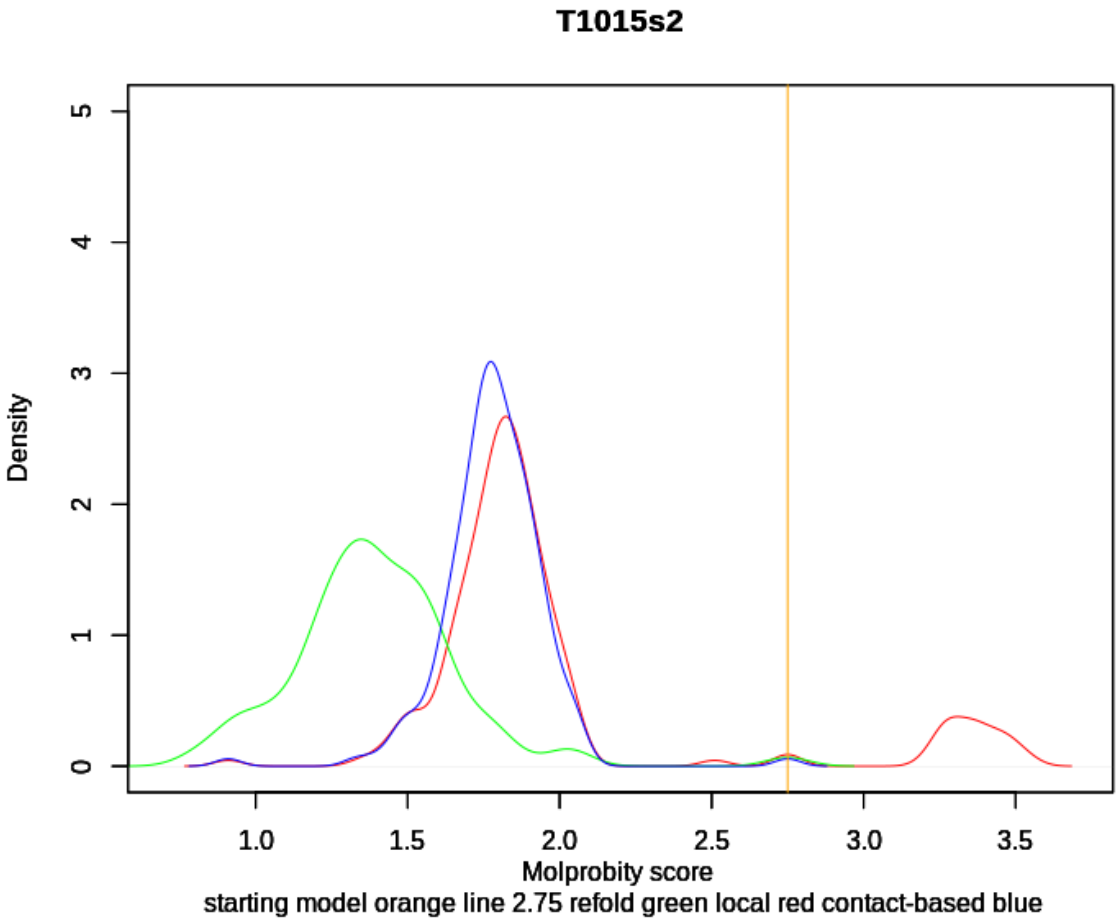

Figure 5. 9 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol a TBM target according to Molprobity score.

Performance of methods on T1015s2 (a TBM CASP13 target) according to Molprobity score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the Molprobity score of the initial structure was 2.75 and lower Molprobity scores are better).

5.5 Conclusions

In this Chapter, prior knowledge from contact prediction data and local quality scores has been applied in refinement pipelines in order to increase the accuracy of the 3D models beyond that of the initial structures. Predicted residue-residue contacts have been used in order to build 3D models for over two decades (Schaarschmidt et al. 2018; Skolnick, Kolinski, and Ortiz 1997; Taylor et al. 2014). In recent years, the contact prediction methods based on the machine learning and coevolution-derived approaches have made significant progress towards higher accuracy. In CASP13, the top-performing groups managed to reach around 70% precision. The integration of such contact prediction methods in refinement pipelines may be a major contributor towards more consistent sampling. Therefore, here we describe the first attempt at using predicted contacts to provide a reliable guide to the MD-based refinement protocol.

Although the restraint strategy based on the fixed local quality assessment score has made significant progress in terms of preventing refinement models from structural deviations, it was worth investigating the potential of integrating state-of-the-art contact predictions. The contact predictions made by MetaPSICOV were used to produce the original CDA score method used by ModFOLD6 by considering the distance between residues in the predicted 3D model and predicted contacts. In this chapter, the CDA scoring method was upgraded using DMP (Kandathil et al., 2019a; Maghrabi & McGuffin, 2017).

The CDA method was then used to guide the original MD-based protocol of ReFOLD via gradual restraints based on the distribution of scores. The CDA score also varies from 0 to 1, therefore a gradual restraint strategy could be applied with the magnitude of restraints depending on the CDA score. Here we chose to apply a gradual restraint based on the CDA score as we previously learned that using a fixed threshold for the whole structure was not always appropriate, particularly in the case of FM/TBM targets (Chapter 3).

In the final stage, the performance of the new contact-assisted MD-based protocol was benchmarked using to GDT-HA and Molprobity scores to measure improvements in observed

model quality. The performance of the contact-assisted MD-based protocol was then compared with the fixed local quality assessment guided protocol, which was tested in CASP13, and the original MD-based protocol of ReFOLD, which was tested in CASP12.

The contact-assisted MD-based protocol showed roughly similar performance to the fixed local quality assessment guided according to the cumulative minimum, and mean GDT-HA scores, and both protocols performed much better than the original MD-based protocol of ReFOLD in terms of preventing 3D models from undesired structural deviations. It is evident that the contact-assisted and the fixed local quality assessment guided restraint strategies managed to prevent more 3D models from detrimental structural deviations compared to the weak harmonic restraint on all C-alphas applied during the original MD-based protocol of ReFOLD.

The percentage of the improved models is also an important criteria for the refinement pipeline, as higher populations of better models allow for better odds in the scoring and selection stage. It was discovered that the contact-assisted MD-based protocol managed to generate more improved models in contrast with the two other protocols. The percentage of improved models increased from 29.53% to 31.31% by applying the fixed local quality assessment guided restraint strategy, and this further increased to 35.73% with the application of the contact-assisted restraint strategy for all targets. The contact-assisted restraint strategy provided more consistent refinement compared to the fixed local quality assessment guided restraint strategy. This is likely due to the fact that the contact prediction methods may be relatively more accurate than local QA scoring methods in cases of low similarity between the target and known structures and sequences (i.e., FM targets or domains).

Chapter 6 Synthesis, Conclusions and Next Directions

Work presented in this chapter has been submitted in the following paper:

Recep Adiyaman and Liam James McGuffin, 2021. ReFOLD3: refinement of 3D protein models with gradual restraints based on predicted local quality and residue contacts. Submitted to Nucleic Acid Research (Web Server Issue 2021).

6.1 Synopsis of Study

The incorporation of MD-based protocols for the refinement of the predicted 3D structures has been effective since CASP10, and such approaches have showed promising performance in consistently increasing the accuracy of refinement targets. Our group developed the ReFOLD server to refine predicted 3D models with much less computational effort in comparison with other MD-based protocols tested in CASP12. Nevertheless, undesired structural deviations from the native structure were seen in 3D models generated by ReFOLD during the CASP12 experiment (Shuid et al., 2017). This thesis has focused on the further development of ReFOLD by exploiting our knowledge of predicted local model quality scores and residue contacts to guide the refinement of 3D models towards the native basin.

6.1.1 The Usage of the Local Quality Assessment to Provide Guidance to the Original MD-Based Protocol of ReFOLD

Model Quality Assessment Programs (MQAPs) are used to identify the most native like structure among decoys generated by TBM and/or FM methods. These programs are also capable of producing local quality estimation scores along with the global score. The local quality estimation of predicted structures aims to provide the accuracy of each residue in 3D models versus the native structure. The ModFOLD server, developed by our group, has been continuously tested in the CASP and CAMEO experiments and it ranks among the top few groups in terms of assessing the global and local quality of the predicted 3D models (Maghrabi & McGuffin, 2019, 2017). The per-residue accuracy score produced by the ModFOLD server shows the predicted C-alpha distance of each residue in a 3D model from the equivalent residue in the native structure. In the first part of the study, the per-residue accuracy score was used to provide reliable guidance to the original MD-based protocol of ReFOLD, directing models closer towards the native basin.

In Chapter 2, the local quality assessment guided restraint strategy was applied by determining a threshold according to the distribution of the per-residue accuracy scores produced by ModFOLD6 scores were used identify the poorly predicted regions in 3D models for selective refinement, while

restraints were imposed on the well-predicted regions, during MD simulations (Maghrabi & McGuffin, 2017; Shuid et al., 2017). The performance of the local quality assessment guided restraint strategy was also compared with the original MD-based protocol of ReFOLD for the generation of improved targets on CASP12 initial structures. It is clear that the 3D models generated by our local quality assessment guided MD-based protocol are much closer to the native basin than those from the original MD-based protocol of ReFOLD, according to the observed quality scores, especially for TBM targets or domains. This means that the per-residue accuracy score produced by ModFOLD6 can be used to successfully direct the generation of the 3D models towards the native basin. The work presented in Chapter 2 describes the very first attempt to utilise local model quality predictions in order to guide an MD-based refinement pipeline to produce models that are closer to native structures.

6.1.2 The Performance of Our Refinement Pipeline in CASP13 and the Gradual Restraint Strategy Based on the Local Quality Estimation

The accuracy of the local quality estimation score was significantly improved in ModFOLD7 by our group (Maghrabi & McGuffin, 2019). Therefore, ModFOLD7 was used to upgrade the local quality assessment guided MD-based protocol to guide the MD simulation and identify the most-native like structures. The upgraded version of the MD-based protocol was also used to refine CASP13 structures in the regular prediction and refinement categories, as described in Chapter 3. The refinement pipeline was ranked among the top 10 approaches in CASP13. This shows that our deployment of local quality scores to guide our rapid MD-based protocol had enabled us to be more competitive.

ModFOLD7 was also used to select the most improved models among the 3D models generated by our refinement pipeline in CASP13. While ModFOLD6 selected the improved 3D models compared to the initial structure for 31% of the targets, which were refined in the study of Chapter 2, ModFOLD7 managed to select the improved models for 41% of the refinement targets in CASP13, according to the GDT-HA scores. This means that the upgraded version of the

ModFOLD server performed much better in terms of the identification of the optimal model from among those generated by the local quality assessment guided restraint strategy.

The local quality assessment guided restraint strategy was implemented by applying a threshold based on the per-residue accuracy score in CASP13. It should be noted that the CASP13 targets were relatively larger multi-domain structures in comparison with the previous CASP targets. Furthermore, the domains might have been predicted by different prediction methods depending on the usage of available templates. For this reason, we developed a gradual restraint based on the per-residue accuracy score to consider the required level of refinement for each residue rather than the application of a blanket threshold to the whole structure. Our application of these gradual restraints has also been a unique pioneering strategy for MD-simulations.

Although the gradual and the fixed restraint strategy based on the local quality estimation showed similar performances according to the observed scores, the population of the improved models had increased from ~28.86% to ~34.36% following the application of the gradual restraint strategy. Our prediction pipeline, including the IntFOLD, ModFOLD, FunFOLD and ReFOLD servers, has been designed to provide an understanding of protein structures and functions. The application of the gradual restraint strategy may also boost the overall performance of the prediction pipeline in terms of increasing the accuracy of the predicted 3D structures.

The gradual restraint strategy was also used to refine the SARS-CoV-2 targets with the usage of ModFOLD8 to generate the per-residue accuracy score and identify the improved models for the CASP Commons COVID-19 initiative. Our independent pipeline has helped us to generate the initial models for the SARS-CoV-2 targets using the IntFOLD server, identify the best predicted 3D models using the ModFOLD server, and then refine the best-predicted structures using the gradual restraint strategy. Using our complete pipeline, we managed to provide a considerable proportion of the top 10 predicted structures for the SARS-CoV-2 targets according to the initial CASP official estimates of model accuracy (official results according to the observed score are not yet available at the time of writing this section). This success highlights the importance of the role of our prediction pipelines for the elucidation of the structures for key protein targets whose experimental structures are not yet solved.

6.1.3 Increasing the Accuracy of the Predicted Protein-Ligand Binding Sites

The determination of the protein-ligand binding sites is a vital part of the elucidation of protein functions. The FunFOLD server was developed by the McGuffin group for the prediction of the protein-ligand binding sites, utilising the knowledge of the available structures (Roche et al., 2012a, 2013b; Roche, Tetchner, et al., 2011; Roche & McGuffin, 2016a). In Chapter 4, we developed the binding-site focused MD-based protocol to improve the local quality of the binding sites rather than the whole structure. For the application of the binding-site focused MD-based protocol, the binding sites predicted by FunFOLD3 were further refined by restraining the rest of the structure during the MD-simulations. The MCC and BDT scores were also produced to investigate the improvement in the predicted binding site residues compared with the observed binding site residues.

The binding site-focused MD-based performed well in terms of increasing the accuracy of all predicted binding site regions, whether they were initially well or poorly predicted according to the BDT and MCC scores (Roche et al., 2010). The prediction methods used for generating the initial structures were found to be a determinant of the quality of the predicted binding sites, as well as the availability of templates, yet the binding site-focused strategy showed a considerable improvement across TBM and FM target categories. The development of the binding site-focused MD-based protocol shows promise, but it should be tested on additional data sets (e.g. CASP14) to further evaluate the significance its reliability across a larger number of targets.

6.1.4 The Development of the Contact-Assisted MD-Based Protocol

Although the contact prediction category was introduced in CASP2, the residue-residue contact prediction methods had only reached a useful level of accuracy by CASP11, through the utilisation of machine learning and coevolution-derived approaches (Schaarschmidt et al., 2018; Shrestha et al.,2019). The contact prediction methods based on the deep neural network algorithms have pushed the accuracy further, reaching up to 70% precision in the last CASP experiment (Shrestha et al., 2019). Contact predictions have been useful for the prediction of the 3D models, drug design

(Kliger et al., 2009), and model quality estimation (Miller & Eisenberg, 2008; Z. Wang et al., 2011).

The residue-residue contact predictions were first used for the generation of the Contact Distance Agreement (CDA) score by ModFOLD6 for the quality estimation of the 3D protein models in our prediction pipeline (Maghrabi & McGuffin, 2017). The CDA score relies on the agreement between the contact predictions made by MetaPSICOV and the contacts measured according to the Euclidean distance (in Å) between residues in the predicted 3D model (Jones et al., 2015; Jones & Kandathil, 2018; Kosciolek & Jones, 2016; Maghrabi & McGuffin, 2017).

DeepMetaPSICOV (DMP), which is the upgraded version of MetaPSICOV, was ranked among the top 5 contact prediction approaches and achieved an average precision of ~60% in CASP13 (Kandathil et al., 2019a). We proposed to utilise the CDA score, which was generated using DMP, to provide guidance for the original MD-based protocol. A gradual restraint strategy was also applied to restrain the accurate contact predictions, according to the CDA scores, by considering the degree of refinement required for each residue. The performance of the contact-assisted MD-based protocol was also compared with the fixed local quality assessment guided MD-based protocol, which was tested in CASP13, and the original MD-based protocol of ReFOLD which was tested in CASP12.

The fixed local quality assessment guided restraint strategy managed to prevent the 3D models from the structural drifts which were observed in the 3D models generated by the original MD-based protocol of ReFOLD (see Chapter 2). The contact-assisted MD-based protocol also performed similarly according to the observed scores. The application of the contact-assisted MD-based protocol made significant progress in terms of the population of the improved models. The percentage of the improved models generated by the original MD-based protocol of ReFOLD was 29.53%, and this increased to 31.31% with the application of the fixed local quality assessment guided restraint strategy. The improvement was also further increased to 35.73% by the contact-assisted MD-based protocol for all CASP13 targets. Thus, it is evident that the contact-assisted MD-based protocol significantly increased the population of the improved models, and this enables a higher probability of identifying improving models in the scoring stage.

6.1.5 The Participation of Our Refinement Pipeline in CASP14

Our refinement pipeline was also tested in CASP14 from May to September in 2020, but the official result was not released by CASP at the time of writing this section. For the refinement of 3D models of proteins, we used a modified version of our automated ReFOLD method. Our new refinement pipeline, ReFOLD3, consisted of four protocols that were also similar to the protocol in CASP13 (see Chapter 3). The major improvement for ReFOLD version 3 was the accommodation of the two MD-based strategies developed in Chapter 3 and Chapter 5 (Figure 6.1). The first protocol used a rapid iterative strategy (i3Drefine), and the second and third protocols both employed a more CPU/GPU intensive molecular dynamic simulation strategy.

The second protocol included the application of the gradual restraint strategy, which was guided by the per-residue accuracy scores obtained from ModFOLD8. The per-residue accuracy scores were used for the identification of the poorly modelled regions, which were then targeted for refinement to increase the accuracy of the 3D model, as described in Chapter 3.

For the third protocol, the contact-assisted MD-based protocol was applied to refine the initial structures (see Chapter 5). DeepMetaPSICOV was also run for each target sequence, and then the Contact Distance Agreement (CDA) scores were generated for each model to guide the MD-based protocol. Another gradual restraint strategy was implemented by considering the distribution of the CDA scores during the MD simulation, as described in Chapter5.

Refined models generated from the first three protocols were then assessed and ranked using ModFOLD8_rank. The fourth protocol was a combination of the approaches, where the top-ranked model from the 2nd and 3rd protocol was then further refined using i3Drefine. Finally, all the refined models generated by each of these protocols and the starting model were pooled and re-ranked again using ModFOLD8_rank, and then the final top 5 models were submitted. All of the refined models were additionally ranked by ModFOLD6_rank and ModFOLD7_rank. This is so that we can compare the performance of the different ModFOLD versions in terms of the selection of refinement models, when the native structures are released by the CASP assessors (at the time

of writing the prediction part of the experiment is over, but the results based in the observed structures are not yet available).

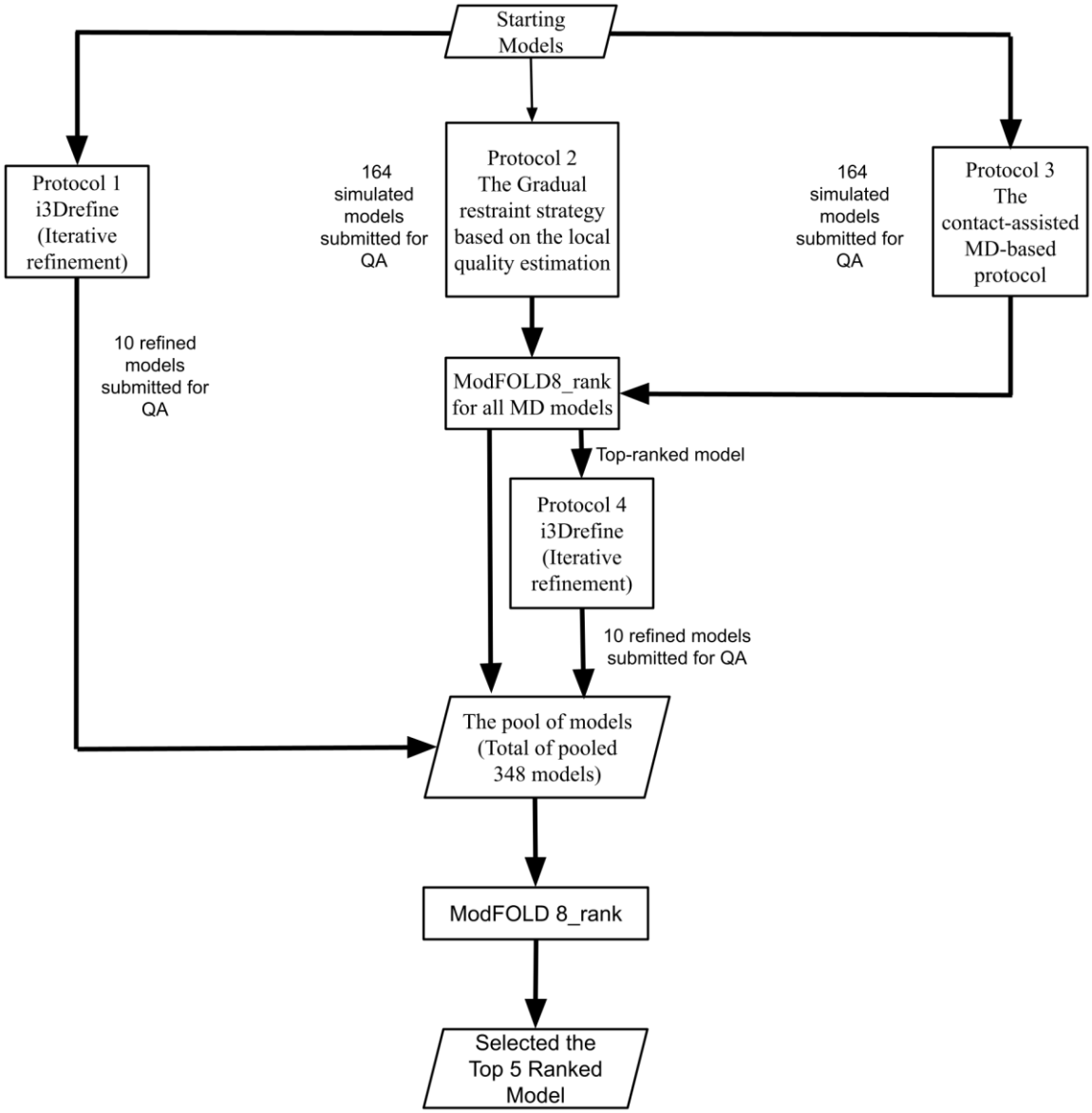

Figure 6. 1 Flowchart of our CASP14 refinement pipeline.
The refinement of the starting model using i3Drefine (Protocol 1), the gradual restraint strategy based on the local quality estimation (Protocol 2), the contact-assisted MD-based protocol (Protocol 3), and the second round of i3Drefine iterative refinement strategy (Protocol 4). All refined models were ranked by the MoldFOLD8 server using the ModFOLD8_rank option (optimised for selecting the best top model).

6.2 Conclusions

The ReFOLD server was developed to refine 3D protein models without requiring the huge computational-resources typically required by other MD-based protocols. This progress has made the refinement of thousands of structures possible with less computational effort. Through our server we provide free access to rapid refinement to researchers all around the world. However, despite many initial successes using the original MD-based protocol of ReFOLD, some undesired structural deviations from the native basin were observed in many of the 3D models that were generated during CASP12 experiment.

Our initial aim was to prevent 3D models from structural deviations by utilising the per-residue accuracy scores produced by the ModFOLD server. The initial aim was met by applying a threshold that was based on the distribution of the per-residue accuracy scores to selectively refine the poorly predicted regions during the MD simulations. It is clear that the usage of the local quality estimation has provided a reliable guidance to the MD-based protocol, producing models that are closer to the observed structures (Chapter 2). The local quality assessment guided MD-based protocol also showed a competitive performance when it was tested in the CASP13 experiment, where it ranked among the top 10 refinement approaches.

Following our analysis of the CASP13 results, we found that the determination of a single threshold based on the per-residue accuracy score was less applicable in the case of multi-domain targets, particularly those with a mixture of domains of different difficulty and/or those with large deviations in model quality. For this reason, a gradual restraint strategy was first proposed to consider a more targeted level of refinement that would be appropriate for each residue in the 3D models. The application of the gradual restraint strategy based on the local quality estimation had the effect of increasing the population of the improved models. The population of the improved models generated by the original MD-based protocol of ReFOLD was 29.53%, following the application of the fixed restraint strategy this increased by ~2% to 31.31%, and applying the gradual restraint strategy based on the local quality estimation managed to increase it by a further 3% to 34.36%. The gradual restraint based on the local quality estimation was also used to refine

our 3D models of the SARS-CoV-2 targets as part of the CASP Commons COVID-19 initiative. A significant proportion of the top 10 scoring models were submitted by our group, according to the CASP official quality estimations results, so our overall pipeline showed impressive performance compared with the many different approaches in the initiative.

Highly accurate contact predictions have also boosted the performance of many prediction pipelines since CASP13. In this thesis, we also described the first attempt to utilize the contact predictions in a refinement pipeline to direct the generation of the 3D models closer towards experimental accuracy. Although the contact-assisted MD-based protocol performed similarly to the local quality assessment guided restraint strategies, according to the observed scores, the population of the improved models was further increased to 35.73% for all targets. Therefore, it can be said that the contact-assisted MD-based protocol outperformed the local quality assessment guided MD-based protocol, in terms of increasing the population of improved models. With the application of the contact-assisted MD-based protocol we have seen considerable progress towards a more consistent refinement pipeline. Contact predictions may provide more reliable guidance for refinement where there is low similarity between the target sequence and known structures, such as in the case of targets containing FM domains.

Unlike other MD-based protocols developed in this study, the binding site-focused MD-based protocol was developed to refine the binding site regions rather than the whole protein structure. It should be noted that the quality of all predicted binding sites was improved by the binding site-focused MD-based protocol. The integration of the binding-site MD-based protocol with the FunFOLD server may also provide a more accurate prediction of protein-ligand binding site regions to elucidate protein-ligand interactions at an atomic level.

6.3 Future Directions

Overall, three different MD-based protocols were developed for the refinement of the whole structure, and one protocol was developed for the refinement of the predicted binding sites. A more consistent refinement of 3D models has also been achieved utilising the local quality estimation

and contact prediction methods to guide the original MD-based protocol of ReFOLD. Nevertheless, there are still many ways that we may be able to further improve the performance of the refinement pipeline that can be considered as our future goals:

The gradual restraint strategy based on the per-residue accuracy score produced by ModFOLD8 and the contact-assisted MD-based protocol were used to refine the 3D models in CASP14, so that the our newly developed protocols could be blind tested by independent assessors. After the official results are released by the assessors, the performance of the MD-based protocol will be analysed on the CASP14 dataset. In the short term, the ReFOLD server can be updated with one of or both of these MD-based protocols by considering their relative performance in CASP14 and the computational resource availability.

The weak harmonic positional and gradual restraint strategies were also applied during MD simulations, but the strength of the restraints seems to be an important parameter for a successful refinement. Therefore, other kinds of restraint parameters could be explored in order to more fully develop the restraint strategy.

Several parameters, such as temperature, duration of the simulation and force field parameters from the original ReFOLD methods were kept fixed and were also used by the new MD-based protocols. This was so we could control for these parameters and fairly test our new restraint strategies. However, these parameters may also need to be optimised, and this optimisation might contribute further to the improvement in the quality of refined models.

We proposed that an iterative MD-based protocol with gradual restraint strategy may show a better performance to avoid structural deviations, but it is worthy of note that the usage of restraints may also limit the extent of refinement. Such structural deviations might have also been caused by force field inaccuracies. The latest versions of CHARMM (Huang et al., 2017) and Amber (Maier et al., 2015) force fields might have the potential for directing the generation of the 3D models closer towards the native basin. The MD simulations were also run using NAMD (Phillips et al., 2005). The performance of the latest version of the force fields and other software for molecular

mechanics modelling such as OpenMM (Eastman et al., 2013, 2017) and GROMACS (Abraham et al., 2015) can be investigated further.

The residue-residue contact predictions made by DeepMetaPSICOV were utilised to produce the CDA score (Kandathil et al., 2019a; Maghrabi & McGuffin, 2017), then guide the original MD-based protocol of ReFOLD. Other contact prediction methods or a consensus method for the prediction of residue-residue contacts should be investigated in order to provide more reliable guidance to the MD-based protocol.

There are also many different local quality assessment approaches along with the ModFOLD server that should be tested to guide the MD-based protocols. Therefore, the potential of the quality assessment approaches such as ProQ4 (Hurtado et al., 2018), and QMEANDisCo (Studer et al., 2020) could be explored to provide better guidance as an alternative to the ModFOLD server for providing reliable guidance, ranking and selection of the final refined models.

When the restraints were applied, the generation of the improved models becomes better, but the reliable selection of improved models is still challenging. The ReFOLD method involves the submission of the refined models to ModFOLD6 to assess the quality of the refined models. In CASP14, we also submitted the 3D models generated by the refinement pipeline to ModFOLD6, 7 and 8 to compare the performance of these versions when the native structures are released by CASP (at the time of writing this section the native structures are not available yet).

Although ModFOLD7 performed better than ModFOLD6 in terms of the selection of the improved models generated in the refinement pipeline, the ModFOLD server was trained to identify the most native-like structures generated in the main prediction pipeline, which included a variety of models with a large range of quality, rather than the refinement pipeline, which include very similar models with a narrow range of quality (see Chapter 2 and Chapter 3). Nevertheless, the ModFOLD server has not found to be sufficiently successful to select optimal models as in the prediction pipeline. Furthermore, alternative versions of ModFOLD may also be developed specifically for the selection of improved models, generated by the MD-based refinement approaches, which may have smaller differences. Different energy function-based methods have also been used to rank the

3D models by many prediction groups in the refinement category of CASP experiments. It should be noted that the ModFOLD servers regular outperforms all of these older energy-based functions in quality estimation benchmarks (see CAMEO website). A special optimised version of the ModFOLD server for use with refinement pipelines, may have the potential of outperforming the energy functions further still.

With the development of the binding site-focused MD-based protocol, we successfully improved the quality of the binding sites predicted by the FunFOLD server, which may provide a better understanding of protein-ligand binding site interactions. The binding-site focused MD-based protocol may also be integrated with the FunFOLD and IntFOLD servers, to enable more accurate binding site predictions (Roche et al., 2013a; Roche & McGuffin, 2016a).

The CASP assessors have sometimes provided predictors with clues about the focused regions, which need to be refined in certain targets, and we can also start to use that information for a more targeted refinement strategy in the next CASP experiments. In the past, we have shied away from using such information for the development of our automated tools, which are intended to be integrated into server pipelines, as most general users would not have such information. However, there is the possibility of developing servers that can take this information from user inputs, as some advanced users may know which regions of their model/s they wish to target for further refinement. For example, we were successful in refining the specific regions of the 3D models with the application of our binding site-focused MD-based protocol. So, we can also provide the option of the refinement of the user-selected regions rather than the whole 3D model, for the cases where users may need more targeted refinement for different purposes.

Finally, the IntFOLD5 server was integrated with the original ReFOLD server, providing users with an option for refinement of the generated 3D models via a simple “refinement button” (McGuffin et al., 2019). The ReFOLD server will be upgraded to use the MD-based protocols developed in this study and will also be integrated with future versions of the IntFOLD server.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindah, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Adamczak, R., Porollo, A., & Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, *59*(3), 467–475. https://doi.org/10.1002/prot.20441

Adhikari, B., Bhattacharya, D., Cao, R., & Cheng, J. (2015). CONFOLD: Residue-Residue Contact-guided ab initio Protein Folding. *Proteins*, *83*(8), 1436–1449. https://doi.org/10.1002/prot.24829

Adhikari, B., & Cheng, J. (2016). Protein residue contacts and prediction methods. In *Methods Mol Biol* (Vol. 1415, pp. 463–476). https://doi.org/10.1007/978-1-4939-3572-7_24

Adhikari, B., Hou, J., & Cheng, J. (2018). DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, *34*(9), 1466–1472. https://doi.org/10.1093/bioinformatics/btx781

Adiyaman, R., & McGuffin, L. J. (2019). Methods for the Refinement of Protein Structure 3D Models. *International Journal of Molecular Sciences*, *20*(9). https://doi.org/10.3390/ijms20092301

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125

Allen, F., Almasi, G., Andreoni, W., Beece, D., Berne, B. J., Bright, A., Brunheroto, J., Cascaval, C., Castanos, J., Coteus, P., Crumley, P., Curioni, A., Denneau, M., Donath, W., Eleftheriou, M., Fitch, B., Fleischer, B., Georgiou, C. J., Germain, R., … Zhou, R. (2001). Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, *40*(2), 310–327. https://doi.org/10.1147/sj.402.0310

Aloy, P., Stark, A., Hadley, C., & Russell, R. B. (2003). Predictions Without Templates: New Folds, Secondary Structure, and Contacts in CASP5. *Proteins: Structure, Function and Genetics*, *53*(SUPPL. 6), 436–456. https://doi.org/10.1002/prot.10546

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. In *Nucleic Acids Research* (Vol. 25, Issue 17, pp. 3389–3402). Nucleic Acids Res. https://doi.org/10.1093/nar/25.17.3389

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, *36*(Database issue), D419-25. https://doi.org/10.1093/nar/gkm993

Anfinsen. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230. https://doi.org/10.1126/SCIENCE.181.4096.223

Anfinsen, Haber, E., Sela, M., & White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(9), 1309–1314. https://doi.org/10.1073/pnas.47.9.1309

Antczak, P. L. M., Ratajczak, T., Blazewicz, J., & Lukasiak, P. (2015). SphereGrinder-reference

structure-based tool for quality assessment of protein structural models. *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*, 665–668. https://doi.org/10.1109/BIBM.2015.7359765

Ashraf, G., Greig, N., Khan, T., Hassan, I., Tabrez, S., Shakil, S., Sheikh, I., Zaidi, S., Akram, M., Jabir, N., Firoz, C., Naeem, A., Alhazza, I., Damanhouri, G., & Kamal, M. (2014). Protein Misfolding and Aggregation in Alzheimer's Disease and Type 2 Diabetes Mellitus. *CNS & Neurological Disorders - Drug Targets*, *13*(7), 1280–1293. https://doi.org/10.2174/18715273136661409170955l4

Bacardit, J., Widera, P., Márquez-Chamorro, A., Divina, F., Aguilar-Ruiz, J. S., & Krasnogor, N. (2012). Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. *Bioinformatics (Oxford, England)*, *28*(19), 2441–2448. https://doi.org/10.1093/bioinformatics/bts472

Baldwin, R. L. (2007). Energetics of Protein Folding. *Journal of Molecular Biology*, *371*(2), 283–301. https://doi.org/10.1016/j.jmb.2007.05.078

Benkert, Tosatto, & Schomburg. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins-Structure Function and Bioinformatics*, *71*(1), 261–277. https://doi.org/10.1002/prot.21715

Benkert, Tosatto, & Schwede. (2009). Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 173–180. https://doi.org/10.1002/prot.22532

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. *European Journal of Biochemistry*, *80*(2), 319–324. https://doi.org/10.1111/j.1432-1033.1977.tb11885.x

Best, R. B., Buchete, N.-V., & Hummer, G. (2008). Are Current Molecular Dynamics Force Fields too Helical? *Biophysical Journal*, *95*(1), L07–L09. https://doi.org/10.1529/biophysj.108.132696

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., & MacKerell, A. D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation*, *8*(9), 3257–3273. https://doi.org/10.1021/ct300400x

Bhattacharya, D., & Cheng, J. (2013a). i3Drefine Software for Protein 3D Structure Refinement and Its Assessment in CASP10. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0069648

Bhattacharya, D., & Cheng, J. (2013b). 3Drefine: Consistent Protein Structure Refinement by Optimizing Hydrogen Bonding Network and Atomic-Level Energy Minimization. *Proteins*, *81*(1), 119–131. https://doi.org/10.1002/prot.24167

Bhattacharya, D., Nowotny, J., Cao, R., & Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Research*, *44*(April), W406-409. https://doi.org/10.1093/nar/gkw336

Björkholm, P., Daniluk, P., Kryshtafovych, A., Fidelis, K., Andersson, R., & Hvidsten, T. R. (2009). Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. *Bioinformatics (Oxford, England)*, *25*(10),

1264–1270. https://doi.org/10.1093/bioinformatics/btp149

Bohr, J., Bohr, H., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B., & Petersen, S. B. (1993). Protein structures from distance inequalities. *Journal of Molecular Biology*, *231*(3), 861–869. https://doi.org/10.1006/jmbi.1993.1332

Bonneau, R., Tsai, J., Ruczinski, I., & Baker, D. (2001). Functional Inferences from Blind ab Initio Protein Structure Predictions. *Journal of Structural Biology*, *134*(2–3), 186–190. https://doi.org/10.1006/JSBI.2000.4370

Bourne, P. E., & Shindyalov, I. N. (2005). *Structure Comparison and Alignment* (pp. 321–337). John Wiley & Sons, Inc. https://doi.org/10.1002/0471721204.ch16

Bragg, W. H., Bragg Apr, W. L., H Bragg, B. W., & Professor of Physics, C. (1913). The reflection of X-rays by crystals. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *88*(605), 428–438. https://doi.org/10.1098/rspa.1913.0040

Brändén, C.-I., & Tooze, J. (1991). *Introduction to protein structure*. Garland Pub.

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, *33*(10), 3390–3400. https://doi.org/10.1093/nar/gki615

Brylinski, M., & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(1), 129–134. https://doi.org/10.1073/pnas.0707684105

Buchan, D. W. A. A., & Jones, D. T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*, *86*(Suppl Suppl 1), 78–83. https://doi.org/10.1002/prot.25379

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, *41*(W1), W349–W357. https://doi.org/10.1093/nar/gkt381

Buslje, C. M., Santos, J., Delfino, J. M., & Nielsen, M. (2009). Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, *25*(9), 1125–1131. https://doi.org/10.1093/bioinformatics/btp135

Cabra, V., & Samsó, M. (2015). Do's and don'ts of cryo-electron microscopy: A primer on sample preparation and high quality data collection for macromolecular 3D reconstruction. *Journal of Visualized Experiments*, *95*. https://doi.org/10.3791/52311

Cao, Terada, Nakamura, & Shimizu. (2003). Refinement of Comparative-Modeling Structures by Multicanonical Molecular Dynamics. *Genome Informatics*, *14*, 484–485. https://doi.org/10.11234/gi1990.14.484

Cao, Y., & Li, L. (2014). Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics*, *30*(12), 1674–1680. https://doi.org/10.1093/bioinformatics/btu104

Chen, Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., Richardson, D. C., & Richardson, D. C. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 12–21. https://doi.org/10.1107/S0907444909042073

Chen, & Brooks. (2007). Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins: Structure, Function, and Bioinformatics*, *67*(4), 922–930. https://doi.org/10.1002/prot.21345

Chen, Brooks, & Khandogin. (2008). Recent advances in implicit solvent-based methods for

biomolecular simulations. *Current Opinion in Structural Biology*, *18*(2), 140–148. https://doi.org/10.1016/J.SBI.2008.01.003

Chen, P., Huang, J. Z., & Gao, X. (2014). LigandRFs: Random forest ensemble to identify ligand-binding residues from sequence information alone. *BMC Bioinformatics*, *15*(15), S4. https://doi.org/10.1186/1471-2105-15-S15-S4

Cheng, J., & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, *8*, 113. https://doi.org/10.1186/1471-2105-8-113

Cheng, J., Choe, M., Elofsson, A., Han, K., Hou, J., Maghrabi, A. H. A., McGuffin, L. J., Menéndez-Hurtado, D., Olechnovič, K., Schwede, T., Studer, G., Uziela, K., Venclovas, Č., & Wallner, B. (2019). Estimation of model accuracy in CASP13. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1361–1377. https://doi.org/10.1002/prot.25767

Cheng, J., Wang, Z., Tegge, A. N., & Eickholt, J. (2009). Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 181–184. https://doi.org/10.1002/prot.22429

Cheng, Joung, & Lee. (2017). A Simple and Efficient Protein Structure Refinement Method. *Journal of Chemical Theory and Computation*, *13*(10), 5146–5162. https://doi.org/10.1021/acs.jctc.7b00470

Chiang, Y.-S., Gelfand, T. I., Kister, A. E., & Gelfand, I. M. (2007). New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins*, *68*(4), 915–921. https://doi.org/10.1002/prot.21473

Cho, H. J., Hyun, J. K., Kim, J. G., Jeong, H. S., Park, H. N., You, D. J., & Jung, H. S. (2013). Measurement of ice thickness on vitreous ice embedded cryo-EM grids: investigation of optimizing condition for visualizing macromolecules. *Journal of Analytical Science and Technology*, *4*(1), 1–5. https://doi.org/10.1186/2093-3371-4-7

Chopra, G., Kalisman, N., & Levitt, M. (2010). Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins: Structure, Function and Bioinformatics*, *78*(12), 2668–2678. https://doi.org/10.1002/prot.22781

Chou, P. Y., & Fasman, G. D. (1978). Empirical Predictions of Protein Conformation. *Annual Review of Biochemistry*, *47*(1), 251–276. https://doi.org/10.1146/annurev.bi.47.070178.001343

Cong, Q., Kinch, L. N., Pei, J., Shi, S., Grishin, V. N., Li, W., & Grishin, N. V. (2011). An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics*, *27*(24), 3371–3378. https://doi.org/10.1093/bioinformatics/btr572

Costa, T. R. D., Ignatiou, A., & Orlova, E. V. (2017). Structural analysis of protein complexes by cryo electron microscopy. In *Methods in Molecular Biology* (Vol. 1615, pp. 377–413). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7033-9_28

Cozzetto, D., Kryshtafovych, A., Ceriani, M., & Tramontano, A. (2007). Assessment of predictions in the model quality assessment category. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 175–183. https://doi.org/10.1002/prot.21669

Davis, I. W., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*, *32*(Web Server), W615–W619. https://doi.org/10.1093/nar/gkh398

De Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. In *Nature Reviews Genetics* (Vol. 14, Issue 4, pp. 249–261). Nature Publishing Group. https://doi.org/10.1038/nrg3414

Delano L.W. (2002). The PyMOL Molecular Graphics System. In *http://www.pymol.org*. DeLano Scientific. https://ci.nii.ac.jp/naid/10020095229

Di Lena, P., Nagata, K., & Baldi, P. (2012). Deep architectures for protein contact map prediction. *Bioinformatics (Oxford, England)*, *28*(19), 2449–2457. https://doi.org/10.1093/bioinformatics/bts475

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., & Voelz, V. A. (2007). The protein folding problem: when will it be solved? In *Current Opinion in Structural Biology* (Vol. 17, Issue 3, pp. 342–346). Curr Opin Struct Biol. https://doi.org/10.1016/j.sbi.2007.06.001

DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., & Baker, D. (2009). Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. *Journal of Molecular Biology*, *392*(1), 181–190. https://doi.org/10.1016/j.jmb.2009.07.008

Dorn, Silva, E., Buriol, & Lamb. (2014). Three-dimensional protein structure prediction:Methods and computational strategies. *Computational Biology and Chemistry*, *53*(PB), 251–276. https://doi.org/10.1016/j.compbiolchem.2014.10.001

Drenth, J. (1999). *Principles of protein X-ray crystallography*. Springer. https://books.google.co.uk/books/about/Principles_of_Protein_X_ray_Crystallogra.html?id=ABjCdPuly4IC

Duarte, J. M., Sathyapriya, R., Stehr, H., Filippis, I., & Lappe, M. (2010). Optimal contact definition for reconstruction of Contact Maps. *BMC Bioinformatics*, *11*, 283. https://doi.org/10.1186/1471-2105-11-283

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C. H., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., & Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8

Eastman, P., Friedrichs, M. S., Chodera, J. D., Radmer, R. J., Bruns, C. M., Ku, J. P., Beauchamp, K. A., Lane, T. J., Wang, L.-P., Shukla, D., Tye, T., Houston, M., Stich, T., Klein, C., Shirts, M. R., & Pande, V. S. (2013). OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*, *9*(1), 461–469. https://doi.org/10.1021/ct300857j

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, *13*(7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Eickholt, J., & Cheng, J. (2013). A study and benchmark of DNcon: A method for protein residue-residue contact prediction using deep networks. *BMC Bioinformatics*, *14*(SUPPL.14), S12. https://doi.org/10.1186/1471-2105-14-S14-S12

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods in Enzymology*, *277*, 396–404. http://www.ncbi.nlm.nih.gov/pubmed/9379925

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(1). https://doi.org/10.1103/PhysRevE.87.012707

Ekins, S., Mestres, J., & Testa, B. (2007). *In silico* pharmacology for drug discovery: applications to targets and beyond. *British Journal of Pharmacology*, *152*(1), 21–37. https://doi.org/10.1038/sj.bjp.0707306

Erdin, S., Ward, R. M., Venner, E., & Lichtarge, O. (2010). Evolutionary Trace Annotation of Protein Function in the Structural Proteome. *Journal of Molecular Biology*, *396*(5), 1451–1473. https://doi.org/10.1016/j.jmb.2009.12.037

Ezkurdia, L., Grana, O., Izarzugaza, J. M. G., & Tress, M. L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. In *Proteins: Structure, Function and Bioinformatics* (Vol. 77, Issue SUPPL. 9, pp. 196–209). https://doi.org/10.1002/prot.22554

Fan, H., & Mark, A. E. (2004). Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science*, *13*(1), 211–220. https://doi.org/10.1110/ps.03381404

Fariselli, P., Olmea, O., Valencia, A., & Casadio, R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins: Structure, Function and Genetics*, *45*(SUPPL. 5), 157–162. https://doi.org/10.1002/prot.1173

Feig, M. (2016). Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *Journal of Chemical Information and Modeling*, *56*(7), 1304–1312. https://doi.org/10.1021/acs.jcim.6b00222

Feig, M. (2017). Computational protein structure refinement: almost there, yet still so far to go. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *7*(3), e1307. https://doi.org/10.1002/wcms.1307

Feig, M., & Mirjalili, V. (2016). Protein Structure Refinement via Molecular-Dynamics Simulations: What works and what does not? *Proteins*, *84*(1), 282–292.

Fischer, Barret, Bryson, Elofsson, Godzik, Jones, Karplus, Kelley, Maccallum, Pawowski, Rost, Rychlewski, & Sternberg. (1999). CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function and Genetics*, *37*(SUPPL. 3), 209–217. https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<209::AID-PROT27>3.0.CO;2-Y

Fletterick, R. J. (1992). Introduction to protein structure, by Carl Branden and John Tooze. New York: Garland Publishing Company, 302 pages, $27.95 (paper), 1991. *Proteins: Structure, Function, and Genetics*, *12*(2), 200–200. https://doi.org/10.1002/prot.340120213

Fuller, J. C., Martinez, M., Henrich, S., Stank, A., Richter, S., & Wade, R. C. (2015). LigDig: a web server for querying ligand-protein interactions. *Bioinformatics (Oxford, England)*, *31*(7), 1147—1149. https://doi.org/10.1093/bioinformatics/btu784

Gallo Cassarino, Bordoli, L., & Schwede, T. (2014). *Assessment of ligand binding site predictions in CASP10*. *82*(SUPPL.2), 154–163. https://doi.org/10.1002/prot.24495

Gil, N., & Fiser, A. (2019). The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. *Bioinformatics*, *35*(1), 12–19. https://doi.org/10.1093/bioinformatics/bty523

Ginalski, K., Elofsson, A., Fischer, D., & Rychlewski, L. (2003). *No Title*. *19*(8). https://doi.org/10.1093/bioinformatics/btg124

Giorgetti, A., Raimondo, D., Miele, A. E., & Tramontano, A. (2005). Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics*, *21*(Suppl 2), 72–76. https://doi.org/10.1093/bioinformatics/bti1112

Göbel, U., Sander, C., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue

contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, *18*(4), 309–317. https://doi.org/10.1002/prot.340180402

Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., & Walker, R. C. (2012). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. *Journal of Chemical Theory and Computation*, *8*(5), 1542–1555. https://doi.org/10.1021/ct200909j

Graña, O., Baker, D., MacCallum, R. M., Meiler, J., Punta, M., Rost, B., Tress, M. L., & Valencia, A. (2005). CASP6 assessment of contact prediction. *Proteins: Structure, Function and Genetics*, *61*(SUPPL. 7), 214–224. https://doi.org/10.1002/prot.20739

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, *331*(1), 281–299. https://doi.org/10.1016/S0022-2836(03)00670-3

Greener, J. G., Kandathil, S. M., & Jones, D. T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-11994-0

Greenfield, N. J. (2006). Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, *1*(6), 2876–2890. https://doi.org/10.1038/nprot.2006.202

Gromiha, & Selvaraj. (2004). Inter-residue interactions in protein folding and stability. In *Progress in Biophysics and Molecular Biology* (Vol. 86, Issue 2, pp. 235–277). https://doi.org/10.1016/j.pbiomolbio.2003.09.003

Gront, D., Kmiecik, S., Blaszczyk, M., Ekonomiuk, D., & Koliński, A. (2012). Optimization of protein models. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *2*(3), 479–493. https://doi.org/10.1002/wcms.1090

Guo, J. T., Ellrott, K., & Xu, Y. (2008). A Historical Perspective of Template-Based Protein Structure Prediction. In *Protein Structure Prediction* (Vol. 413, pp. 3–42). Humana Press. https://doi.org/10.1007/978-1-59745-574-9_1

Haas, Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., & Schwede, T. (2018). *Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12*. *86*, 387–398. https://doi.org/10.1002/prot.25431

Haas, Gumienny, R., Barbato, A., Ackermann, F., Tauriello, G., Bertoni, M., Studer, G., Smolinski, A., & Schwede, T. (2019). Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1378–1387. https://doi.org/10.1002/prot.25815

Haas, Roth, Arnold, Kiefer, Schmidt, Bordoli, Schwede, Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., & Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, *2013*(8), 1–8. https://doi.org/10.1093/database/bat031

Han, R., Leo-Macias, A., Zerbino, D., Bastolla, U., Contreras-Moreira, B., & Ortiz, A. R. (2008). An efficient conformational sampling method for homology modeling. *Proteins: Structure, Function, and Bioinformatics*, *71*(1), 175–188. https://doi.org/10.1002/prot.21672

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, *34*(23), 4039–4045. https://doi.org/10.1093/bioinformatics/bty481

He, B., Mortuza, S. M., Wang, Y., Shen, H. Bin, & Zhang, Y. (2017). NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics*, *33*(15), 2296–2306. https://doi.org/10.1093/bioinformatics/btx164

Heinemann, U., Frevert, J., Hofman, K.-P., Illing, G., Oschkinat, H., Saenger, W., & Zettl, R. (2002). Linking Structural Biology With Genome Research. In *Genomics and Proteomics* (pp. 179–189). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-46823-9_15

Heo, L., & Feig, M. (2018a). PREFMD: a web server for protein structure refinement via molecular dynamics simulations. *Bioinformatics*, *34*(6), 1063–1065. https://doi.org/10.1093/bioinformatics/btx726

Heo, L., & Feig, M. (2018b). Experimental accuracy in protein structure refinement via molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(52), 13276–13281. https://doi.org/10.1073/pnas.1811364115

Heo, L., Park, H., & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Research*, *41*(Web Server issue), 384–388. https://doi.org/10.1093/nar/gkt458

Heo, L., Shin, W.-H., Lee, M. S., & Seok, C. (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Research*, *42*(Web Server issue), W210-4. https://doi.org/10.1093/nar/gku321

Hernandez, M., Ghersi, D., & Sanchez, R. (2009). SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Research*, *37*(Web Server issue), W413-6. https://doi.org/10.1093/nar/gkp281

Hooft, R. W. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature*, *381*(6580), 272–272. https://doi.org/10.1038/381272a0

Hovan, L., Oleinikovas, V., Yalinca, H., Kryshtafovych, A., Saladino, G., & Gervasio, F. L. (2018). Assessment of the model refinement category in CASP12. *Proteins*, *86*, 152–167. https://doi.org/10.1002/prot.25409

Huang, B., & Schroeder, M. (2006). LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology*, *6*(1), 19. https://doi.org/10.1186/1472-6807-6-19

Huang, Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., & MacKerell, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, *14*(1), 71–73. https://doi.org/10.1038/nmeth.4067

Hughes, K. C., Gao, X., Kim, I. Y., Wang, M., Weisskopf, M. G., Schwarzschild, M. A., & Ascherio, A. (2017). Intake of dairy foods and risk of Parkinson disease. *Neurology*, *89*(1), 46–52. https://doi.org/10.1212/WNL.0000000000004057

Hurtado, D. M., Uziela, K., & Elofsson, A. (2018). *Deep transfer learning in the assessment of the quality of protein models*. http://arxiv.org/abs/1804.06281

Ishitani, R., Terada, T., & Shimizu, K. (2008). Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations. *Molecular Simulation*, *34*(3), 327–336. https://doi.org/10.1080/08927020801930539

Izidoro, S. C., de Melo-Minardi, R. C., & Pappa, G. L. (2015). GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics (Oxford, England)*, *31*(6), 864–870. https://doi.org/10.1093/bioinformatics/btu746

Jagielska, A., Wroblewska, L., & Skolnick, J. (2008). Protein model refinement using an optimized physics-based all-atom force field. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(24), 8268–8273. https://doi.org/10.1073/pnas.0800054105

Jeong, C.-S., & Kim, D. (2012). Reliable and robust detection of coevolving protein residues. *Protein Engineering, Design and Selection*, *25*(11), 705–713. https://doi.org/10.1093/protein/gzs081

Jöbstl, E., Howse, J. R., Fairclough, J. P. A., & Williamson, M. P. (2006). Noncovalent cross-linking of casein by epigallocatechin gallate characterized by single molecule force microscopy. *Journal of Agricultural and Food Chemistry*, *54*(12), 4077–4081. https://doi.org/10.1021/jf053259f

Jones. (2000). A practical guide to protein structure prediction. *Methods in Molecular Biology (Clifton, N.J.)*, *143*, 131–154. https://doi.org/10.1385/1-59259-368-2:131

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*(2), 195–202. https://doi.org/10.1006/jmbi.1999.3091

Jones, D. T. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function and Genetics*, *45*(S5), 127–132. https://doi.org/10.1002/prot.1171

Jones, D. T., Buchan, D. W. A., Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, *28*(2), 184–190. https://doi.org/10.1093/bioinformatics/btr638

Jones, D. T., & Cozzetto, D. (2015). DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, *31*(6), 857–863. https://doi.org/10.1093/bioinformatics/btu744

Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics (Oxford, England)*, *34*(19), 3308–3315. https://doi.org/10.1093/bioinformatics/bty341

Jones, D. T., Singh, T., Kosciolek, T., & Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, *31*(7), 999–1006. https://doi.org/10.1093/bioinformatics/btu791

Jonic, S., & Vénien-Bryan, C. (2009). Protein structure determination by electron cryo-microscopy. *Current Opinion in Pharmacology*, *9*(5), 636–642. https://doi.org/10.1016/J.COPH.2009.04.006

Joo, K., Lee, J., Lee, S., Seo, J.-H., Lee, S. J., & Lee, J. (2007). High accuracy template based modeling by global optimization. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 83–89. https://doi.org/10.1002/prot.21628

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*(2), 926. https://doi.org/10.1063/1.445869

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S., & Rost, B. (2014). FreeContact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, *15*(1), 85. https://doi.org/10.1186/1471-2105-15-85

Kalisman, N., Levi, A., Maximova, T., Reshef, D., Zafriri-Lynn, S., Gleyzer, Y., & Keasar, C. (2005). MESHI: A new library of Java classes for molecular modeling. *Bioinformatics*, *21*(20), 3931–3932. https://doi.org/10.1093/bioinformatics/bti630

Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based

residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15674–15679. https://doi.org/10.1073/pnas.1314045110

Kandathil, S. M., Greener, J. G., & Jones, D. T. (2019a). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins*, *87*(12), 1092–1099. https://doi.org/10.1002/prot.25779

Kandathil, S. M., Greener, J. G., & Jones, D. T. (2019b). Recent developments in deep learning applied to protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1179–1189. https://doi.org/10.1002/prot.25824

Kannan, S., & Zacharias, M. (2010). Application of biasing-potential replica-exchange simulations for loop modeling and refinement of proteins in explicit solvent. *Proteins: Structure, Function, and Bioinformatics*, *78*(13), 2809–2819. https://doi.org/10.1002/prot.22796

Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, *37*(Web Server), W492–W497. https://doi.org/10.1093/nar/gkp403

Kastritis, P. L., Rodrigues, J. P. G. L. M., & Bonvin, A. M. J. J. (2014). HADDOCK2P2I: A biophysical model for predicting the binding affinity of protein-protein interaction inhibitors. *Journal of Chemical Information and Modeling*, *54*(3), 826–836. https://doi.org/10.1021/ci4005332

Khoury, G. A., Smadbeck, J., Kieslich, C. A., Koskosidis, A. J., Guzman, Y. A., Tamamis, P., & Floudas, C. A. (2017). Princeton_TIGRESS 2.0: High refinement consistency and net gains through support vector machines and molecular dynamics in double-blind predictions during the CASP11 experiment. *Proteins: Structure, Function, and Bioinformatics*, *85*(6), 1078–1098. https://doi.org/10.1002/prot.25274

Khoury, G. A., Tamamis, P., Pinnaduwage, N., Smadbeck, J., Kieslich, C. A., & Floudas, C. A. (2014). Princeton_TIGRESS: Protein geometry refinement using simulations and support vector machines. *Proteins: Structure, Function and Bioinformatics*, *82*(5), 794–814. https://doi.org/10.1002/prot.24459

Kim, D. E., Blum, B., Bradley, P., & Baker, D. (2009). Sampling Bottlenecks in De novo Protein Structure Prediction. *Journal of Molecular Biology*, *393*(1), 249–260. https://doi.org/10.1016/J.JMB.2009.07.063

Kliger, Y., Levy, O., Oren, A., Ashkenazy, H., Tiran, Z., Novik, A., Rosenberg, A., Amir, A., Wool, A., Toporik, A., Schreiber, E., Eshel, D., Levine, Z., Cohen, Y., Nold-Petry, C., Dinarello, C. A., & Borukhov, I. (2009). Peptides modulating conformational changes in secreted chaperones: From in silico design to preclinical proof of concept. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13797–13801. https://doi.org/10.1073/pnas.0906514106

Kosciolek, T., & Jones, D. T. (2016). Accurate contact predictions using covariation techniques and machine learning. *Proteins*, *84*, 145–151. https://doi.org/10.1002/prot.24863

Kryshtafovych, A., Barbato, A., Fidelis, K., Monastyrskyy, B., Schwede, T., & Tramontano, A. (2014). Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins*, *82 Suppl 2*(0 2), 112–126. https://doi.org/10.1002/prot.24347

Kryshtafovych, A., Fidelis, K., & Tramontano, A. (2011). Evaluation of model quality predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 91–106. https://doi.org/10.1002/prot.23180

Kryshtafovych, A., Monastyrskyy, B., & Fidelis, K. (2016). CASP11 statistics and the prediction center evaluation system. *Proteins: Structure, Function, and Bioinformatics*, *84*, 15–19.

https://doi.org/10.1002/prot.25005
Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., & Tramontano, A. (2017). Assessment of model accuracy estimations in CASP12. *Proteins: Structure, Function, and Bioinformatics*. https://doi.org/10.1002/prot.25371
Kryshtafovych, A., Venclovas, Č., Fidelis, K., & Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins: Structure, Function, and Bioinformatics*, *61*(S7), 225–236. https://doi.org/10.1002/prot.20740
Kuhlman, & Baker. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(19), 10383–10388. https://doi.org/10.1073/PNAS.97.19.10383
Kuhlman, Dantas, Ireton, Varani, Stoddard, & Baker. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)*, *302*(5649), 1364–1368. https://doi.org/10.1126/science.1089427
Kumar, A., Campitelli, P., Thorpe, M. F., & Ozkan, S. B. (2015). Partial unfolding and refolding for structure refinement: A unified approach of geometric simulations and molecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, *83*(12), 2279–2292. https://doi.org/10.1002/prot.24947
Lapedes, A. S., Giraud, B., Liu, L., & Stormo, G. D. (1999). *Correlated mutations in models of protein sequences: phylogenetic and structural effects*. 236–256. https://doi.org/10.1214/LNMS/1215455556
Larsen, A. B., Wagner, J. R., Jain, A., & Vaidehi, N. (2014). Protein Structure Refinement of CASP Target Proteins Using GNEIMO Torsional Dynamics Method. *Journal of Chemical Information and Modeling*, *54*(2), 508–517. https://doi.org/10.1021/ci400484c
Larsson, P., Skwark, M. J., Wallner, B., & Elofsson, A. (2009). Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 167–172. https://doi.org/10.1002/prot.22476
Laskowski, Macarthur, M., Moss, D., & Thornton, J. (1993). PROCHECK : a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, *26*. http://www.citeulike.org/user/pkolb/article/1720734
Laskowski, R. A., Watson, J. D., & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research*, *33*(Web Server), 89–93. https://doi.org/10.1093/nar/gki414
Laskowski, Rullmannn, MacArthur, Kaptein, & Thornton. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, *8*(4), 477–486. https://doi.org/10.1007/BF00228148
Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D., & Kuhlman, B. (2013). Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods in Enzymology*, *523*, 109–143. https://doi.org/10.1016/B978-0-12-394292-0.00006-0
Lee, Heo, & Seok. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins: Structure, Function, and Bioinformatics*, *84*, 293–301. https://doi.org/10.1002/prot.24858
Lee, J., Wu, S., & Zhang, Y. (2009). Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics* (pp. 3–25). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9058-5_1

Lee, Tsai, Baker, & Kollman. (2001). Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology*, *313*(2), 417–430. https://doi.org/10.1006/JMBI.2001.5032

Lesk. (1997). CASP2: Report on ab initio predictions. *Proteins: Structure, Function and Genetics*, *29*, 151–166. https://doi.org/10.1002/(SICI)1097-0134(1997)1+<151::AID-PROT20>3.0.CO;2-M

Lesk, Conte, & Hubbard. (2001). Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins: Structure, Function and Genetics*, *45*, 98–118. https://doi.org/10.1002/prot.10056

Levitt, M., & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(11), 5913–5920. https://doi.org/10.1073/pnas.95.11.5913

Li, Y., Fang, Y., & Fang, J. (2011). Predicting residue-residue contacts using random forest models. *Bioinformatics (Oxford, England)*, *27*(24), 3379–3384. https://doi.org/10.1093/bioinformatics/btr579

Lin, M. S., & Head-Gordon, T. (2011). Reliable protein structure refinement using a physical energy function. *Journal of Computational Chemistry*, *32*(4), 709–717. https://doi.org/10.1002/jcc.21664

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS ONE*, *7*(2), 32131. https://doi.org/10.1371/journal.pone.0032131

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science*, *334*(6055), 517–520. https://doi.org/10.1126/science.1208351

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., & Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, *78*(8), 1950–1958. https://doi.org/10.1002/prot.22711

Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems*, *6*(1), 65-74.e3. https://doi.org/10.1016/j.cels.2017.11.014

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Research*, *28*(1), 257–259. http://www.ncbi.nlm.nih.gov/pubmed/10592240

Loncharich, R. J., Brooks, B. R., & Pastor, R. W. (1992). Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. *Biopolymers*, *32*(5), 523–535. https://doi.org/10.1002/bip.360320508

López, Ezkurdia, I., & Tress, M. L. (2009). Assessment of ligand binding residue predictions in CASP8. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 138–146. https://doi.org/10.1002/prot.22557

López, G., Valencia, A., & Tress, M. L. (2007). firestar--prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Research*, *35*(Web Server issue), W573-7. https://doi.org/10.1093/nar/gkm297

Lu, H., & Skolnick, J. (2003). Application of statistical potentials to protein structure refinement from low resolutionab initio models. *Biopolymers*, *70*(4), 575–584. https://doi.org/10.1002/bip.10537

Lundström, J., Rychlewski, L., Bujnicki, J., & Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Science : A Publication of the*

Protein Society, *10*(11), 2354–2362. https://doi.org/10.1101/ps.08501.are

MacCallum, J. L., Hua, L., Schnieders, M. J., Pande, V. S., Jacobson, M. P., & Dill, K. A. (2009). Assessment of the protein-structure refinement category in CASP8. *Proteins*, *77*(S9), 66–80. https://doi.org/10.1002/prot.22538

MacCallum, J. L., Pérez, A., Schnieders, M. J., Hua, L., Jacobson, M. P., & Dill, K. A. (2011). Assessment of protein structure refinement in CASP9. *Proteins*, *79*(S10), 74–90. https://doi.org/10.1002/prot.23131

MacKerell, A. D., Banavali, N., & Foloppe, N. (2001). Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, *56*(4), 257–265. https://doi.org/10.1002/1097-0282(2000)56:4<257::AID-BIP10029>3.0.CO;2-W

MacKerell, A. D., Feig, M., & Brooks, C. L. (2004). Extending the treatment of backbone energetics in protein force fields. *J. Comp. Chem.*, *25*(11), 1400–1415. https://doi.org/Doi 10.1002/Jcc.20065

Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E., & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology*, *316*(1), 139–154. https://doi.org/10.1006/jmbi.2001.5327

Maghrabi, A. H. A., & McGuffin, L. J. (2019). Estimating the quality of 3D protein models using the ModFOLD7 server. *Methods in Molecular Biology*, *2165*, 69–81. https://doi.org/10.1007/978-1-0716-0708-4_4

Maghrabi, & McGuffin, L. J. (2017). ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Research*, *45*(1), W416-421. https://doi.org/10.1093/nar/gkx332

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

Mandle, Jain, & Shrivastava. (2012). Protein structure prediction using support vector machine. *International Journal on Soft Computing*, *3*(1), 67–78. https://doi.org/0.5121/ijsc.2012.3106

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*, *6*(12), e28766. https://doi.org/10.1371/journal.pone.0028766

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

McGuffin. (2007). Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics*, *8*(August), 345. https://doi.org/10.1186/1471-2105-8-345

McGuffin. (2008a). Aligning Sequences to Structures. In *Protein Structure Prediction* (pp. 61–90). Humana Press. https://doi.org/10.1007/978-1-59745-574-9_3

McGuffin. (2008b). Protein Fold Recognition and Threading. In *Computational Structural Biology* (pp. 37–60). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812778789_0002

McGuffin. (2009). Prediction of global and local model quality in CASP8 using the ModFOLD server. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 185–190.

https://doi.org/10.1002/prot.22491
McGuffin. (2010). Model Quality Prediction. In *Introduction to Protein Structure Prediction* (pp. 323–342). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470882207.ch15
McGuffin, Adiyaman, Maghrabi, Shuid, Brackenridge, Nealon, & Philomina. (2019). IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Research*, *47*(W1), 408–413. https://doi.org/10.1093/nar/gkz322
McGuffin, Atkins, J. D., 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*(W1), W169-73. https://doi.org/10.1093/nar/gkv236
McGuffin, Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)*, *16*(4), 404–405. https://doi.org/10.1093/bioinformatics/16.4.404
McGuffin, Buenavista, M. T., & Roche, D. B. (2013). The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Research*, *41*(Web Server issue), W368-372. https://doi.org/10.1093/nar/gkt294
McGuffin, L. J. (2008c). The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics*, *24*(4), 586–587. https://doi.org/10.1093/bioinformatics/btn014
McGuffin, & Roche, D. B. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, *26*(2), 182–188. https://doi.org/10.1093/bioinformatics/btp629
McGuffin, & Roche, D. B. (2011). Automated tertiary structure prediction with accurate local model quality assessment using the intfold-ts method. *Proteins: Structure, Function and Bioinformatics*, *79*(SUPPL. 10), 137–146. https://doi.org/10.1002/prot.23120
Meiler, J., & Baker, D. (2003). Rapid protein fold determination using unassigned NMR data. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15404–15409. https://doi.org/10.1073/pnas.2434121100
Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., & Elofsson, A. (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics (Oxford, England)*, *30*(17), i482-8. https://doi.org/10.1093/bioinformatics/btu458
Miller, C. S., & Eisenberg, D. (2008). Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics (Oxford, England)*, *24*(14), 1575–1582. https://doi.org/10.1093/bioinformatics/btn248
Mirjalili, V., & Feig, M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. *Journal of Chemical Theory and Computation*, *9*(2), 1294–1303. https://doi.org/10.1021/ct300962x
Mirjalili, V., Noyes, K., & Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins*, *82*(SUPPL.2), 196–207. https://doi.org/10.1002/prot.24336
Mirny, L., & Domany, E. (1996). Protein fold recognition and dynamics in the space of contact maps. *Proteins: Structure, Function, and Bioinformatics*, *26*(4), 391–410. https://doi.org/10.1002/(SICI)1097-0134(199612)26:4<391::AID-PROT3>3.0.CO;2-F
Misura, K. M. S. S., & Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins: Structure, Function and Genetics*, *59*(1), 15–29. https://doi.org/10.1002/prot.20376
Modi, V., & Dunbrack, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins*, *April*, 260–281. https://doi.org/10.1002/prot.25048

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2014). Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, *82*(SUPPL.2), 138–153. https://doi.org/10.1002/prot.24340

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, *15*(3), 285–289. https://doi.org/10.1016/j.sbi.2005.05.011

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., & Tramontano, A. (2007). Critical assessment of methods of protein structure prediction—Round VII. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 3–9. https://doi.org/10.1002/prot.21767

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., & Tramontano, A. (2009). Critical assessment of methods of protein structure prediction-Round VIII. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 1–4. https://doi.org/10.1002/prot.22589

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function, and Bioinformatics*, *82*(S2), 1–6. https://doi.org/10.1002/prot.24452

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, *84*(S1), 4–14. https://doi.org/10.1002/prot.25064

Moult, J., Fidelis, K., Rost, B., Hubbard, T., & Tramontano, A. (2005). Critical assessment of methods of protein structure prediction (CASP)—Round 6. *Proteins: Structure, Function, and Bioinformatics*, *61*(S7), 3–7. https://doi.org/10.1002/prot.20716

Murata, K., & Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1862*(2), 324–334. https://doi.org/10.1016/J.BBAGEN.2017.07.020

Niggemann, M., & Steipe, B. (2000). Exploring local and non-local interactions for protein stability by structural motif engineering. *Journal of Molecular Biology*, *296*(1), 181–195. https://doi.org/10.1006/jmbi.1999.3385

Nugent, T., Cozzetto, D., & Jones, D. T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function and Bioinformatics*, *82*(SUPPL.2), 98–111. https://doi.org/10.1002/prot.24377

Olechnovič, K., Kulberkytė, E., Venclovas, Č., Kulberkyte, E., & Venclovas, Č. (2013). *CAD-score: A new contact area difference-based function for evaluation of protein structural models*. *81*(1), 149–162. https://doi.org/10.1002/prot.24172

Olechnovič, K., & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins: Structure, Function and Bioinformatics*, *85*(6), 1131–1145. https://doi.org/10.1002/prot.25278

Olmea, O., & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding and Design*, *2*(3), S25–S32. https://doi.org/10.1016/S1359-0278(97)00060-6

Olson, M. A., & Lee, M. S. (2014). Evaluation of Unrestrained Replica-Exchange Simulations Using Dynamic Walkers in Temperature Space for Protein Structure Refinement. *PLoS ONE*, *9*(5), e96638. https://doi.org/10.1371/journal.pone.0096638

Oren, Dale, Yael, Dongli, Sharon, Srinivasa, Alexander, Pradyumna, Merav, Anurag, Raphael, Michael, & Noiman. (2006). *An Integrated in Silico 3D Model-Driven Discovery of a Novel, Potent, and Selective Amidosulfonamide 5-HT1A Agonist (PRX-00023) for the Treatment of*

Anxiety and Depression. https://doi.org/10.1021/JM0508641

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., & Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, *Suppl 3*, 149–170. https://doi.org/10.1002/(sici)1097-0134(1999)37:3+<149::aid-prot20>3.3.co;2-8

Ovchinnikov, S., Kim, D. E., Wang, R. Y.-R., Liu, Y., DiMaio, F., & Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins: Structure, Function, and Bioinformatics*, *84*, 67–75. https://doi.org/10.1002/prot.24974

Ovchinnikov, S., Park, H., Kim, D. E., DiMaio, F., & Baker, D. (2018). *Protein structure prediction using Rosetta in CASP12*. *86*(Suppl 1), 113–121. https://doi.org/10.1002/prot.25390

Ovchinnikov, S., Park, H., Varghese, N., Huang, P. S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, *355*(6322), 294–298. https://doi.org/10.1126/science.aah4043

Pande, V. S., Grosberg, A. Y., Tanaka, T., & Rokhsar, D. S. (1998). Pathways for protein folding: Is a new view needed? *Current Opinion in Structural Biology*, *8*(1), 68–79. https://doi.org/10.1016/S0959-440X(98)80012-2

Park, Gangupomu, Wagner, Jain, & Vaidehi. (2012). Structure Refinement of Protein Low Resolution Models Using the GNEIMO Constrained Dynamics Method. *The Journal of Physical Chemistry B*, *116*(8), 2365–2375. https://doi.org/10.1021/jp209657n

Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., & DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, *12*(12), 6201–6212. https://doi.org/10.1021/acs.jctc.6b00819

Park, H., DiMaio, F., & Baker, D. (2015). The Origin of Consistent Protein Structure Refinement from Structural Averaging. *Structure*, *23*(6), 1123–1128. https://doi.org/10.1016/J.STR.2015.03.022

Pauling, Corey, & Branson. (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), 205–211. https://doi.org/10.1073/PNAS.37.4.205

Pavlopoulou, A., & Michalopoulos, I. (2011). State-of-the-art bioinformatics protein structure prediction tools (Review). *International Journal of Molecular Medicine*, *28*(3), 295–310. https://doi.org/10.3892/ijmm.2011.705

Pawlowski, M., Gajda, M. J., Matlak, R., & Bujnicki, J. M. (2008). MetaMQAP: A meta-server for the quality assessment of protein models. *BMC Bioinformatics*, *9*(1), 403. https://doi.org/10.1186/1471-2105-9-403

Pelton, J. T., & McLean, L. R. (2000). Spectroscopic Methods for Analysis of Protein Secondary Structure. *Analytical Biochemistry*, *277*(2), 167–176. https://doi.org/10.1006/abio.1999.4320

Peng, J., & Xu, J. (2009). Boosting Protein Threading Accuracy. *Research in Computational Molecular Biology : ... Annual International Conference, RECOMB ... : Proceedings. RECOMB (Conference : 2005-)*, *5541*, 31–45. https://doi.org/10.1007/978-3-642-02008-7_3

Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. New Science Press.

Pettitt, C. S., McGuffin, L. J., & Jones, D. T. (2005). Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics*, *21*(17), 3509–3515. https://doi.org/10.1093/bioinformatics/bti540

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16), 1781–1802. https://doi.org/10.1002/jcc.20289

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., & Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, *30*(12), 1771–1773. https://doi.org/10.1093/bioinformatics/btu097

Pollastri, G., & McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, *21*(8), 1719–1720. https://doi.org/10.1093/bioinformatics/bti203

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J., & Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, *450*(7167), 259–264. https://doi.org/10.1038/nature06249

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*, 95–99. http://www.ncbi.nlm.nih.gov/pubmed/13990617

Randall, A., & Baldi, P. (2008). SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Structural Biology*, *8*(1), 52. https://doi.org/10.1186/1472-6807-8-52

Rangwala, H., & Karypis, G. (George). (2010). *Introduction to protein structure prediction : methods and algorithms*. Wiley.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics*, *80*(8), 2071–2079. https://doi.org/10.1002/prot.24098

Read, R. J., Sammito, M. D., Kryshtafovych, A., & Croll, T. I. (2019). Evaluation of model refinement in CASP13. *Proteins*, *87*(12), 1249–1262. https://doi.org/10.1002/prot.25794

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Research Computing Documentation contributors. (2018). *Amber 12*. https://wiki.rc.usf.edu/index.php?title=Amber12&oldid=1808

Rhizobium, G. E. (2013). Complete Genome Sequence of the Sesbania Symbiont and Rice. *Nucleic Acids Research*, *1*(1256879), 13–14. https://doi.org/10.1093/nar

Roberts, V. A., Thompson, E. E., Pique, M. E., Perez, M. S., & Ten Eyck, L. F. (2013). DOT2: Macromolecular docking with improved biophysical models. *Journal of Computational Chemistry*, *34*(20), 1743–1758. https://doi.org/10.1002/jcc.23304

Robertson, M. J., Tirado-Rives, J., & Jorgensen, W. L. (2015). Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation*, *11*(7), 3499–3509. https://doi.org/10.1021/acs.jctc.5b00356

Roche, Brackenridge, D. A., & 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), 29829–29842. https://doi.org/10.3390/ijms161226202

Roche, Buenavista, M., & McGuffin, L. (2013a). Predicting protein structures and structural annotation of proteomes. *Encylopedia of Biophysics*, 2061–2068. https://doi.org/10.1007/978-3-642-16712-6_418

Roche, Buenavista, M. T., & McGuffin, L. J. (2012a). *FunFOLDQA : A Quality Assessment Tool*

for Protein-Ligand Binding Site Residue Predictions FunFOLDQA : A Quality Assessment Tool for Protein- Ligand Binding Site Residue Predictions. *August 2016*. https://doi.org/10.1371/journal.pone.0038219

Roche, Buenavista, M. T., & McGuffin, L. J. (2013b). The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Research*, *41*(Web Server issue), 303–307. https://doi.org/10.1093/nar/gkt498

Roche, Buenavista, M. T., Tetchner, S. J., & 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*(SUPPL. 2), 171–176. https://doi.org/10.1093/nar/gkr184

Roche, Buenavista, & McGuffin. (2014). *Assessing the Quality of Modelled 3D Protein Structures Using the ModFOLD Server* (pp. 83–103). https://doi.org/10.1007/978-1-4939-0366-5_7

Roche, D. B., Buenavista, M. T., & McGuffin, L. J. (2012b). Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*, *28*(14), 1851–1857. https://doi.org/10.1093/bioinformatics/bts292

Roche, & McGuffin. (2016a). In silico identifi cation and characterization of protein-ligand binding sites. In *Methods in Molecular Biology* (Vol. 1414, pp. 1–21). Humana Press Inc. https://doi.org/10.1007/978-1-4939-3569-7_1

Roche, & McGuffin, L. (2016b). Toolbox for Protein Structure Prediction. In Barry L. Stoddard (Ed.), *Computational Design of Ligand Binding Proteins Binding Proteins* (Vol. 7, Issue 5, pp. 363–377). Humana Press. https://doi.org/10.1007/978-1-4939-3145-3_23

Roche, Tetchner, S. J., & McGuffin, L. J. (2010). The binding site distance test score: A robust method for the assessment of predicted protein binding sites. *Bioinformatics*, *26*(22), 2920–2921. https://doi.org/10.1093/bioinformatics/btq543

Roche, 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*(1), 160. https://doi.org/10.1186/1471-2105-12-160

Rodrigues, J. P. G. L. M., Levitt, M., & Chopra, G. (2012). KoBaMIN: A knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Research*, *40*(W1), 323–328. https://doi.org/10.1093/nar/gks376

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, *383*, 66–93. https://doi.org/10.1016/S0076-6879(04)83004-0

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725–738. https://doi.org/10.1038/nprot.2010.5

Rykunov, D., & Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, *11*(1), 128. https://doi.org/10.1186/1471-2105-11-128

Sanger, F. (1959). Chemistry of Insulin. *Science*, *129*(3359).

Sanger, F., & Tuppy, H. (1951). The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 1. *Biochemical J*, *49*(4), 463–481. https://doi.org/10.1016/0006-3002(53)90071-7

Sankararaman, S., Kolaczkowski, B., & Sjölander, K. (2009). INTREPID: A web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Research*, *37*(SUPPL. 2), 390–395. https://doi.org/10.1093/nar/gkp339

Sankararaman, S., Sha, F., Kirsch, J. F., Jordan, M. I., & Sjölander, K. (2010). Active site

prediction using evolutionary and structural information. *Bioinformatics*, *26*(5), 617–624. https://doi.org/10.1093/bioinformatics/btq008

Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., & Bonvin, A. M. J. J. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function and Bioinformatics*, *86*(October 2017), 51–66. https://doi.org/10.1002/prot.25407

Schmidt, T., Haas, J., Cassarino, T. G., & Schwede, T. (2011). Assessment of ligand-binding residue predictions in CASP9. *Proteins: Structure, Function and Bioinformatics*, *79*(SUPPL. 10), 126–136. https://doi.org/10.1002/prot.23174

Schmidt, & Urlaub. (2017). Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. In *Current Opinion in Structural Biology* (Vol. 46, pp. 157–168). Elsevier Ltd. https://doi.org/10.1016/j.sbi.2017.10.005

Schneider, M., & Brock, O. (2014). Combining physicochemical and evolutionary information for protein contact prediction. *PLoS ONE*, *9*(10), e108438. https://doi.org/10.1371/journal.pone.0108438

Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics (Oxford, England)*, *30*(21), 3128–3130. https://doi.org/10.1093/bioinformatics/btu500

Seffernick, J. T., & Lindert, S. (2020). Hybrid methods for combined experimental and computational determination of protein structure. In *Journal of Chemical Physics* (Vol. 153, Issue 24, p. 240901). American Institute of Physics Inc. https://doi.org/10.1063/5.0026025

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Senior, Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1141–1148. https://doi.org/10.1002/prot.25834

Shackelford, G., & Karplus, K. (2007). Contact prediction using mutual information and neural nets. *Proteins: Structure, Function and Genetics*, *69*(SUPPL. 8), 159–164. https://doi.org/10.1002/prot.21791

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shindyalov, I. N., Kolchanov, N. A., & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering, Design and Selection*, *7*(3), 349–358. https://doi.org/10.1093/protein/7.3.349

Shrestha, R., Fajardo, E., Gil, N., Fidelis, K., Kryshtafovych, A., Monastyrskyy, B., & Fiser, A. (2019). Assessing the accuracy of contact predictions in CASP13. *Proteins*, *87*(12), 1058–1068. https://doi.org/10.1002/prot.25819

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 Research*, *45*, 422–428. https://doi.org/10.1093/nar/gkx249

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., & Orengo, C. A. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, *43*(D1), D376–D381. https://doi.org/10.1093/nar/gku947

Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M., & Gray, J. J. (2010). A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins: Structure, Function, and Bioinformatics*, *78*(15), 3115–3123. https://doi.org/10.1002/prot.22765

Skolnick, J., Kolinski, A., & Ortiz, A. R. (1997). MONSSTER: A method for folding globular proteins with a small number of distance restraints. *Journal of Molecular Biology*, *265*(2), 217–241. https://doi.org/10.1006/jmbi.1996.0720

Skwark, M. J., Raimondi, D., Michel, M., & Elofsson, A. (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology*, *10*(11), e1003889. https://doi.org/10.1371/journal.pcbi.1003889

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews*, *66*(1), 334–395. https://doi.org/10.1124/pr.112.007336

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*(7), 951–960. https://doi.org/10.1093/bioinformatics/bti125

Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., & Baker, D. (2013). High-Resolution Comparative Modeling with RosettaCM. *Structure*, *21*(10), 1735–1742. https://doi.org/10.1016/j.str.2013.08.005

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-04964-5

Stoker, H. S. (Howard S. (2013). *Organic and biological chemistry* (Alyssa White (ed.); 6th ed.). Cengage Learning.

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics (Oxford, England)*, *36*(6), 1765–1771. https://doi.org/10.1093/bioinformatics/btz828

Stumpff-Kane, A. W., Maksimiak, K., Lee, M. S., & Feig, M. (2007). Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), 1345–1356. https://doi.org/10.1002/prot.21674

Summa, C. M., & Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(9), 3177–3182. https://doi.org/10.1073/pnas.0611593104

Talavera, Laskowski, & Thornton. (2009). WSsas: A web service for the annotation of functional residues through structural homologues | Request PDF. *Bioinformatics*, *25*(9). https://www.researchgate.net/publication/24146311_WSsas_A_web_service_for_the_annotation_of_functional_residues_through_structural_homologues

Taylor, T. J., Bai, H., Tai, C. H., & Lee, B. (2014). Assessment of CASP10 contact-assisted predictions. *Proteins: Structure, Function and Bioinformatics*, *82*(SUPPL.2), 84–97. https://doi.org/10.1002/prot.24367

Terashi, G., & Kihara, D. (2018). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function, and*

Bioinformatics, *86*, 189–201. https://doi.org/10.1002/prot.25373

Tetchner, S., Kosciolek, T., & Jones, D. T. (2014). Opportunities and limitations in applying coevolution-derived contacts to protein. In *Bio-Algorithms and Med-Systems* (Vol. 10, Issue 4, pp. 243–254). Walter de Gruyter GmbH. https://doi.org/10.1515/bams-2014-0013

Tompa, P., & Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *The FASEB Journal*, *18*(11), 1169–1175. https://doi.org/10.1096/fj.04-1584rev

Torchala, M., Bates, P. A., McGuffin, L. J., Buenavista, M. T., & Roche, D. B. (2013). Predicting the structure of protein–protein complexes using the Swarmdock web server. *Nucleic Acids Research*, *1137*(Web Server issue), 368–372. https://doi.org/10.1007/978-1-4939-0366-5_13

Tramontano, A., Cozzetto, D., Giorgetti, A., & Raimondo, D. (2008). The Assessment of Methods for Protein Structure Prediction. In *Protein Structure Prediction* (pp. 43–57). Humana Press. https://doi.org/10.1007/978-1-59745-574-9_2

Tyka, M. D., Keedy, D. A., André, I., Dimaio, F., Song, Y., Richardson, D. C., Richardson, J. S., & Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, *405*(2), 607–618. https://doi.org/10.1016/j.jmb.2010.11.008

UniProt Consortium. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *36*(Database), D190–D195. https://doi.org/10.1093/nar/gkm895

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(Database issue), D204-12. https://doi.org/10.1093/nar/gku989

UniProt Consortium, T. U. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, *39*(Database issue), D214-9. https://doi.org/10.1093/nar/gkq1020

Uversky, V. N. (2002). What does it mean to be natively unfolded? In *European Journal of Biochemistry* (Vol. 269, Issue 1, pp. 2–12). John Wiley & Sons, Ltd. https://doi.org/10.1046/j.0014-2956.2001.02649.x

Uziela, K., Hurtado, D. M., Wallner, B., & Elofsson, A. (2016). *ProQ3D: Improved model quality assessments using Deep Learning*. http://arxiv.org/abs/1610.05189

Uziela, K., Shu, N., Wallner, B., & Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Nature Publishing Group*, *6*(August), 1–10. https://doi.org/10.1038/srep33509

Uziela, K., & Wallner, B. (2016). ProQ2: Estimation of model accuracy implemented in Rosetta. *Bioinformatics*, *32*(9), 1411–1413. https://doi.org/10.1093/bioinformatics/btv767

Vendruscolo, M., & Domany, E. (2000). Protein folding using contact maps. In *Vitamins and Hormones* (Vol. 58, pp. 171–212). https://doi.org/10.1016/s0083-6729(00)58025-x

Vitkup, D., Melamud, E., Moult, J., & Sander, C. (2001). Completeness in structural genomics. *Nature Structural Biology*, *8*(6), 559–566. https://doi.org/10.1038/88640

Vullo, A., Walsh, I., & Pollastri, G. (2006). A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, *7*(1), 180. https://doi.org/10.1186/1471-2105-7-180

Wallner, B., & Elofsson, A. (2003). Can correct protein models be identified? *Protein Science : A Publication of the Protein Society*, *12*(5), 1073–1086. https://doi.org/10.1110/ps.0236803

Wallner, B., Fang, H., & Elofsson, A. (2003). Automatic Consensus-Based Fold Recognition Using Pcons, ProQ, and Pmodeller. *Proteins: Structure, Function and Genetics*, *53*(SUPPL. 6), 534–541. https://doi.org/10.1002/prot.10536

Wallner, & Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science*, *15*(4), 900–913. https://doi.org/10.1110/ps.051799606

Wallner, & Elofsson, A. (2007). Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 184–193. https://doi.org/10.1002/prot.21774

Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, *13*(1), e1005324. https://doi.org/10.1371/journal.pcbi.1005324

Wang, Y., Shi, Q., Yang, P., Zhang, C., Mortuza, S. M., Xue, Z., Ning, K., & Zhang, Y. (2019). Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biology*, *20*(1), 229. https://doi.org/10.1186/s13059-019-1823-z

Wang, Z., Eickholt, J., & Cheng, J. (2011). APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics*, *27*(12), 1715–1716. https://doi.org/10.1093/bioinformatics/btr268

Wang, Z., Tegge, A. N., & Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Structure, Function and Bioinformatics*, *75*(3), 638–647. https://doi.org/10.1002/prot.22275

Ward, J. J., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, *19*(13), 1650–1655. https://doi.org/10.1093/bioinformatics/btg223

Wass, & Sternberg. (2008). *ConFunc - functional annotation in the twilight zone*. Bioinformatics. https://www.semanticscholar.org/paper/ConFunc-functional-annotation-in-the-twilight-zone-Wass-Sternberg/6be6387300477c3df61091ed1bec7150531f6d1a

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., & Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(1), 67–72. https://doi.org/10.1073/pnas.0805923106

Westbrook, J., Feng, Z., Chen, L., Yang, H., & Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Research*, *31*(1), 489–491. https://doi.org/10.1093/nar/gkg068

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(SUPPL.2), 407–410. https://doi.org/10.1093/nar/gkm290

Wieman, H., Tøndel, K., Anderssen, E., & Drabløs, F. (2004). Homology-based modelling of targets for rational drug design. *Mini Reviews in Medicinal Chemistry*, *4*(7), 793–804. http://www.ncbi.nlm.nih.gov/pubmed/15379646

Wierschin, T., Wang, K., Welter, M., Waack, S., & Stanke, M. (2015). Combining features in a graphical model to predict protein binding sites. *Proteins*, *83*(5), 844–852. https://doi.org/10.1002/prot.24775

Williamson, M. (2012a). Protein Domains. In *How Proteins Work* (pp. 78–115). Garland Science. https://doi.org/10.1201/9781136665493-5

Williamson, M. (2012b). Techniques for Studying Proteins. In *How Proteins Work* (pp. 396–451). Garland Science. https://doi.org/10.1201/9781136665493-14

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein

structure-function paradigm. *Journal of Molecular Biology*, *293*(2), 321–331. https://doi.org/10.1006/jmbi.1999.3110

Wu, Peng, Z., Zhang, Y., & Yang, J. (2018). COACH-D: Improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Research*, *46*(W1), W438–W442. https://doi.org/10.1093/nar/gky439

Wu, S., Szilagyi, A., & Zhang, Y. (2011). Improving protein structure prediction using multiple sequence-based contact predictions. *Structure*, *19*(8), 1182–1191. https://doi.org/10.1016/j.str.2011.05.004

Wu, S., & Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England)*, *24*(7), 924–931. https://doi.org/10.1093/bioinformatics/btn069

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current Protein & Peptide Science*, *7*(3), 217–227. http://www.ncbi.nlm.nih.gov/pubmed/16787261

Xu, D., & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*, *101*(10), 2525–2534. https://doi.org/10.1016/j.bpj.2011.10.024

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., & Rost, B. (2014). PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, *42*(Web Server issue), W337-43. https://doi.org/10.1093/nar/gku366

Yang, J., Roy, A., Zhang, Y., Yang, Roy, & Zhang. (2013). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)*, *29*(20), 2588–2595. https://doi.org/10.1093/bioinformatics/btt447

Yang, J., Wang, Y., & Zhang, Y. (2016). ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *Journal of Molecular Biology*, *428*(4), 693–701. https://doi.org/10.1016/j.jmb.2015.09.024

Yang, Roy, & Zhang. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, *41*(Database issue), D1096-103. https://doi.org/10.1093/nar/gks966

Yang, & Zhou. (2008a). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function, and Bioinformatics*, *72*(2), 793–803. https://doi.org/10.1002/prot.21968

Yang, & Zhou. (2008b). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Science : A Publication of the Protein Society*, *17*(7), 1212–1219. https://doi.org/10.1110/ps.033480.107

Ye, Feenstra, Heringa, Ijzerman, & Marchiori. (2008). Multi-RELIEF: A method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, *24*(1), 18–25. https://doi.org/10.1093/bioinformatics/btm537

Yip, K. M., Fischer, N., Paknia, E., Chari, A., & Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, *587*(7832), 157–161. https://doi.org/10.1038/s41586-020-2833-4

Yu, D. J., Hu, J., Li, Q. M., Tang, Z. M., Yang, J. Y., & Shen, H. Bin. (2015). Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. *IEEE Transactions on Nanobioscience*, *14*(1), 44–57. https://doi.org/10.1109/TNB.2015.2394328

Yu, D. J., Hu, J., Yang, J. Y., Shen, H. Bin, Tang, J., & Yang, J. Y. (2013). Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(4), 994–1008. https://doi.org/10.1109/TCBB.2013.104

Zhang. (2009). Protein structure prediction: when is it useful? In *Current Opinion in Structural Biology* (Vol. 19, Issue 2, pp. 145–155). https://doi.org/10.1016/j.sbi.2009.02.005

Zhang, Liang, & Zhang, Y. (2011). Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure*, *19*(12), 1784–1795. https://doi.org/10.1016/J.STR.2011.09.022

Zhang, & Skolnick. (2004a). SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, *25*(6), 865–871. https://doi.org/10.1002/jcc.20011

Zhang, & Skolnick. (2004b). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Zhang, & Zhang. (2010). A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLoS ONE*, *5*(10), e15386. https://doi.org/10.1371/journal.pone.0015386

Zhang, Zheng, W., Mortuza, S. M., Li, Y., & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, *36*(7), 2105–2112. https://doi.org/10.1093/bioinformatics/btz863

Zheng, Zhang, C., Wuyun, Q., Pearce, R., Li, Y., & 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), W429–W436. https://doi.org/10.1093/nar/gkz384

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science : A Publication of the Protein Society*, *11*(11), 2714–2726. https://doi.org/10.1110/ps.0217002

Zhu, Fan, Periole, Honig, & Mark. (2008). Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins: Structure, Function, and Bioinformatics*, *72*(4), 1171–1188. https://doi.org/10.1002/prot.22005

Zhu, X., Xiong, Y., & Kihara, D. (2014). Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. *Bioinformatics*, *31*(5), 707–713. https://doi.org/10.1093/bioinformatics/btu724

Appendices

Appendix 1

		GDT-HA score													
CASP TARGET			The original MD-based protocol of ReFOLD						The local quality assessment guided MD-based protocol (threshold is 3 Ångströms)						Wilcoxon test
Target ID by domain	CASP Category	Starting model	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Significance
T0859	Regular	0.1681	0.1571	-0.011	0.18277	0.01467	0.2146	0.0465	0.1615	-0.0066	0.17192	0.00382	0.1814	0.0133	***
T0862	Regular	0.4032	0.3575	-0.0457	0.390687	-0.012513	0.4328	0.0296	0.371	-0.0322	0.396678	-0.006522	0.414	0.0108	*
T0866	Regular	0.2391	0.187	-0.0521	0.222985	-0.016115	0.2543	0.0152	0.2304	-0.0087	0.241027	0.001927	0.2565	0.0174	***
T0880	Regular	0.0596	0.0544	-0.0052	0.0597755	0.0001755	0.0674	0.0078	0.0544	-0.0052	0.0576659	-0.0019341	0.0622	0.0026	***
T0886	Regular	0.1681	0.1638	-0.0043	0.179839	0.011739	0.1987	0.0306	0.1627	-0.0054	0.174943	0.006843	0.1921	0.024	***
T0897	Regular	0.0582	0.0534	-0.0048	0.0595611	0.0013611	0.0687	0.0105	0.0563	-0.0019	0.0587799	0.0005799	0.062	0.0038	*
T0904	Regular	0.2267	0.1752	-0.0515	0.213548	-0.013152	0.2508	0.0241	0.2122	-0.0145	0.221043	-0.005657	0.2323	0.0056	***
T0915	Regular	0.2808	0.2451	-0.0357	0.268702	-0.012098	0.3068	0.026	0.2614	-0.0194	0.273628	-0.007172	0.289	0.0082	***
TR594	Refinement	0.3427	0.323	-0.0197	0.352589	0.009889	0.3876	0.0449	0.3287	-0.014	0.344157	0.001457	0.3652	0.0225	**
TR862	Refinement	0.4032	0.3495	-0.0537	0.397913	-0.005287	0.457	0.0538	0.3656	-0.0376	0.394813	-0.008387	0.422	0.0188	*
TR866	Refinement	0.6082	0.5433	-0.0649	0.597188	-0.011012	0.6587	0.0505	0.5769	-0.0313	0.602386	-0.005814	0.6346	0.0264	***
TR869	Refinement	0.2885	0.2476	-0.0409	0.272706	-0.015794	0.2933	0.0048	0.262	-0.0265	0.276501	-0.011999	0.2909	0.0024	***
TR870	Refinement	0.25	0.2064	-0.0436	0.249615	-0.000385	0.305	0.055	0.2431	-0.0069	0.257492	0.007492	0.2729	0.0229	***
TR905	Refinement	0.3244	0.2448	-0.0796	0.282299	-0.042101	0.3223	-0.0021	0.2851	-0.0393	0.298757	-0.025643	0.3223	-0.0021	***
The Cumulative scores		3.8208	3.3081	-0.5127	3.7301776	-0.0906224	4.218	0.3972	3.5713	-0.2495	3.7697908	-0.0510092	3.9974	0.1766	

Table S. 1 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 FM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. Ho: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 2

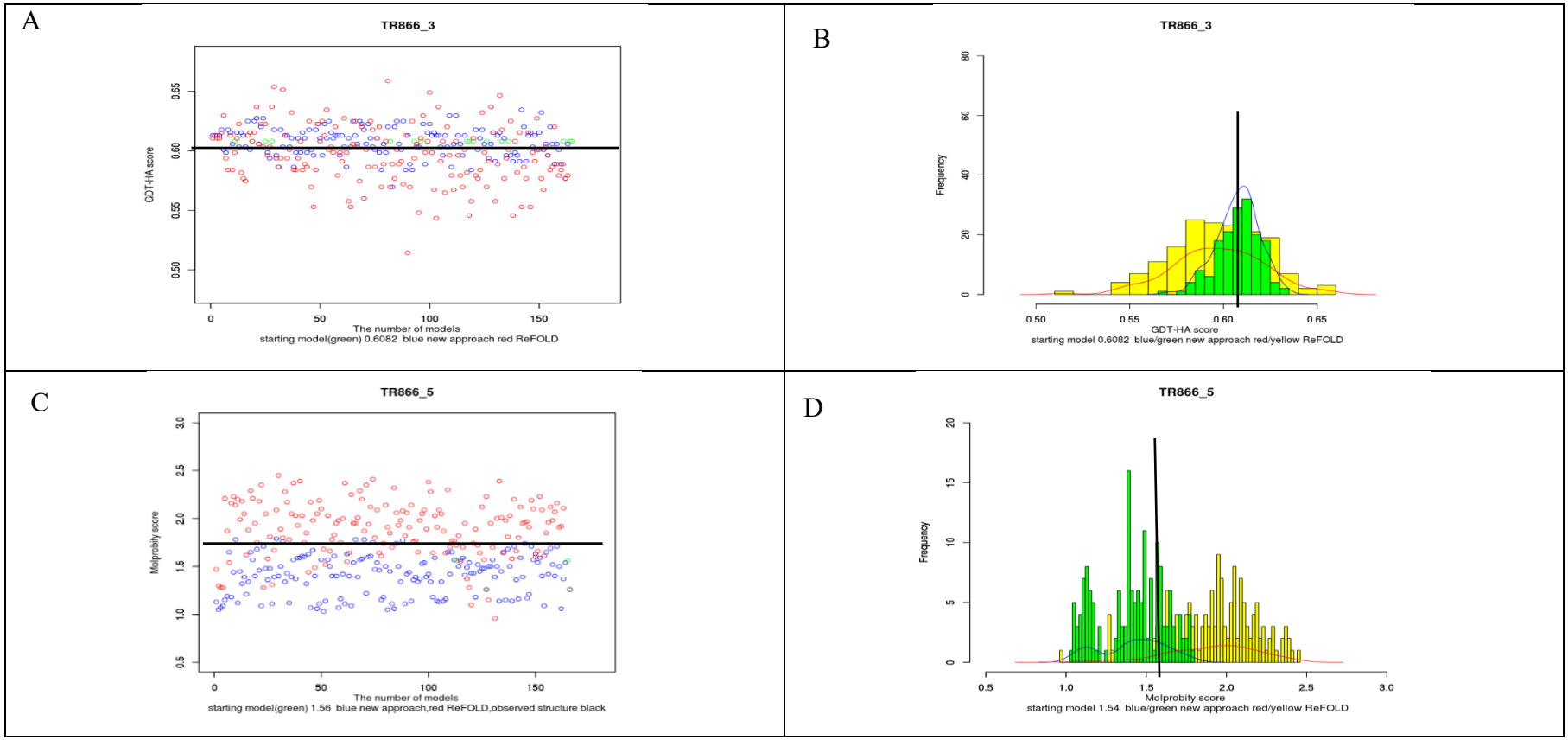

Figure S. 1 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM target.

Performance of methods on TR866 (an FM category CASP12 refinement target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 3

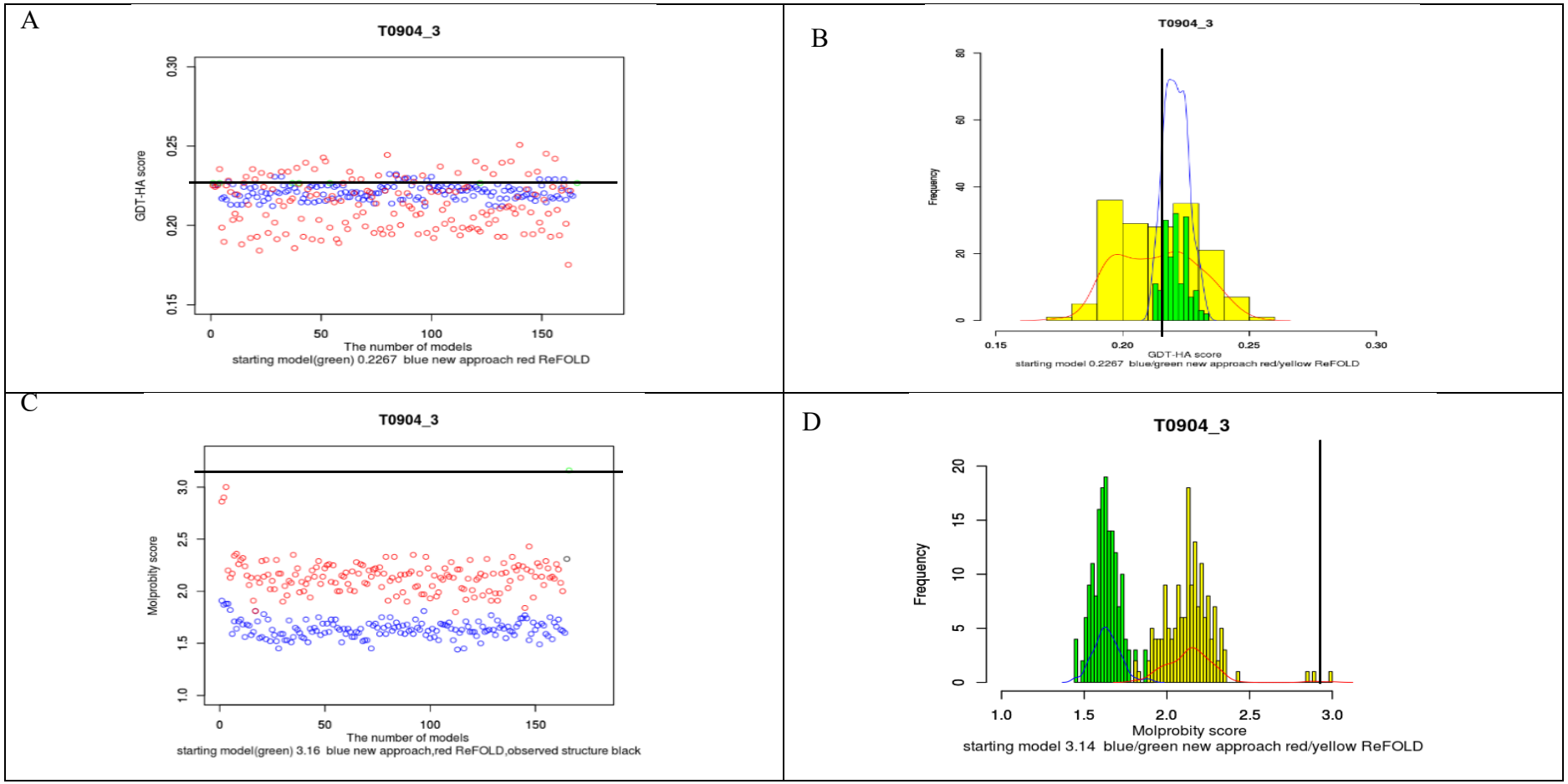

Figure S. 2 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM target.

Performance of methods on T0904 (an FM category CASP12 target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 4

		GDT-HA score													
CASP TARGET			The original MD-based protocol of ReFOLD						The local quality assessment guided MD-based protocol (threshold is 3 Ångströms)						Wilcoxon test
Target ID by domain	CASP Category	Starting model	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Significance
T0890	Regular	0.1742	0.1516	-0.0226	0.162086	-0.012114	0.1769	0.0027	0.1609	-0.0133	0.168949	-0.005251	0.1769	0.0027	***
T0898	Regular	0.1599	0.1242	-0.0357	0.140154	-0.019746	0.163	0.0031	0.1475	-0.0124	0.158659	-0.001241	0.1693	0.0094	***
T0899	Regular	0.1628	0.1347	-0.0281	0.152224	-0.010576	0.1707	0.0079	0.1448	-0.018	0.153423	-0.009377	0.1664	0.0036	***
T0909	Regular	0.2703	0.268	-0.0023	0.303593	0.033293	0.3341	0.0638	0.268	-0.0023	0.278384	0.008084	0.2943	0.024	***
T0945	Regular	0.3493	0.3167	-0.0326	0.355063	0.005763	0.378	0.0287	0.3327	-0.0166	0.34364	-0.00566	0.3587	0.0094	**
TR694	Refinement	0.2376	0.2129	-0.0247	0.233538	-0.004062	0.2538	0.0162	0.212	-0.0256	0.22704	-0.01056	0.2414	0.0038	***
TR868	Refinement	0.6143	0.5095	-0.1048	0.568945	-0.045355	0.6738	0.0595	0.5738	-0.0405	0.604776	-0.009524	0.6548	0.0405	**
TR890	Refinement	0.3245	0.2434	-0.0811	0.281735	-0.042765	0.3285	0.004	0.2899	-0.0346	0.305804	-0.018696	0.3271	0.0026	**
TR896	Refinement	0.468	0.3837	-0.0843	0.421912	-0.046088	0.468	0	0.407	-0.061	0.443089	-0.024911	0.4826	0.0146	***
TR898	Refinement	0.2524	0.2123	-0.0401	0.245933	-0.006467	0.2901	0.0377	0.2358	-0.0166	0.24816	-0.00424	0.2618	0.0094	***
TR901	Refinement	0.3061	0.2388	-0.0673	0.273182	-0.032918	0.315	0.0089	0.2836	-0.0225	0.30088	-0.00522	0.3161	0.01	***
TR909	Refinement	0.4257	0.3566	-0.0691	0.393724	-0.031976	0.4399	0.0142	0.4159	-0.0098	0.430218	0.004518	0.4452	0.0195	***
TR945	Refinement	0.412	0.386	-0.026	0.415109	0.003109	0.4493	0.0373	0.3807	-0.0313	0.396554	-0.015446	0.412	0	***
The Cumulative scores		4.1571	3.5384	-0.6187	3.947198	-0.209902	4.4411	0.284	3.8526	-0.3045	4.059576	-0.097524	4.3066	0.1495	

Table S. 2 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 FM/TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. H_o: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H_1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 5

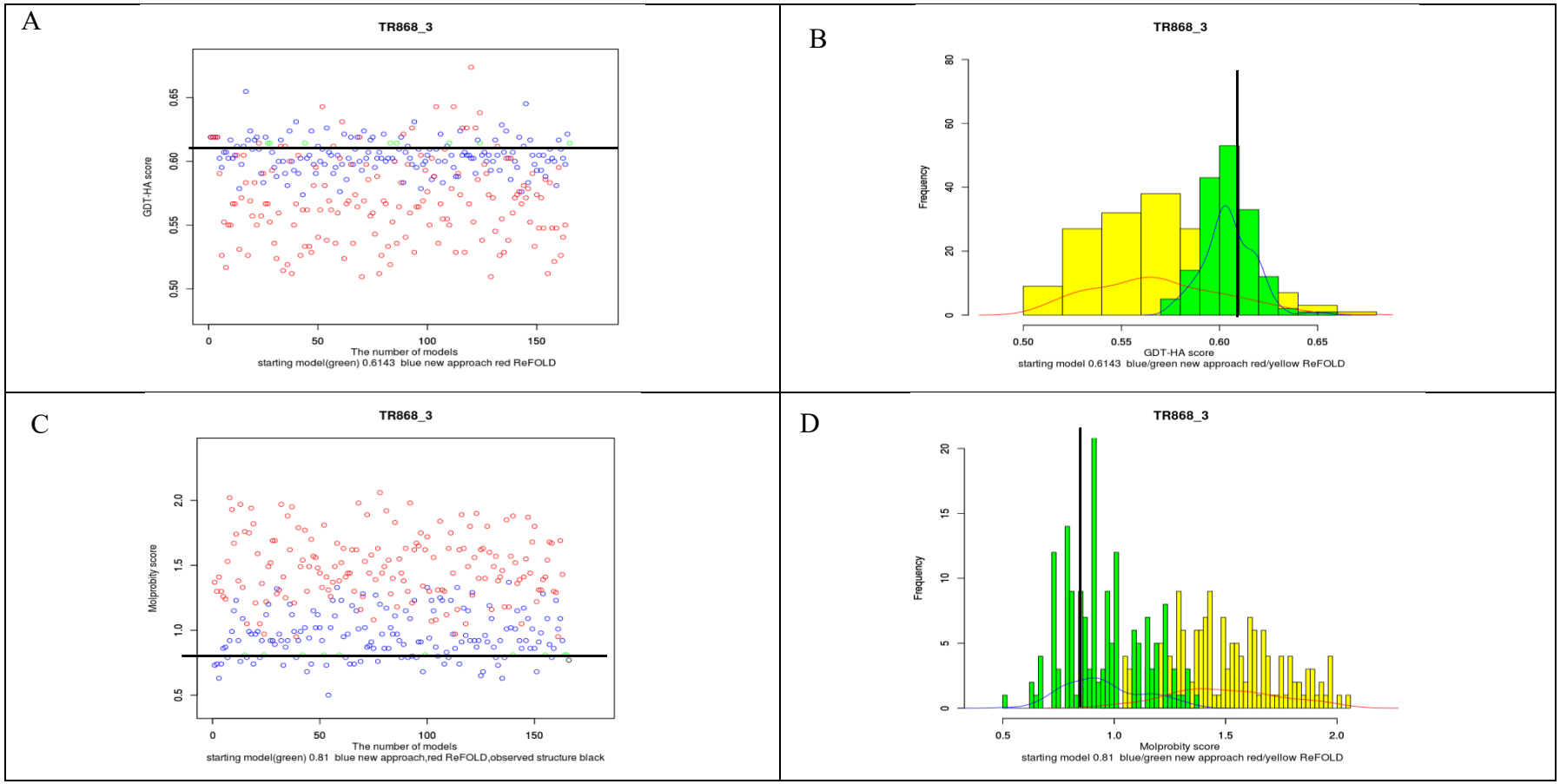

Figure S. 3 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM/TBM target.

Performance of methods on TR868 (an FM/TBM CASP12 refinement target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 6

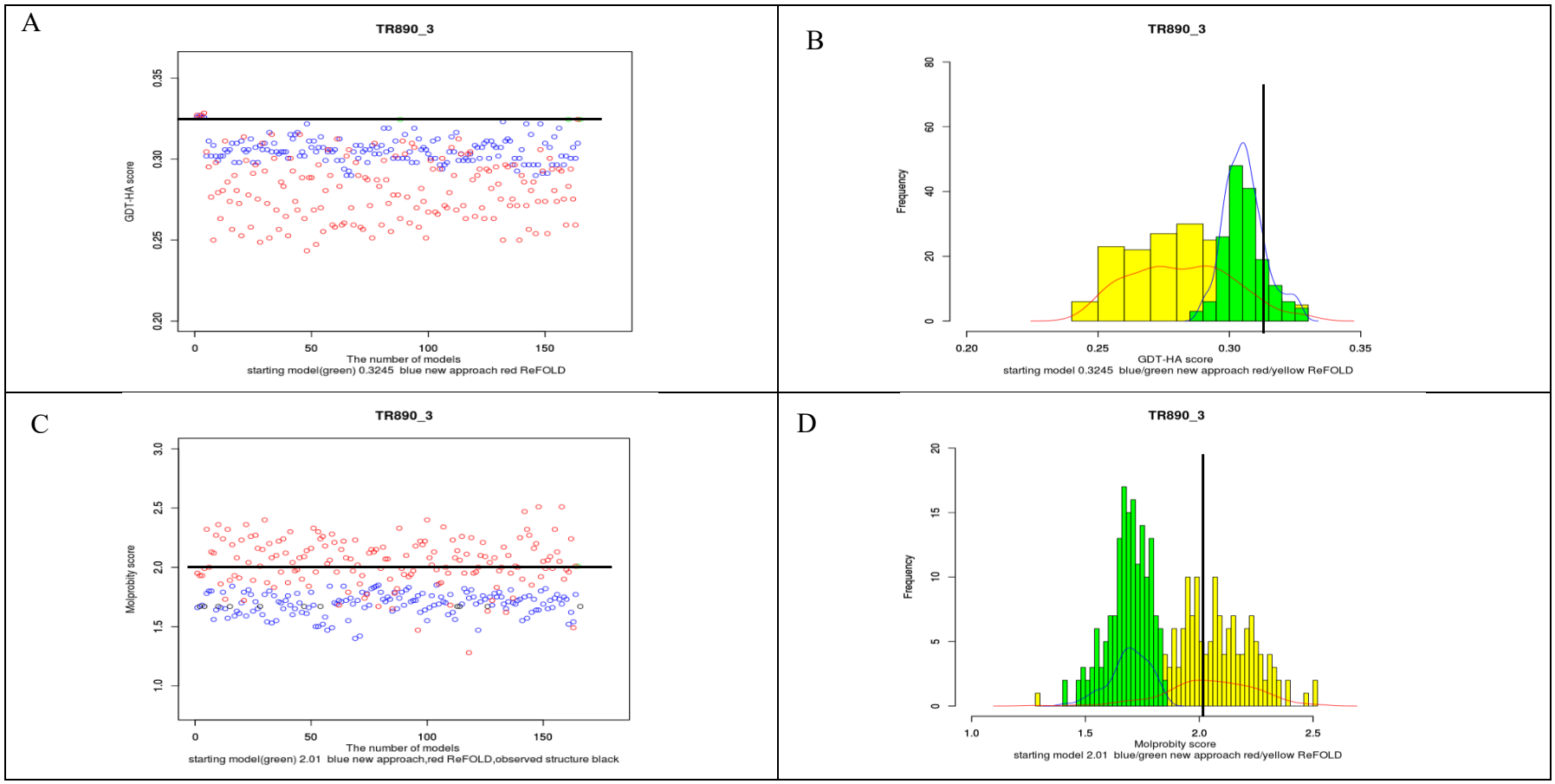

Figure S. 4 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on an FM/TBM target.

Performance of methods on TR890 (an FM/TBM CASP12 refinement target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 7

CASP TARGET		GDT-HA score													Wilcoxon test
			The original MD-based protocol of ReFOLD						The local quality assessment guided MD-based protocol (threshold is 3 Ångströms)						
Target ID by domain	CASP Category	Starting model	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Minimum Score	Diff. Min	Mean Score	Diff. Mean	Maximum Score	Diff. Max	Significance
T0872	Regular	0.4886	0.3977	0.0909	0.437362	-0.051238	0.4943	0.0057	0.4432	-0.0454	0.472564	-0.016036	0.5028	0.0142	***
T0882	Regular	0.5791	0.5538	0.0253	0.621886	0.042786	0.6994	0.1203	0.557	-0.0221	0.588816	0.009716	0.6297	0.0506	***
T0895	Regular	0.5167	0.4229	0.0938	0.473702	-0.042998	0.525	0.0083	0.4771	-0.0396	0.501923	-0.014777	0.5292	0.0125	***
T0911	Regular	0.2972	0.1961	0.1011	0.226078	-0.071122	0.28	-0.0172	0.2794	-0.0178	0.291071	-0.006129	0.3027	0.0055	***
T0913	Regular	0.4149	0.3232	0.0917	0.353898	-0.061002	0.412	-0.0029	0.3802	-0.0347	0.398787	-0.016113	0.4142	-0.0007	***
T0944	Regular	0.5504	0.4545	0.0959	0.512469	-0.037931	0.5514	0.001	0.5	-0.0504	0.533065	-0.017335	0.5534	0.003	***
T0946	Regular	0.2962	0.2166	0.0796	0.240111	-0.056089	0.2945	-0.0017	0.2551	-0.0411	0.27321	-0.02299	0.2954	-0.0008	***
T0947	Regular	0.4814	0.4143	0.0671	0.453246	-0.028154	0.5086	0.0272	0.4329	-0.0485	0.459048	-0.022352	0.4771	-0.0043	***
T0948	Regular	0.5235	0.4144	0.1091	0.477226	-0.046274	0.5352	0.0117	0.4698	-0.0537	0.495054	-0.028446	0.5268	0.0033	***
TR520	Refinement	0.581	0.4315	0.1495	0.523768	-0.057232	0.5818	0.0008	0.5156	-0.0654	0.543085	-0.037915	0.581	0	***
TR872	Refinement	0.5682	0.4716	0.0966	0.552178	-0.016022	0.6307	0.0625	0.5199	-0.0483	0.541712	-0.026488	0.5739	0.0057	***
TR877	Refinement	0.4894	0.4155	0.0739	0.453761	-0.035639	0.4842	-0.0052	0.4472	-0.0422	0.466093	-0.023307	0.4877	-0.0017	***
TR879	Refinement	0.633	0.4341	0.1989	0.497477	-0.135523	0.6261	-0.0069	0.5614	-0.0716	0.583705	-0.049295	0.6261	-0.0069	***
TR881	Refinement	0.479	0.3205	0.1585	0.410781	-0.068219	0.4715	-0.0075	0.4257	-0.0533	0.452469	-0.026531	0.4802	0.0012	***
TR882	Refinement	0.6899	0.6013	0.0886	0.668846	-0.021054	0.7437	0.0538	0.6677	-0.0222	0.71137	0.02147	0.75	0.0601	***
TR885	Refinement	0.7837	0.6418	0.1419	0.706848	-0.076852	0.7885	0.0048	0.6995	-0.0842	0.747322	-0.036378	0.7909	0.0072	***
TR891	Refinement	0.7567	0.6071	0.1496	0.693176	-0.063524	0.7701	0.0134	0.683	-0.0737	0.721636	-0.035064	0.7589	0.0022	***
TR893	Refinement	0.6908	0.605	0.0858	0.661648	-0.029152	0.7219	0.0311	0.6272	-0.0636	0.655179	-0.035621	0.6997	0.0089	**
TR895	Refinement	0.5146	0.3896	-0.125	0.450184	-0.064416	0.4917	-0.0229	0.4688	-0.0458	0.492889	-0.021711	0.5188	0.0042	***
TR913	Refinement	0.4534	0.3587	0.0947	0.401011	-0.052389	0.4482	-0.0052	0.4164	-0.037	0.436608	-0.016792	0.4586	0.0052	***
TR917	Refinement	0.6535	0.5799	0.0736	0.628904	-0.024596	0.6829	0.0294	0.61	-0.0435	0.637371	-0.016129	0.6618	0.0083	***
TR920	Refinement	0.6039	0.484	0.1199	0.548031	-0.055869	0.6279	0.024	0.5434	-0.0605	0.568527	-0.035373	0.6016	-0.0023	***
TR921	Refinement	0.4801	0.4203	0.0598	0.477502	-0.002598	0.5236	0.0435	0.4402	-0.0399	0.465543	-0.014557	0.4928	0.0127	***
T0922	Refinement	0.7581	0.5766	0.1815	0.709603	-0.048497	0.8024	0.0443	0.7177	-0.0404	0.766796	0.008696	0.8105	0.0524	***
TR928	Refinement	0.4274	0.3087	0.1187	0.356789	-0.070611	0.4267	-0.0007	0.3783	-0.0491	0.396442	-0.030958	0.4245	-0.0029	***
TR942	Refinement	0.3333	0.2752	0.0581	0.304669	-0.028631	0.3424	0.0091	0.3081	-0.0252	0.320707	-0.012593	0.332	-0.0013	***
TR944	Refinement	0.5603	0.4684	0.0919	0.51819	-0.04211	0.5642	0.0039	0.501	-0.0593	0.529974	-0.030326	0.5583	-0.002	***
TR947	Refinement	0.5157	0.4557	-0.06	0.502683	-0.013017	0.5357	0.02	0.49	-0.0257	0.511695	-0.004005	0.5329	0.0172	***
TR948	Refinement	0.5956	0.5352	0.0604	0.591222	-0.004378	0.6527	0.0571	0.5705	-0.0251	0.594791	-0.000809	0.6242	0.0286	***
The Cumulative scores		15.7156	12.7742	-2.944	14.453249	-1.262351	16.2173	0.5017	14.3863	-1.3293	15.157452	-0.558148	15.9957	0.2801	

Table S. 3 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. Ho: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 8

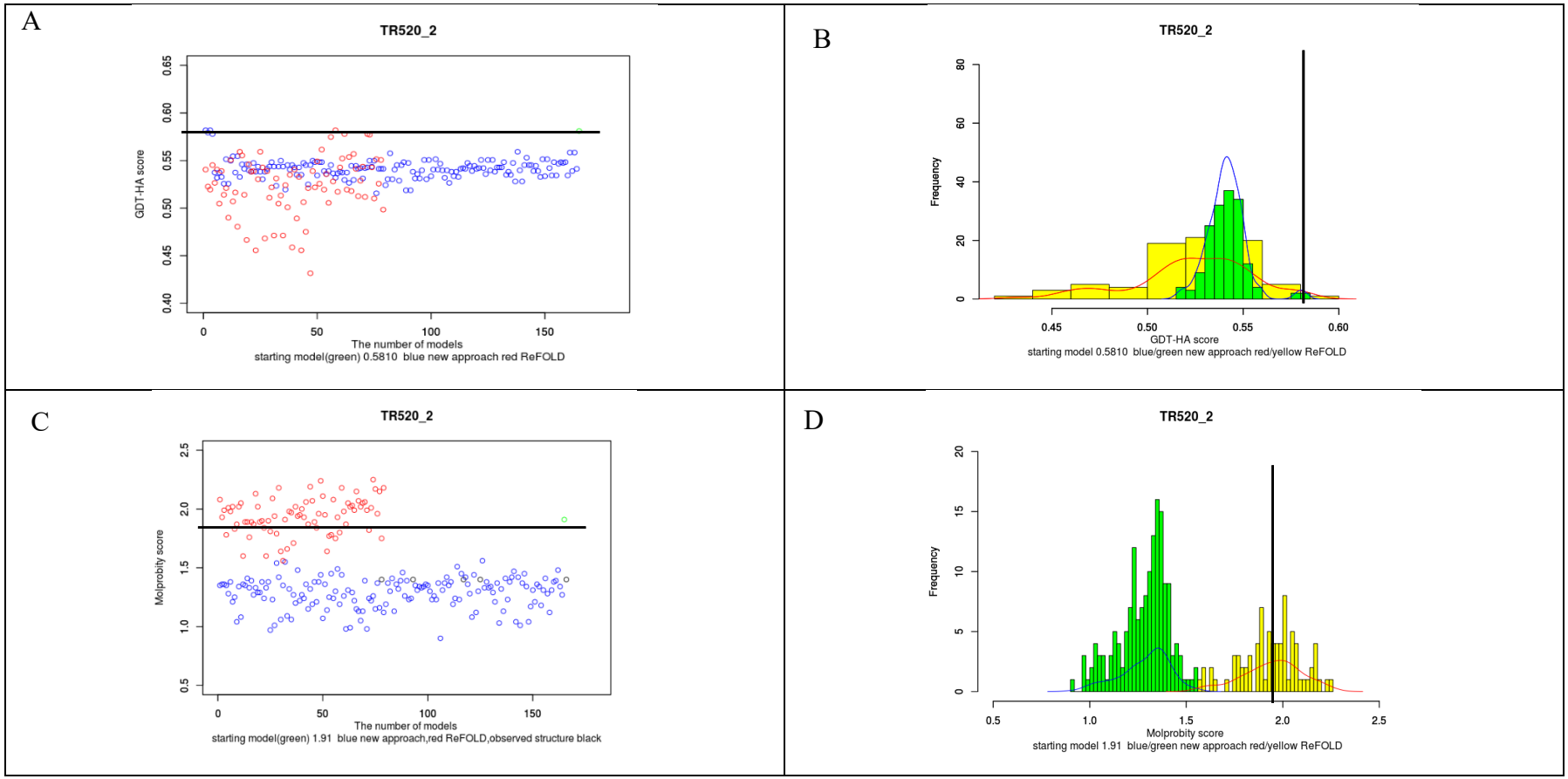

Figure S. 5 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on a TBM target.

Performance of methods on TR520 (a TBM CASP12 refinement target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 9

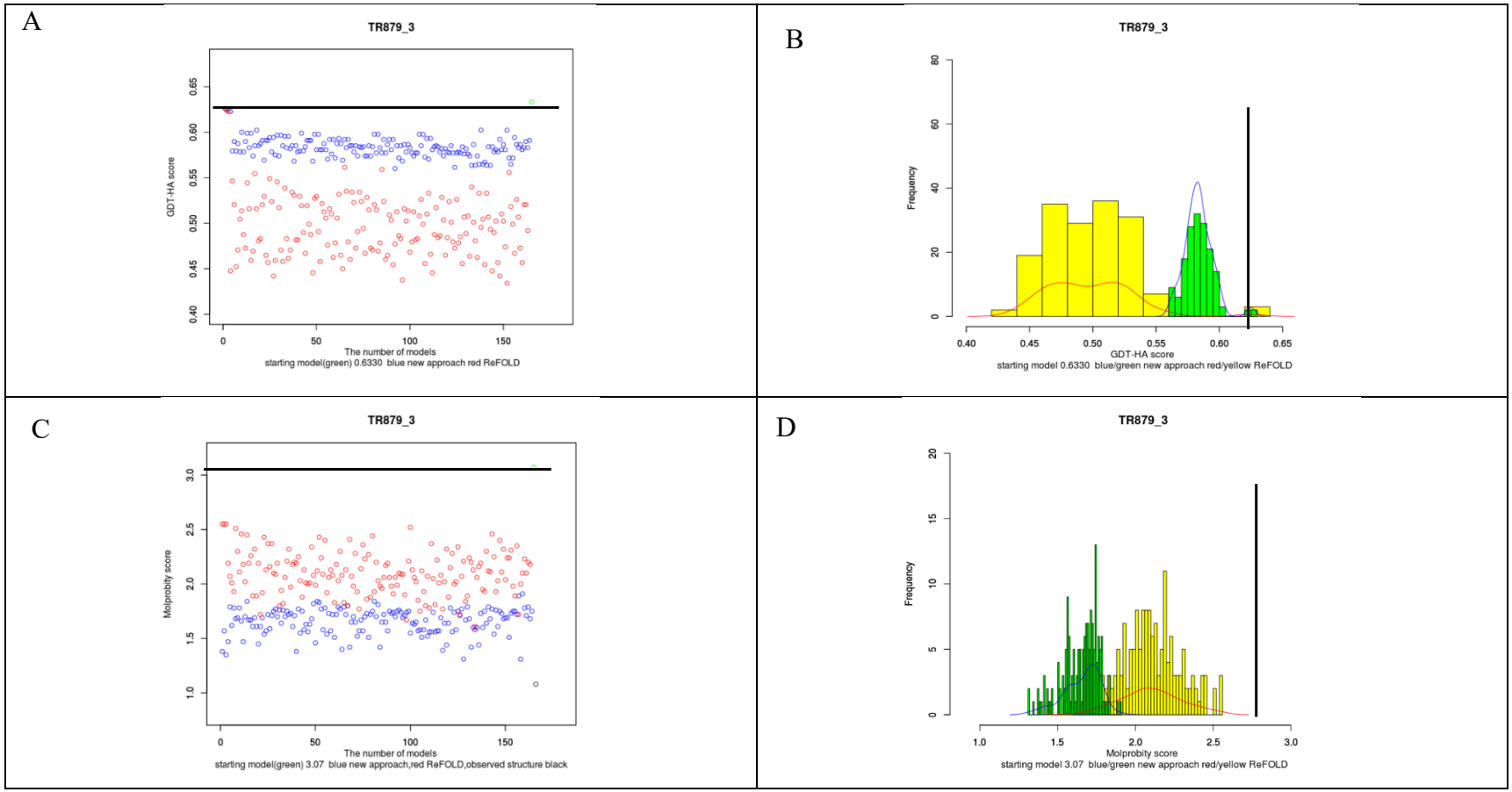

Figure S. 6 A comparison of the original MD-based protocol of ReFOLD and the local quality assessment guided MD-based protocol on a TBM target.

Performance of methods on TR879 (a TBM CASP12 refinement target) according to GDT-HA score and Molprobity Score. (A) The blue points indicate scores for the models generated using the local quality assessment guided MD-based protocol, the red points indicate scores for the models generated using the original MD-based protocol of ReFOLD, and the black line represents the starting model score. The points above the black line indicate the improved models. (B) The blue line and green bars represent the scores of models generated using the local quality assessment guided MD-based protocol, the red line and yellow bars represent models generated using the original MD-based protocol of ReFOLD and the black line represents the starting model (higher GDT-HA scores are better), (C) and (D) ditto but according to the Molprobity Score (lower scores are better).

Appendix 10

CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Maximum vs Starting
FM	**0.03571**	0.07571	0.9995	**6.104e-05**
TBM	**9.115e-07**	**3.459e-06**	0.9415	**0.0005068**
FM/TBM	**0.001662**	**0.04529**	0.9787	**0.0008281**
ALL	**1.383e-10**	**7.007e-06**	0.9998	**1.267e-08**

Table S. 4 Calculated pairwise p-values for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 targets according to GDT-HA score.

H_o: The scores of the targets refined by the local quality assessment guided MD-based protocol are equal or lower in quality than those refined by the original MD-based protocol of ReFOLD. H_1: The scores of targets refined by the local quality assessment guided MD-based protocol are higher quality models than those refined by the original MD-based protocol of ReFOLD. The maximum score of the models generated by the local quality assessment guided MD-based protocol were also compared with the score of the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface, higher GDT-HA scores are better)

Appendix 11

CASP TARGET		Molprobity score													Wilcoxon test
			The original MD-based protocol of ReFOLD						The local quality assessment guided MD-based protocol Threshold is 3 Ångströms						
Target ID by domain	CASP Category	Starting model	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max	Significance
T0859	Regular	2.88	1.91043	0.96957	1.04	1.84	3.03	0.15	1.65713	1.22287	1.17	1.71	2	0.88	***
T0862	Regular	2.72	1.984	0.736	1.62	1.1	2.43	0.29	1.71963	1.00037	1.56	1.16	1.79	0.93	***
T0866	Regular	3.29	2.03113	1.25887	1.59	1.7	2.87	0.42	1.76933	1.52067	1.21	2.08	2.15	1.14	***
T0880	Regular	3.08	2.31331	0.76669	1.55	1.53	2.7	0.38	1.74982	1.33018	1.26	1.82	1.81	1.27	***
T0886	Regular	3.23	2.15975	1.07025	1.54	1.69	3.21	0.02	1.78854	1.44146	1.52	1.71	2.17	1.06	***
T0897	Regular	1.03	1.83235	0.80235	1.02	0.01	0.97	0.06	1.23537	0.20537	0.89	0.14	0.8	0.23	***
T0904	Regular	3.16	2.14509	1.01491	1.8	1.36	2.86	0.3	1.6375	1.5225	1.44	1.72	1.91	1.25	***
T0915	Regular	1.82	1.5254	0.2946	0.94	0.88	2.09	0.27	1.17085	0.64915	0.76	1.06	1.87	-0.05	***
TR594	Refinement	2.91	1.85626	1.05374	1.22	1.69	2.45	0.46	1.72317	1.18683	1.2	1.71	2.01	0.90	***
TR862	Refinement	2.26	1.42969	0.83031	0.56	1.7	2.19	0.07	1.24848	1.01152	0.78	1.48	1.55	0.71	***
TR866	Refinement	1.56	1.9038	-0.3438	0.96	0.6	2.23	0.67	1.45	0.11	1.07	0.49	1.73	-0.17	***
TR869	Refinement	1.98	1.7692	0.2108	1.01	0.97	2.24	0.26	1.36421	0.61579	0.94	1.04	1.72	0.26	***
TR870	Refinement	3.61	1.98092	1.62908	1.4	2.21	3.13	0.48	1.55659	2.05341	1.07	2.54	1.86	1.75	***
TR905	Refinement	2.36	2.09521	0.26479	1.68	0.68	2.43	0.07	1.52951	0.83049	1.16	1.2	1.7	0.66	***
The Cumulative scores		35.89	26.93654	10.55816	17.93	17.96	34.83	3.9	21.60013	14.70061	16.03	19.86	25.07	10.82	

Table S. 5 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 FM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. H_o: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H_1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 12

		Molprobity Score													
CASP TARGET		The original MD-based protocol of ReFOLD							The local quality assessment guided MD-based protocol Threshold is 3 Ångströms						Wilcoxon Test
Target ID by domain	CASP Category	Starting model	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max!	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max!	Significance
T0872	Regular	2.62	1.79524	0.82476	0.86	1.76	2.3	0.32	1.3414	1.2786	0.92	1.7	1.8	0.82	***
T0882	Regular	1.73	1.53706	0.19294	0.63	1.1	1.95	-0.22	1.0686	0.6614	0.66	1.07	1.31	0.42	***
T0895	Regular	2.24	1.96374	0.27626	1.18	1.06	2.26	-0.02	1.50561	0.73439	0.88	1.36	1.71	0.53	***
T0911	Regular	3.24	1.91684	1.32316	1.63	1.61	2.46	0.78	1.48238	1.75762	1.23	2.01	1.65	1.59	***
T0913	Regular	3.09	2.27877	0.81123	1.94	1.15	2.71	0.38	1.86073	1.22927	1.62	1.47	2.14	0.95	***
T0944	Regular	1.91	1.94301	0.03301	1.52	0.39	2.07	-0.16	1.43963	0.47037	1.13	0.78	1.53	0.38	***
T0946	Regular	3.74	2.3346	1.4054	1.81	1.93	3.41	0.33	1.88628	1.85372	1.63	2.11	2.19	1.55	***
T0948	Regular	2.72	1.79963	0.92037	1.3	1.42	2.64	0.08	1.62841	1.09159	1.31	1.41	1.9	0.82	***
TR520	Refinement	1.91	1.94304	0.03304	1.56	0.35	2.08	-0.17	1.28591	0.62409	0.9	1.01	1.44	0.47	***
TR872	Refinement	0.5	1.4008	-0.9008	0.66	-0.16	2.13	-1.63	0.93372	-0.43372	0.5	0	1.34	-0.84	***
TR877	Refinement	1.41	1.82299	0.41299	1.16	0.25	2.35	-0.94	1.23122	0.17878	0.77	0.64	1.42	-0.01	***
TR879	Refinement	3.07	2.10135	0.96865	1.61	1.46	2.55	0.52	1.65665	1.41335	1.31	1.76	1.79	1.28	***
TR881	Refinement	2.68	1.96253	0.71747	1.32	1.36	2.41	0.27	1.5203	1.1597	1.14	1.54	1.74	0.94	***
TR882	Refinement	0.5	1.31957	0.81957	2.33	-1.83	2.28	-1.78	0.681037	-0.181037	0.5	0	1.74	-1.24	***
TR885	Refinement	0.82	1.54018	0.72018	2.12	-1.3	2.07	-1.25	0.93878	-0.11878	0.56	0.26	1.22	-0.40	***
TR891	Refinement	1.56	1.64976	0.08976	1	0.56	2.11	-0.55	1.21762	0.34238	0.81	0.75	1.55	0.01	***
TR893	Refinement	1.51	1.63485	0.12485	1.14	0.37	1.99	-0.48	1.05195	0.45805	0.68	0.83	1.27	0.24	***
TR895	Refinement	2.22	1.71413	0.50587	1.19	1.03	2.04	0.18	1.21744	1.00256	0.87	1.35	1.51	0.71	***
TR913	Refinement	1.34	1.98595	0.64595	1.63	-0.29	2.37	-1.03	1.28902	0.05098	1.02	0.32	1.49	-0.15	***
TR917	Refinement	1.36	1.71256	0.35256	1.17	0.19	2.17	-0.81	1.16061	0.19939	0.87	0.49	1.44	-0.08	***
TR920	Refinement	1.61	1.76282	0.15282	1.26	0.35	2.06	-0.45	1.32872	0.28128	0.71	0.9	1.63	-0.02	***
TR921	Refinement	1.61	1.99675	0.38675	1.36	0.25	2.45	-0.84	1.3475	0.2625	0.9	0.71	1.53	0.08	***

T0922	Refinement	1.07	1.77311	0.70311	1.07	0	2.51	-1.44	1.23921	-0.16921	0.65	0.42	1.62	-0.55	***
TR928	Refinement	3.56	2.45543	1.10457	2.1	1.46	3.24	0.32	1.81598	1.74402	1.48	2.08	1.93	1.63	***
TR942	Refinement	2.32	1.79681	0.52319	1.36	0.96	2.56	-0.24	1.53567	0.78433	1.34	0.98	1.83	0.49	***
TR944	Refinement	1.84	1.97859	0.13859	1.44	0.4	2.21	-0.37	1.49024	0.34976	1.16	0.68	1.67	0.17	***
TR947	Refinement	0.86	1.71117	0.85117	1.14	-0.28	2.13	-1.27	1.21372	-0.35372	0.84	0.02	1.47	-0.61	***
TR948	Refinement	1.59	1.60567	0.01567	1.03	0.56	2.04	-0.45	1.09707	0.49293	0.85	0.74	1.32	0.27	***
The Cumulative scores		54.63	51.43695	3.19305	38.52	16.11	65.55	10.92	37.465407	17.164593	27.24	27.39	45.18	9.45	

Table S. 6 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 TBM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. H_o: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H_1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 13

		Molprobity score													
CASP TARGET			The original MD-based protocol of ReFOLD						The local quality assessment guided MD-based protocol Threshold is 3 Ångströms						Wilcoxon test
Target ID by domain	CASP Category	Starting model	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max	Mean Score	Diff. Mean	Minimum Score	Diff. Min	Maximum Score	Diff. Max	Significance
T0890	Regular	2.3	1.74699	0.55301	1.23	1.07	2.04	0.26	1.27024	1.02976	0.89	1.41	1.51	0.79	***
T0898	Regular	3.24	2.18706	1.05294	1.74	1.5	2.67	0.57	1.61866	1.62134	1.28	1.96	1.87	1.37	***
T0909	Regular	3.18	2.29656	0.88344	1.97	1.21	2.7	0.48	1.76707	1.41293	1.21	1.97	1.87	1.31	***
TR694	Refinement	2.66	2.04141	0.61859	1.52	1.14	3.27	-0.61	1.68384	0.97616	1.12	1.54	1.83	0.83	***
TR868	Refinement	0.81	1.49301	-0.68301	0.95	-0.14	1.97	-1.16	0.95	-0.14	0.5	0.31	1.33	-0.52	***
TR890	Refinement	2.01	2.04604	-0.03604	1.28	0.73	2.4	-0.39	1.6928	0.3172	1.4	0.61	1.82	0.19	***
TR896	Refinement	2.14	1.84524	0.29476	1.1	1.04	2.31	-0.17	1.31006	0.82994	1	1.14	1.8	0.34	***
TR898	Refinement	0.66	1.33908	-0.67908	0.66	0	2.06	-1.4	0.827988	0.167988	0.5	0.16	1.36	-0.70	***
TR901	Refinement	2.03	1.89329	0.13671	1.38	0.65	2.15	-0.12	1.37665	0.65335	0.98	1.05	1.68	0.35	***
TR909	Refinement	3.26	2.24025	1.01975	1.75	1.51	2.67	0.59	1.62829	1.63171	1.2	2.06	1.83	1.43	***
TR945	Refinement	2.33	1.82177	0.50823	1.35	0.98	2.21	0.12	1.35579	0.97421	1.11	1.22	1.54	0.79	***
The Cumulative scores		24.62	20.9507	3.6693	14.93	9.69	26.45	-1.83	15.481388	9.138612	11.19	13.43	18.44	6.18	

Table S. 7 Performance summary for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 FM/TBM targets according to Molprobity score.

Table 2.7 One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target. H_o: The scores of the models generated by the local quality assessment guided MD-based protocol are equal or lower in quality than those generated by the original ReFOLD protocol. H_1: The scores of the models generated by the local quality assessment guided MD-based protocol are higher quality models than those generated by the original ReFOLD protocol. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 14

CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Minimum vs Starting
FM	**0.01242**	**3.052e-05**	**3.052e-05**	**0.0003624**
TBM	**3.093e-06**	**1.863e-09**	**1.348e-06**	**2.967e-06**
FM/TBM	**0.002089**	**0.0002441**	**0.0002441**	**0.0002441**
ALL	**4.693e-09**	**1.228e-10**	**1.226e-10**	**2.649e-10**

Table S. 8 Calculated pairwise p-values for the local quality assessment guided MD-based protocol versus the original ReFOLD protocol on the CASP12 targets according to Molprobity score.

H_o: The scores of the targets refined by the local quality assessment guided MD-based protocol are equal or lower in quality than those refined by the original MD-based protocol of ReFOLD. H_1: The scores of targets refined by the local quality assessment guided MD-based protocol are higher quality models than those refined by the original MD-based protocol of ReFOLD. The minimum score of the models generated by the local quality assessment guided MD-based protocol were also compared with the score of the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface. lower Molprobity scores are better).

Appendix 15

CASP TARGET			The local quality assessment guided MD-based protocol Threshold is 3 Ångströms		The local quality assessment guided MD-based protocol Threshold is 5 Ångströms		The local quality assessment guided MD-based protocol Threshold is 8 Ångströms	
Target ID by domain	CASP Category	Starting model	Mean Score	Maximum Score	Mean Score	Maximum Score	Mean Score	Maximum Score
T0859	Regular	0.1681	0.17192	0.1814	0.172224	0.1814	0.172508	0.1858
T0862	Regular	0.4032	0.396678	0.414	0.396597	0.422	0.396843	0.4194
T0866	Regular	0.2391	0.241027	0.2565	0.241905	0.2565	0.239828	0.2543
T0880	Regular	0.0596	0.0576659	0.0622	0.0576896	0.0622	0.0573884	0.0609
T0886	Regular	0.1681	0.174943	0.1921	0.176126	0.1921	0.176764	0.1889
T0897	Regular	0.0582	0.0587799	0.062	0.0586067	0.063	0.0588317	0.0639
T0904	Regular	0.2267	0.221043	0.2323	0.221234	0.2307	0.219848	0.2299
T0915	Regular	0.2808	0.273628	0.289	0.274235	0.2922	0.27428	0.2873
TR594	Refinement	0.3427	0.344157	0.3652	0.344991	0.3708	0.346089	0.3708
TR862	Refinement	0.4032	0.394813	0.422	0.391551	0.422	0.394994	0.4194
TR866	Refinement	0.6082	0.602386	0.6346	0.607648	0.6346	0.604145	0.6322
TR869	Refinement	0.2885	0.276501	0.2909	0.276018	0.2885	0.275846	0.2885
TR870	Refinement	0.25	0.257492	0.2729	0.254413	0.2775	0.251428	0.2752
TR905	Refinement	0.3244	0.298757	0.3223	0.2985	0.3202	0.298863	0.3223
The Cumulative Scores		3.8208	3.7697908	3.9974	3.7717383	4.0137	3.7676561	3.9988

Table S. 9 Performance summary for the local quality assessment guided MD-based protocol with the varying threshold on the CASP12 FM targets according to GDT-HA score. (higher GDT-HA scores are better).

Appendix 16

CASP TARGET			The local quality assessment guided MD-based protocol Threshold is 3 Ångströms		The local quality assessment guided MD-based protocol Threshold is 5 Ångströms		The local quality assessment guided MD-based protocol Threshold is 8 Ångströms	
Target ID by domain	CASP Category	Starting model	Mean Score	Maximum Score	Mean Score	Maximum Score	Mean Score	Maximum Score
T0872	Regular	0.4886	0.472564	0.5028	0.473605	0.5057	0.474388	0.5
T0882	Regular	0.5791	0.588816	0.6297	0.588486	0.6234	0.587465	0.6203
T0895	Regular	0.5167	0.501923	0.5292	0.502213	0.5312	0.504132	0.5312
T0911	Regular	0.2972	0.291071	0.3027	0.289913	0.3009	0.290695	0.307
T0913	Regular	0.4149	0.398787	0.4142	0.399257	0.4157	0.396969	0.4149
T0944	Regular	0.5504	0.533065	0.5534	0.531242	0.5573	0.532495	0.5524
T0946	Regular	0.2962	0.27321	0.2954	0.272807	0.2954	0.274359	0.2954
T0947	Regular	0.4814	0.459048	0.4771	0.459429	0.4829	0.456725	0.48
T0948	Regular	0.5235	0.495054	0.5268	0.496905	0.5268	0.499658	0.5235
TR520	Refinement	0.581	0.543085	0.581	0.543085	0.581		
TR872	Refinement	0.5682	0.541712	0.5739	0.543152	0.5767	0.544709	0.5767
TR877	Refinement	0.4894	0.466093	0.4877	0.466093	0.4877		
TR879	Refinement	0.633	0.583705	0.6261	0.583759	0.6239	0.584782	0.625
TR881	Refinement	0.479	0.452469	0.4802	0.452277	0.4777		
TR882	Refinement	0.6899	0.71137	0.75	0.711462	0.7468		
TR885	Refinement	0.7837	0.747322	0.7909	0.747527	0.7885	0.74688	0.7837
TR891	Refinement	0.7567	0.721636	0.7589	0.720423	0.75		
TR893	Refinement	0.6908	0.655179	0.6997	0.652946	0.6982		
TR895	Refinement	0.5146	0.492889	0.5188	0.490015	0.5104		
TR913	Refinement	0.4534	0.436608	0.4586	0.4402	0.4615	0.436949	0.4564
TR917	Refinement	0.6535	0.637371	0.6618	0.637931	0.6586		
TR920	Refinement	0.6039	0.568527	0.6016	0.569101	0.6005	0.568023	0.6039
TR921	Refinement	0.4801	0.465543	0.4928	0.465543	0.4928		
T0922	Refinement	0.7581	0.766796	0.8105	0.766796	0.8105		
TR928	Refinement	0.4274	0.396442	0.4245	0.396137	0.4245	0.398646	0.4245
TR942	Refinement	0.3333	0.320707	0.332	0.322513	0.3359	0.321614	0.3333
TR944	Refinement	0.5603	0.529974	0.5583	0.52953	0.5583	0.529345	0.5593
TR947	Refinement	0.5157	0.511695	0.5329	0.51118	0.5343	0.510245	0.5343
TR948	Refinement	0.5956	0.594791	0.6242	0.592882	0.6225	0.59376	0.6242
The Cumulative Scores		15.7156	15.157452	15.9957	15.156409	15.9796		

Table S. 10 Performance summary for the local quality assessment guided MD-based protocol with the varying threshold on the CASP12 TBM targets according to GDT-HA score (higher GDT-HA scores are better).

Appendix 17

CASP TARGET			The local quality assessment guided MD-based protocol Threshold is 3 Ångströms		The local quality assessment guided MD-based protocol Threshold is 5 Ångströms		The local quality assessment guided MD-based protocol Threshold is 8 Ångströms	
Target ID by domain	CASP Category	Starting model	Mean Score	Maximum Score	Mean Score	Maximum Score	Mean Score	Maximum Score
T0890	Regular	0.1742	0.168949	0.1769	0.168869	0.1769	0.169681	0.1809
T0898	Regular	0.1599	0.158659	0.1693	0.159651	0.1708	0.15673	0.1661
T0899	Regular	0.1628	0.153423	0.1664	0.153287	0.1628	0.15414	0.1635
T0909	Regular	0.2703	0.278384	0.2943	0.27786	0.2905	0.277429	0.2958
T0945	Regular	0.3493	0.34364	0.3587	0.344235	0.3607	0.345565	0.3633
TR694	Refinement	0.2376	0.22704	0.2414	0.224509	0.2395	0.227157	0.2405
TR868	Refinement	0.6143	0.604776	0.6548	0.603366	0.6381		
TR890	Refinement	0.3245	0.305804	0.3271	0.307387	0.3271	0.308702	0.3271
TR896	Refinement	0.468	0.443089	0.4826	0.442043	0.4826	0.446423	0.4826
TR898	Refinement	0.2524	0.24816	0.2618	0.24881	0.2618	0.249028	0.2618
TR901	Refinement	0.3061	0.30088	0.3161	0.301688	0.3139		
TR909	Refinement	0.4257	0.430218	0.4452	0.42827	0.4414	0.43115	0.4474
TR945	Refinement	0.412	0.396554	0.412	0.39668	0.416	0.395927	0.4113
The Cumulative Scores		4.1571	4.059576	4.3066	4.056655	4.2821		

Table S. 11 Performance summary for the local quality assessment guided MD-based protocol with the varying threshold on the CASP12 FM/TBM targets according to GDT-HA score (higher GDT-HA scores are better)

Appendix 18

CASP TARGET			The local quality assessment guided MD-based protocol Threshold is 3 Ångströms		The local quality assessment guided MD-based protocol Threshold is 5 Ångströms		The local quality assessment guided MD-based protocol Threshold is 8 Ångströms	
Target ID by domain	Prediction Method	Starting model	Mean Score	Minimum Score	Mean Score	Minimum Score	Mean Score	Minimum Score
TR520	TBM	1.91	1.28591	0.9				
TR872	TBM	0.5	0.93372	0.5	0.927927	0.5	0.897439	0.5
TR877	TBM	1.41	1.23122	0.77				
TR879	TBM	3.07	1.65665	1.31	1.60463	1.11	1.64085	1.3
TR881	TBM	2.68	1.5203	1.14	1.53646	1.14		
TR882	TBM	0.5	0.681037	0.5				
TR885	TBM	0.82	0.93878	0.56	0.963293	0.5	0.963963	0.5
TR891	TBM	1.56	1.21762	0.81				
TR893	TBM	1.51	1.05195	0.68	1.08378	0.75		
TR895	TBM	2.22	1.21744	0.87	1.2189	0.87		
TR913	TBM	1.34	1.28902	1.02	1.31677	0.95	1.29927	1.01
TR917	TBM	1.36	1.16061	0.87	1.13957	0.84		
TR920	TBM	1.61	1.32872	0.71	1.32317	0.93	1.35988	0.96
TR921	TBM	1.61	1.3475	0.9				
T0922	TBM	1.07	1.23921	0.65				
TR928	TBM	3.56	1.81598	1.48	1.79451	1.59	1.79616	1.59
TR942	TBM	2.32	1.53567	1.34	1.5578	1.34	1.52945	1.37
TR944	TBM	1.84	1.49024	1.16	1.48555	1.21	1.50567	1.32
TR947	TBM	0.86	1.21372	0.84	1.2003	0.94	1.19878	0.86
TR948	TBM	1.59	1.09707	0.85	1.06445	0.85	1.10622	0.81
TR694	FM/TBM	2.66	1.68384	1.12	1.66591	1.35	1.67689	1.4
TR868	FM/TBM	0.81	0.95	0.5	0.933171	0.5		
TR890	FM/TBM	2.01	1.6928	1.4	1.64177	1.27	1.65805	1.35
TR894	FM/TBM							
TR896	FM/TBM	2.14	1.31006	1	1.31	1	1.3003	0.9
TR898	FM/TBM	0.66	0.827988	0.5	0.868902	0.5	0.854146	0.5
TR901	FM/TBM	2.03	1.37665	0.98				
TR909	FM/TBM	3.26	1.62829	1.2	1.62152	1.24	1.63793	1.28
TR945	FM/TBM	2.33	1.35579	1.11	1.3522	0.94	1.33561	0.96
TR594	FM	2.91	1.72317	1.2	1.71494	1.14	1.68695	1.14
TR862	FM	2.26	1.24848	0.78	1.27518	0.64	1.29433	0.79
TR866	FM	1.56	1.45	1.07	1.41488	1.03	1.4503	1.07
TR869	FM	1.98	1.36421	0.94	1.31354	0.98	1.34043	0.9
TR870	FM	3.61	1.55659	1.07	1.52744	1.08	1.57677	1.19

Table S. 12 Performance summary for the local quality assessment guided MD-based protocol with the varying threshold on the CASP12 targets according to Molprobity Score (lower Molprobity Score are better).

Appendix 19

	GDT-HA score									
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			The percentage of the improved modes		Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Fixed	Gradual	Significance
T0953s2	0.1321	0.128	0.136705	0.1452	0.126	0.134213	0.1421	19.51219512	70.12195122	***
T0955	0.6707	0.622	0.665286	0.6951	0.622	0.664013	0.7073	23.7804878	23.7804878	n.s.
T0957s1	0.2191	0.2037	0.214349	0.2299	0.2037	0.214104	0.2284	7.926829268	7.926829268	n.s.
T0958	0.4481	0.4286	0.452396	0.4773	0.4253	0.450811	0.4838	6.707317073	56.09756098	n.s.
T0960	0.1457	0.137	0.145227	0.1531	0.1357	0.145476	0.1524	35.97560976	42.68292683	n.s.
T0963	0.1511	0.1401	0.146997	0.1566	0.1387	0.148177	0.1587	4.268292683	15.24390244	***
T0981	0.11	0.1058	0.109509	0.1129	0.1042	0.109356	0.1149	33.53658537	35.36585366	n.s.
T0984	0.319	0.3008	0.311644	0.3222	0.2996	0.310373	0.3194	3.048780488	1.219512195	n.s.
T0992	0.6355	0.6005	0.627206	0.6542	0.5981	0.624749	0.6565	16.46341463	14.02439024	n.s.
T1005	0.2991	0.2745	0.29365	0.3037	0.2784	0.292325	0.3044	13.41463415	8.536585366	n.s.
T1022s1	0.2814	0.2601	0.275921	0.2892	0.2635	0.275068	0.2915	10.36585366	9.756097561	n.s.
The Cumulative Scores	3.4118	3.2011	3.37889	3.5394	3.1952	3.368665	3.5594	15.90909091	25.88691796	

Table S. 13 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 FM/TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 20

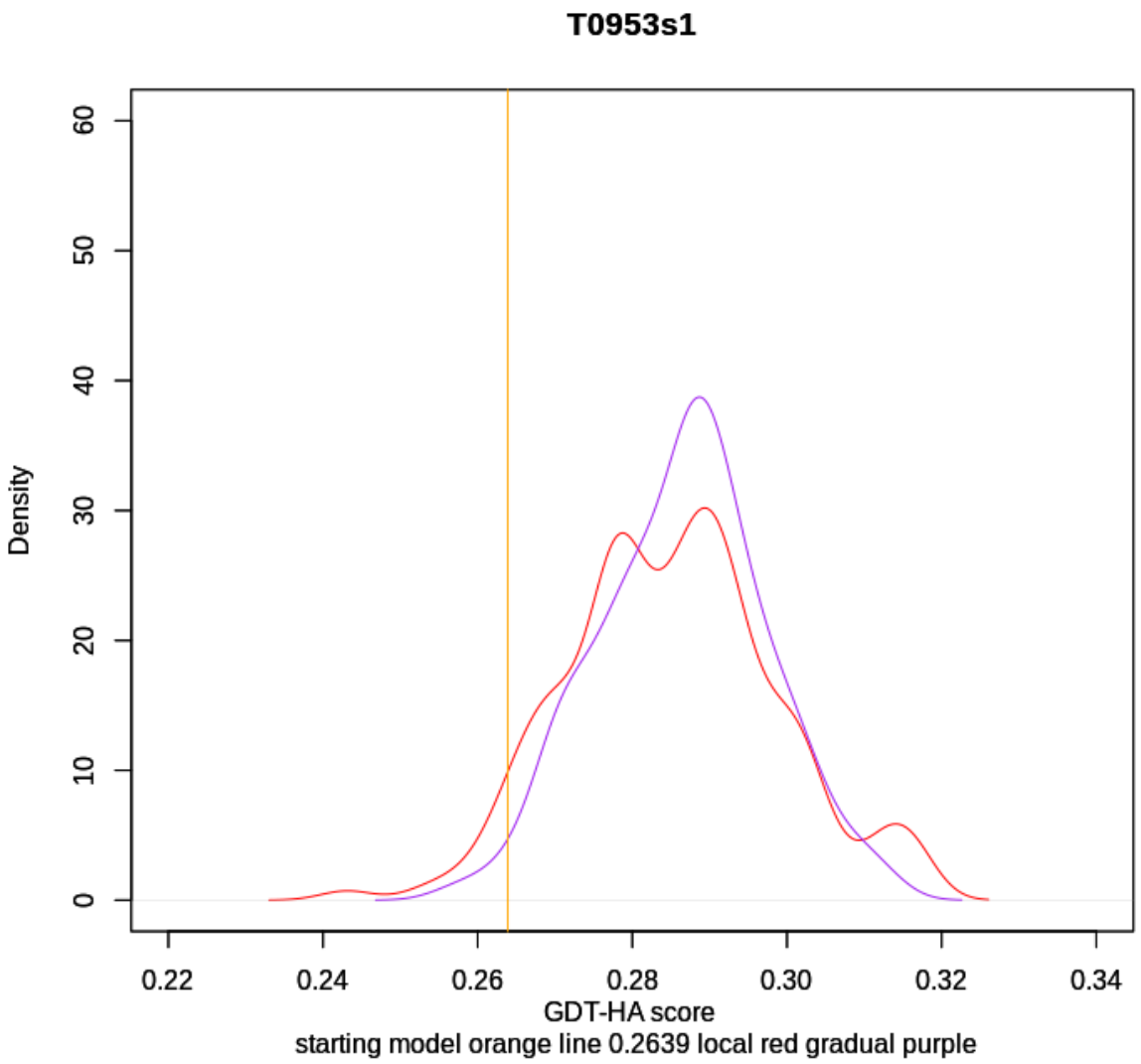

Figure S. 7 A comparison of the gradual restraint strategy and fixed restraint strategy on an FM/TBM target.

Performance of methods on T0953s1 (an FM/TBM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.2639, and higher GDT HA scores are better)

Appendix 21

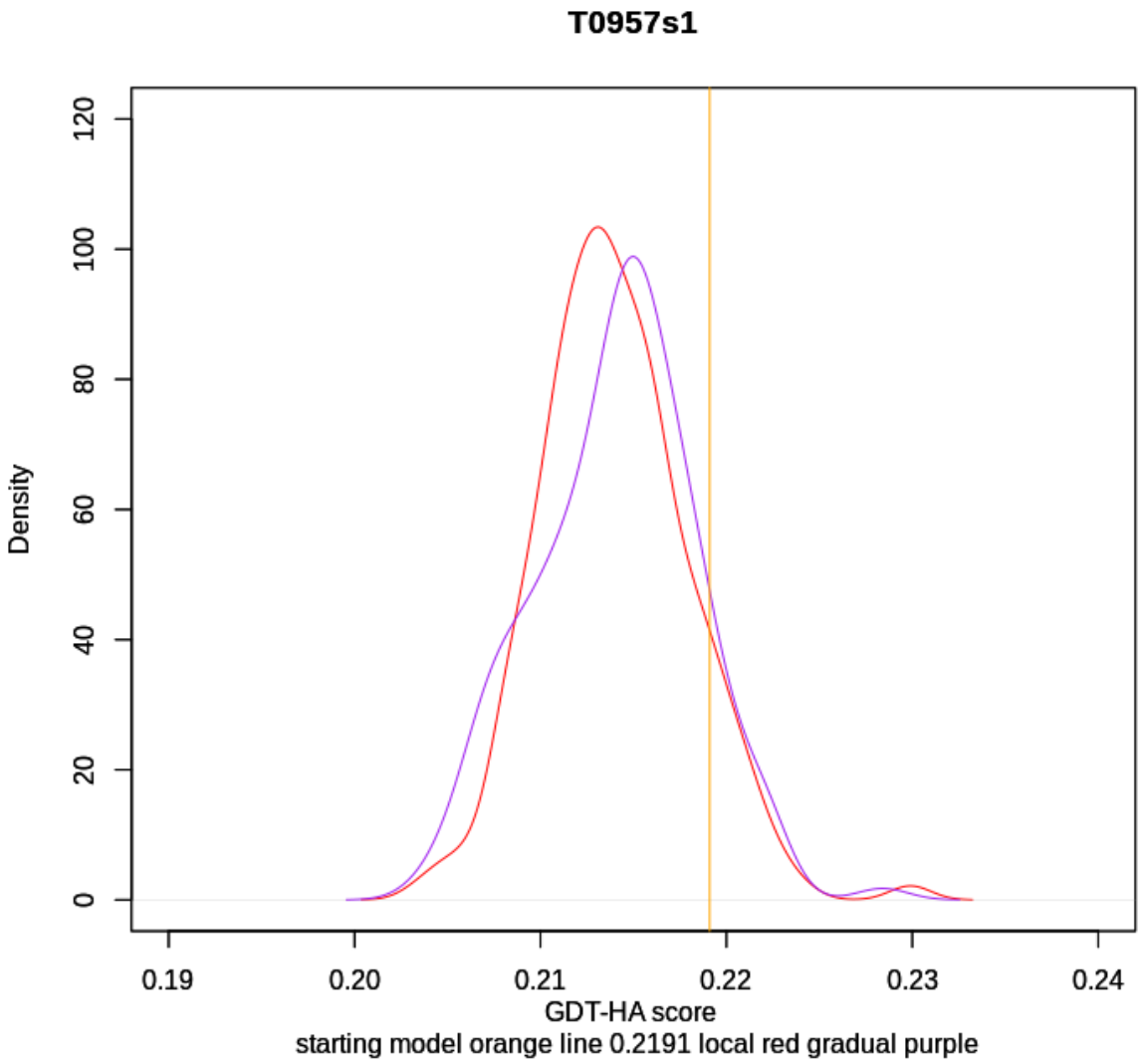

Figure S. 8 A comparison of the gradual restraint strategy and fixed restraint strategy on an FM/TBM target.

Performance of methods on T0957s1 (an FM/TBM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.2191, and higher GDT HA scores are better)

Appendix 22

	GDT-HA score									
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			The percentage of the improved modes		Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Fixed	Gradual	Significance
T1011	0.3598	0.326	0.341817	0.3637	0.3271	0.342325	0.3643	2.43902439	2.43902439	*
T1015s2	0.4632	0.4089	0.445853	0.4729	0.3992	0.446031	0.4922	9.756097561	10.36585366	n.s.
T1021s1	0.448	0.4211	0.448493	0.4782	0.4279	0.449331	0.4748	50.6097561	50.6097561	n.s.
T1021s2	0.4678	0.4427	0.459936	0.4721	0.442	0.459395	0.4742	6.707317073	9.146341463	n.s.
T1022s2	0.3626	0.3537	0.363856	0.374	0.3547	0.363826	0.3745	53.65853659	57.92682927	n.s.
T0974s1	0.6558	0.6413	0.676316	0.7283	0.6232	0.67778	0.7319	87.19512195	88.41463415	n.s.
T0993s1	0.4601	0.4344	0.454698	0.4715	0.4392	0.455711	0.4743	21.34146341	31.70731707	*
T0993s2	0.4796	0.4337	0.458144	0.4847	0.4362	0.461614	0.4949	2.43902439	6.097560976	**
T0995	0.5442	0.5331	0.559068	0.582	0.541	0.557066	0.5797	96.95121951	98.7804878	*
T1004	0.4087	0.3791	0.394914	0.4125	0.381	0.396075	0.4175	4.268292683	6.707317073	n.s.
T1013	0.6422	0.6078	0.631066	0.6534	0.6052	0.629155	0.6491	13.41463415	9.756097561	*
T1016	0.6015	0.5743	0.598407	0.63	0.5743	0.5955	0.6262	35.97560976	37.80487805	n.s.
The Cumulative Scores	5.8935	5.5561	5.832568	6.1233	5.551	5.833809	6.1536	32.06300813	34.14634146	

Table S. 14 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 23

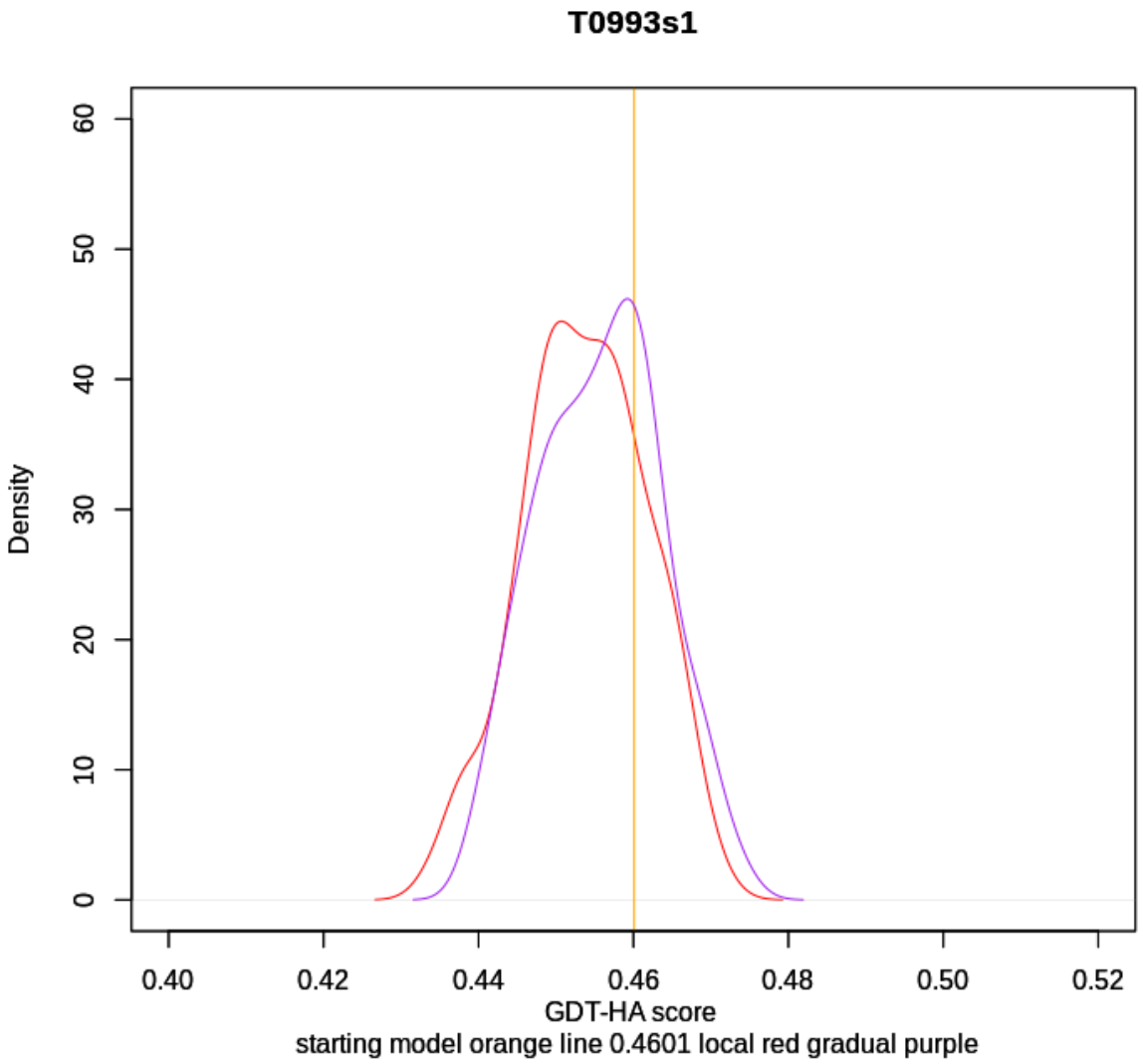

Figure S. 9 A comparison of the gradual restraint strategy and fixed restraint strategy on a TBM target.

Performance of methods on T0993s1 (a TBM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.3598, and higher GDT HA scores are better)

Appendix 24

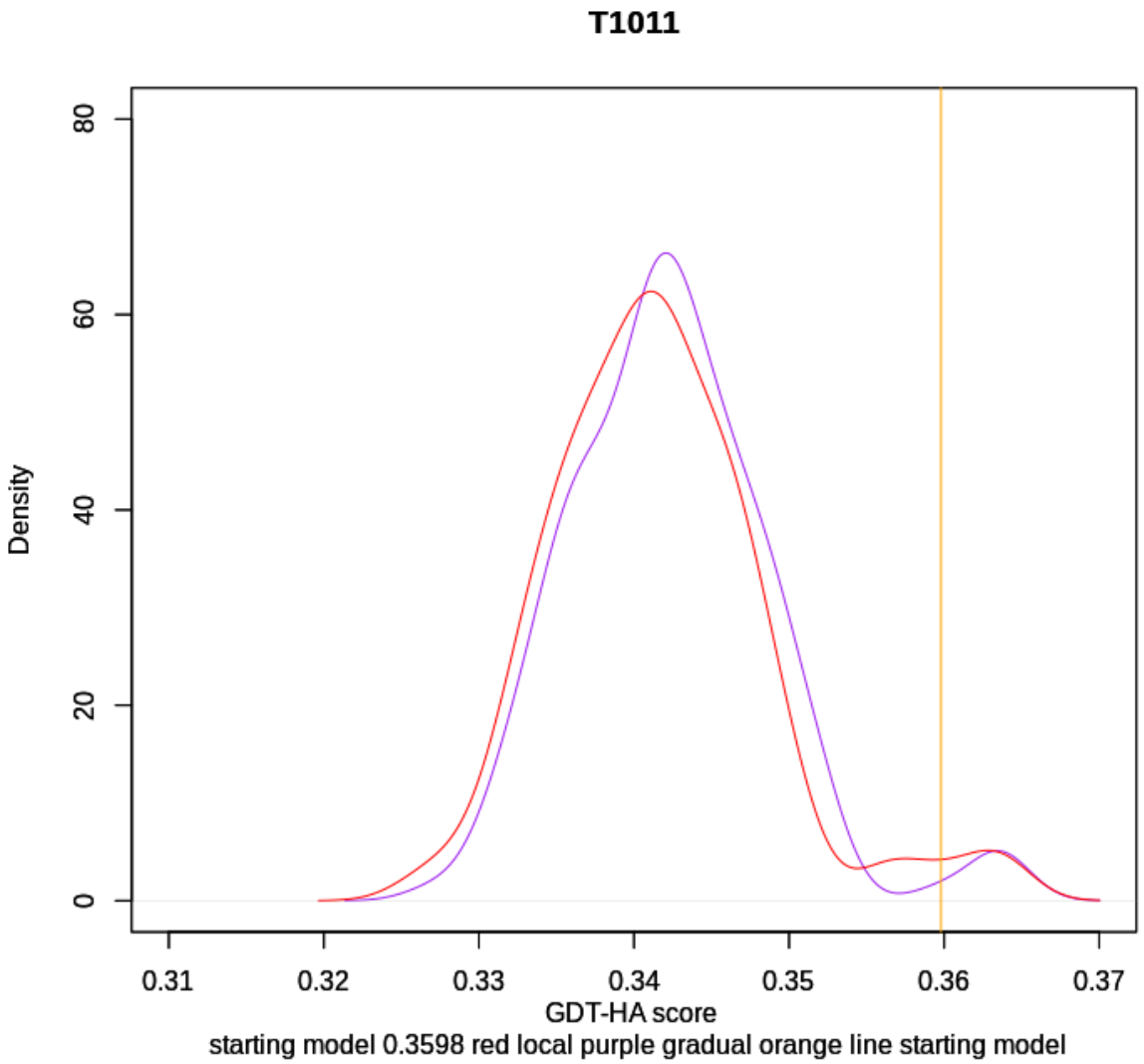

Figure S. 10 A comparison of the gradual restraint strategy and fixed restraint strategy on a TBM target.

Performance of methods on T1011 (a TBM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.3598, and higher GDT HA scores are better)

Appendix 25

	GDT-HA score									
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			The percentage of the improved modes		Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Fixed	Gradual	Significance
T0950	0.1842	0.1798	0.188403	0.1966	0.1791	0.188167	0.1952	86.58536585	85.36585366	n.s.
T0953s1	0.2639	0.2431	0.284723	0.316	0.2569	0.286506	0.3125	95.73170732	98.7804878	n.s.
T0975	0.3256	0.2998	0.316833	0.3292	0.3007	0.316083	0.331	5.487804878	5.487804878	*
T0987	0.1247	0.1175	0.122284	0.1273	0.1161	0.12283	0.1306	17.07317073	18.29268293	**
T0989	0.1524	0.1413	0.151349	0.1596	0.1413	0.149756	0.1575	6.707317073	17.07317073	***
T0991	0.1462	0.1314	0.143451	0.1547	0.1398	0.143461	0.1547	14.63414634	21.34146341	n.s.
T1001	0.5126	0.5018	0.522794	0.545	0.4982	0.523411	0.5504	87.19512195	87.19512195	n.s.
T1010	0.2083	0.1881	0.200436	0.2119	0.1881	0.199572	0.2119	1.829268293	4.87804878	n.s.
T1015s1	0.1989	0.1903	0.201427	0.2216	0.1875	0.200058	0.2216	48.7804878	66.46341463	*
The Cumulative Scores	2.1168	1.9931	2.1317	2.2619	2.0077	2.129844	2.2654	40.44715447	44.98644986	

Table S. 15 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 FM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 26

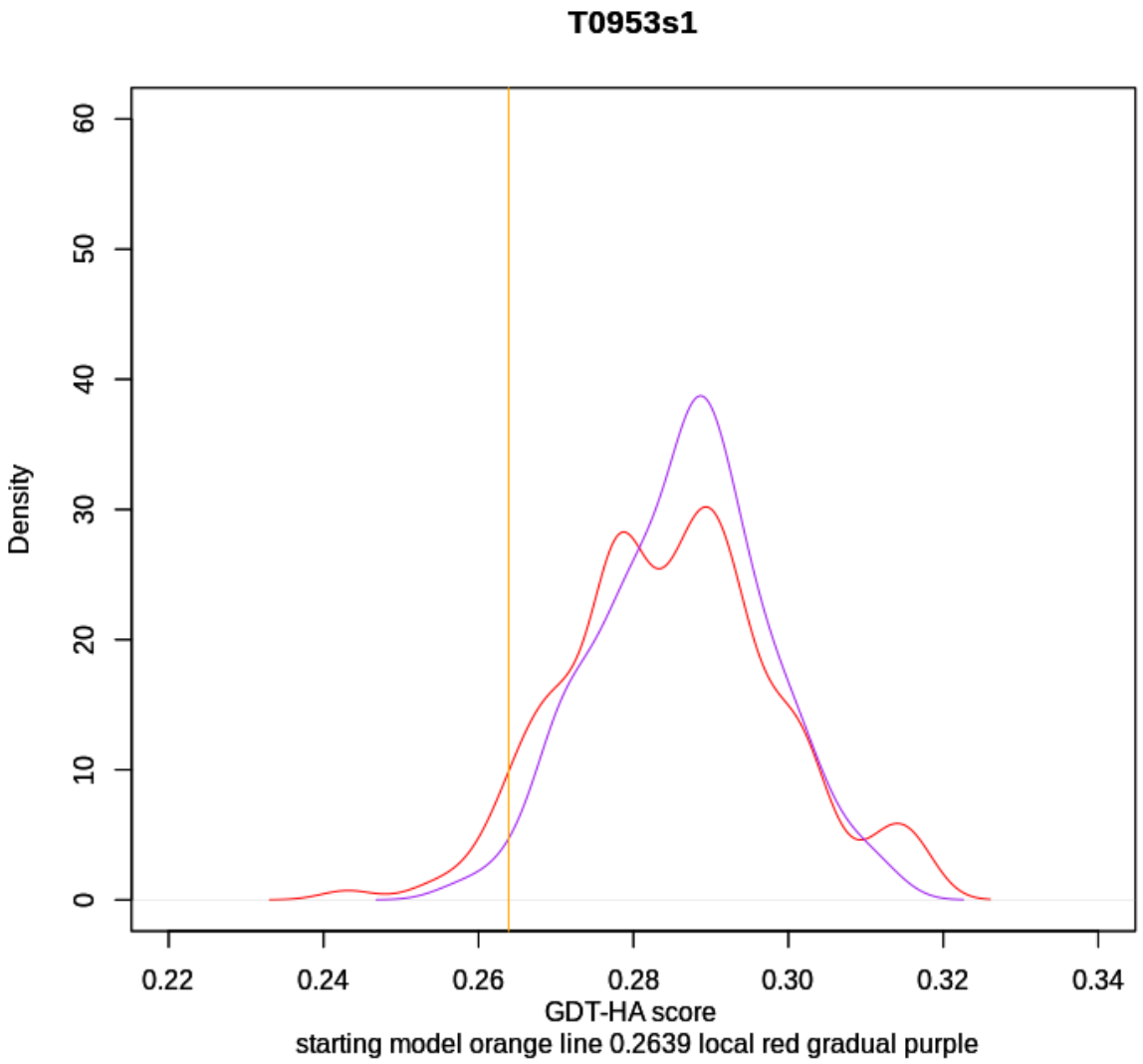

Figure S. 11 A comparison of the gradual restraint strategy and fixed restraint strategy on an FM target.

Performance of methods on T0953s1 (an FM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.3598, and higher GDT HA scores are better)

Appendix 27

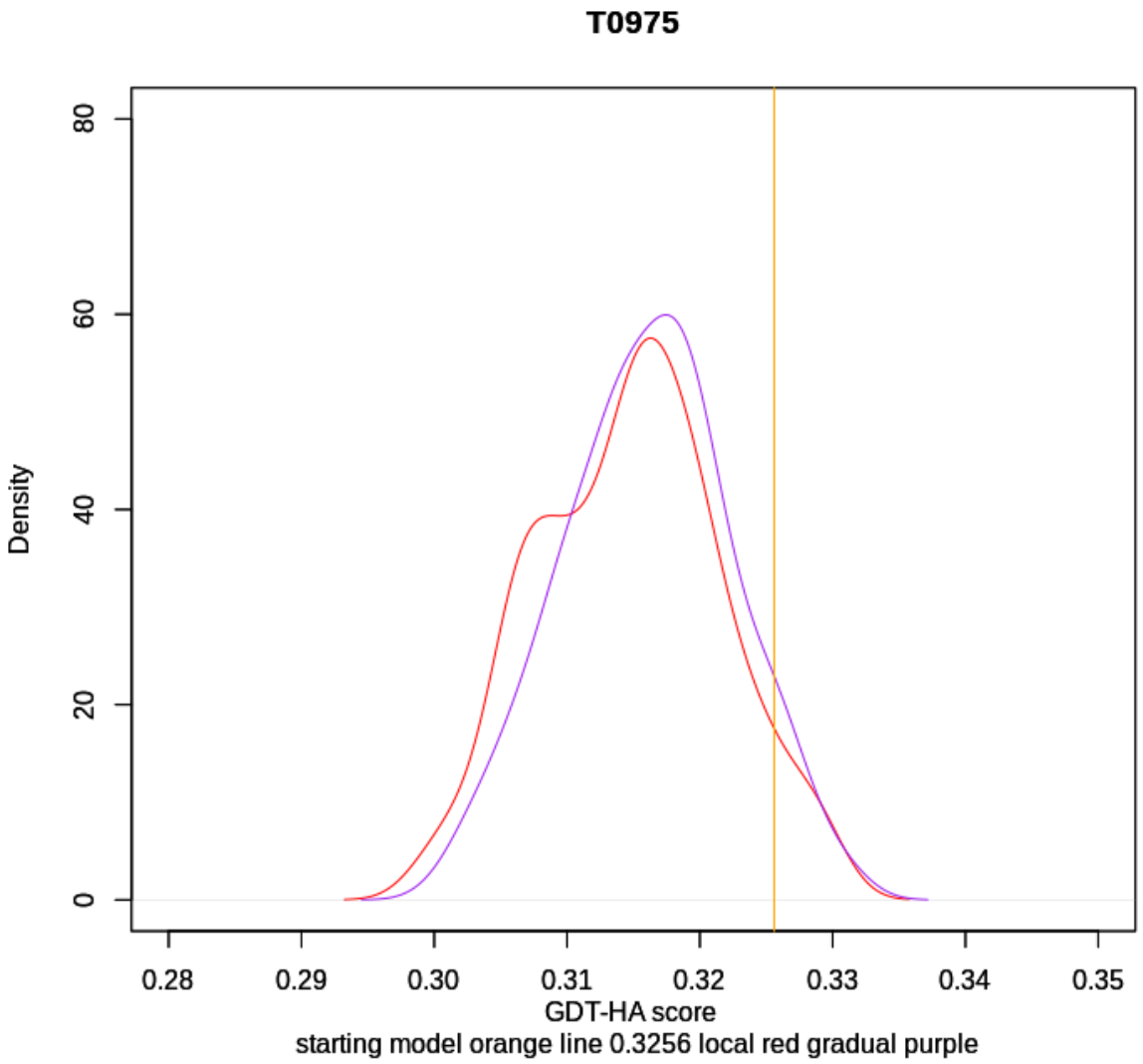

Figure S. 12 A comparison of the gradual restraint strategy and fixed restraint strategy on an FM target.

Performance of methods on T0975 (an FM category CASP13 target) according to GDT-HA score. The purple line represents the gradual restraint models, the red line represents the fixed restraint models, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure is 0.3598, and higher GDT HA scores are better)

Appendix 28

CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Maximum vs Starting
FM	0.2643	0.8623	0.2234	**0.0009766**
TBM	0.4844	0.3177	0.1082	**0.0001221**
FM/TBM	0.8689	0.9954	0.09768	**0.001258**
ALL	0.541	0.9595	**0.02171**	**1.293e-07**

Table S. 16 Calculated pairwise p-values for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 targets according to GDT-HA score.

Ho: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences. The maximum score of the models generated by the gradual restraint strategy were also compared with the score of the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface, higher GDT-HA scores are better).

Appendix 29

	Molprobity Score							
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Mean Score	Minimum Score	Maximum Score	Mean Score	Minimum Score	Maximum Score	Significance
T1011	2.79	1.12	1.55383	1.93	1.06	1.39753	1.58	*
T1015s2	2.75	1.35	1.97463	2.51	1.4	1.80527	2.08	*
T1021s1	1.03	0.87	1.38191	1.63	0.78	1.26234	1.69	**
T1021s2	1.26	0.97	1.52043	1.97	0.99	1.40012	1.63	**
T1022s2	1.27	0.99	1.51787	1.83	1.01	1.37766	1.63	**
T0974s1	0.92	0.5	1.145	1.66	0.5	1.06623	1.59	n.s.
T0993s1	0.73	0.56	1.2483	1.68	0.56	1.16922	1.51	n.s.
T0993s2	3.05	1.07	1.70585	2.16	0.88	1.54862	1.87	***
T0995	3.18	1.18	1.67777	2.08	1.16	1.52369	1.74	***
T1004	0.81	0.85	1.19484	1.7	0.85	1.10476	1.41	*
T1013	2.76	1.36	1.71824	1.99	1.38	1.57641	1.83	*
T1016	1.26	0.81	1.2208	1.55	0.67	1.09768	1.5	*
The Cumulative Scores	21.81	11.63	17.85947	22.69	11.24	16.32953	20.06	

Table S. 17 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 TBM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 30

	Molprobity Score							
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Mean Score	Minimum Score	Maximum Score	Mean Score	Minimum Score	Maximum Score	Significance
T0950	1.06	0.82	1.32861	1.84	0.82	1.25217	1.47	n.s.
T0953s1	2.8	1.01	1.69016	2.57	0.8	1.48572	2	*
T0975	3.82	1.84	2.21681	2.89	1.85	2.00593	2.26	n.s.
T0987	3.74	1.79	2.15803	2.68	1.74	1.96331	2.28	n.s.
T0989	3.31	1.58	2.10321	2.55	1.53	1.88849	2.29	n.s.
T0991	2.86	1.51	1.97899	2.51	1.47	1.84719	2.11	*
T1001	1.17	0.5	1.07787	1.43	0.5	0.996527	1.48	*
T1010	3.49	1.7	2.17787	2.76	1.59	1.97898	2.28	*
T1015s1	0.5	0.57	0.933404	1.55	0.5	0.861497	1.44	*
The Cumulative Scores	22.75	11.32	15.664954	20.78	10.8	14.279814	17.61	

Table S. 18 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 FM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 31

	Molprobity Score							
CASP TARGETS		The fixed restraint strategy			The gradual restraint strategy			Wilcoxon Test gradual vs fixed
Target ID by domain	Starting model	Mean Score	Minimum Score	Maximum Score	Mean Score	Minimum Score	Maximum Score	Significance
T0953s2	3.36	1.75	2.13267	2.48	1.57	1.90108	2.25	***
T0955	2.29	0.86	1.41527	1.72	0.86	1.21443	1.73	n.s.
T0957s1	0.64	0.5	0.978032	1.4	0.5	0.866506	1.27	***
T0958	0.85	0.54	1.05612	1.9	0.54	0.979281	1.69	*
T0960	1.13	0.83	1.42101	2.05	0.89	1.27795	1.56	***
T0963	3.16	1.65	2.02383	2.44	1.63	1.84838	2.08	***
T0981	3.76	1.95	2.2691	2.84	1.92	2.0403	2.38	**
T0984	3.51	1.55	1.88665	2.2	1.51	1.68783	2.07	**
T0992	1	0.88	1.29766	1.6	0.8	1.27084	1.68	n.s.
T1005	3.35	1.73	2.10399	2.52	1.71	1.93018	2.2	*
T1022s1	2.67	1.12	1.74436	2.33	1.08	1.55934	1.85	*
The Cumulative Scores	25.72	13.36	18.328692	23.48	13.01	16.576117	20.76	

Table S. 19 Performance summary for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 FM/TBM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the restraint strategies for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H_1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 32

CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Minimum vs Starting
FM	**0.0009766**	**0.01035**	**0.001953**	**0.009766**
TBM	**0.0001221**	0.06265	**0.00132**	**0.07324**
FM/TBM	**0.0002441**	**0.02879**	**0.001221**	**0.0105**
ALL	**2.328e-10**	**0.0009091**	**1.469e-06**	**0.0004186**

Table S. 20 Calculated pairwise p-values for the gradual restraint strategy versus the fixed restraint strategy on the CASP13 targets according to Molprobity score.

Ho: The scores of the models generated by the gradual restraint strategy are equal or lower in quality than those generated by the fixed restraint strategy. H1: The scores of the models generated by the gradual restraint strategy are higher quality models than those generated by the fixed restraint strategy. P-values ≤ 0.05 indicate significant statistical differences. The minimum score of the models generated by the gradual restraint strategy were also compared with the score of the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface, and lower Molprobity scores are better).

Appendix 33

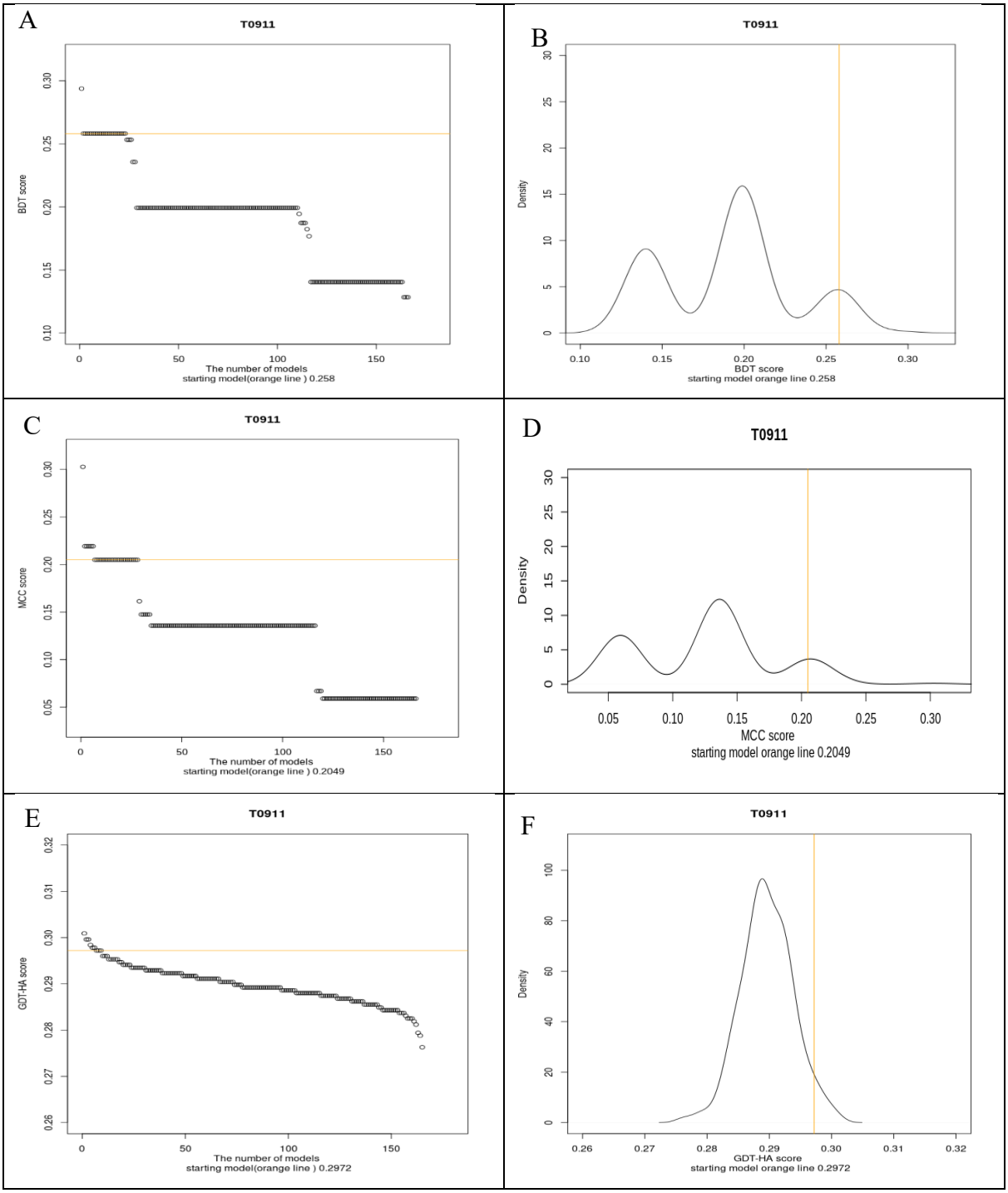

Figure S. 13 The performance of the binding site-focused MD-based protocol for T0911 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

Appendix 34

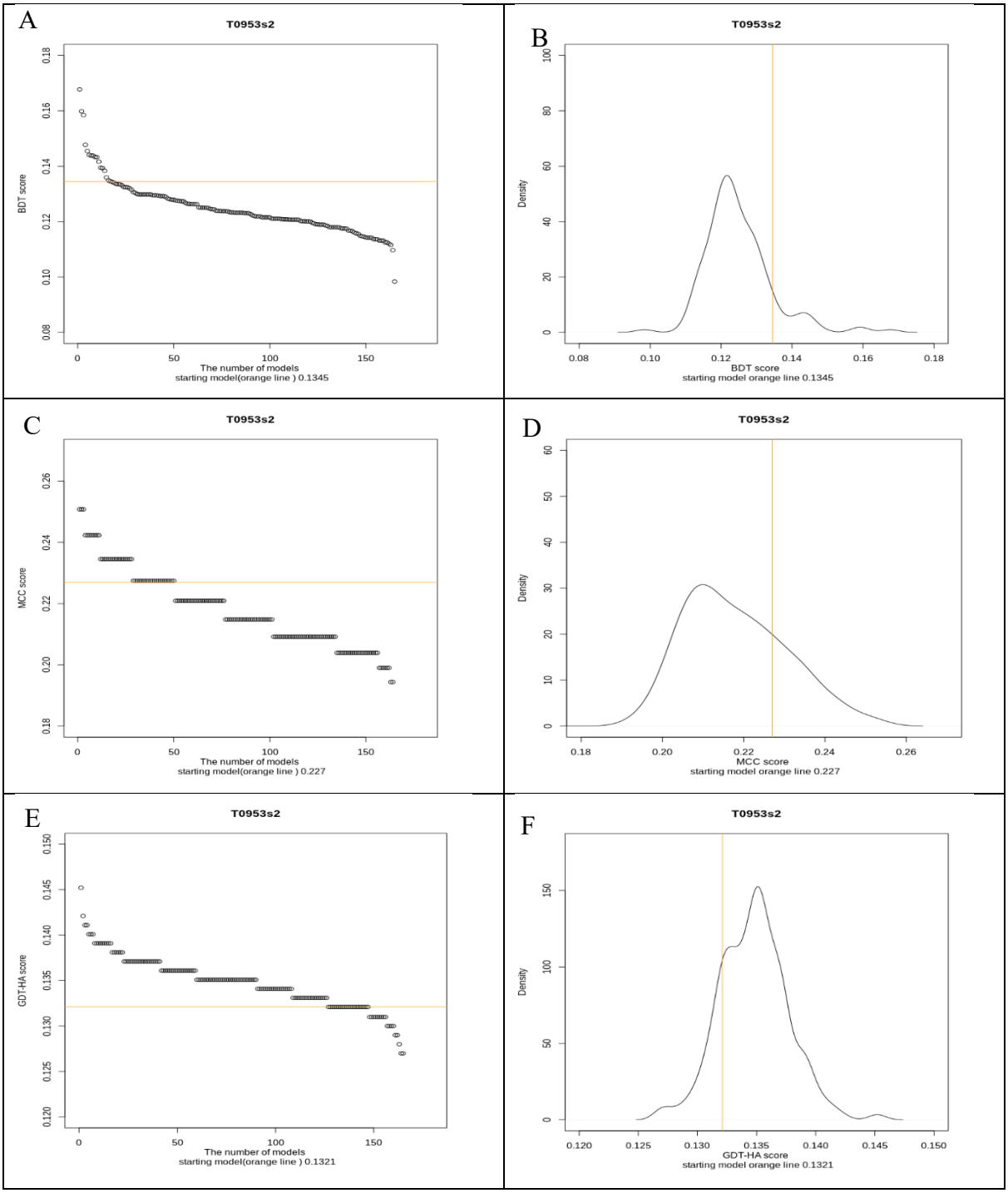

Figure S. 14 Figure 4. 6 The performance of the binding site-focused MD-based protocol for T0953s2 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

Appendix 35

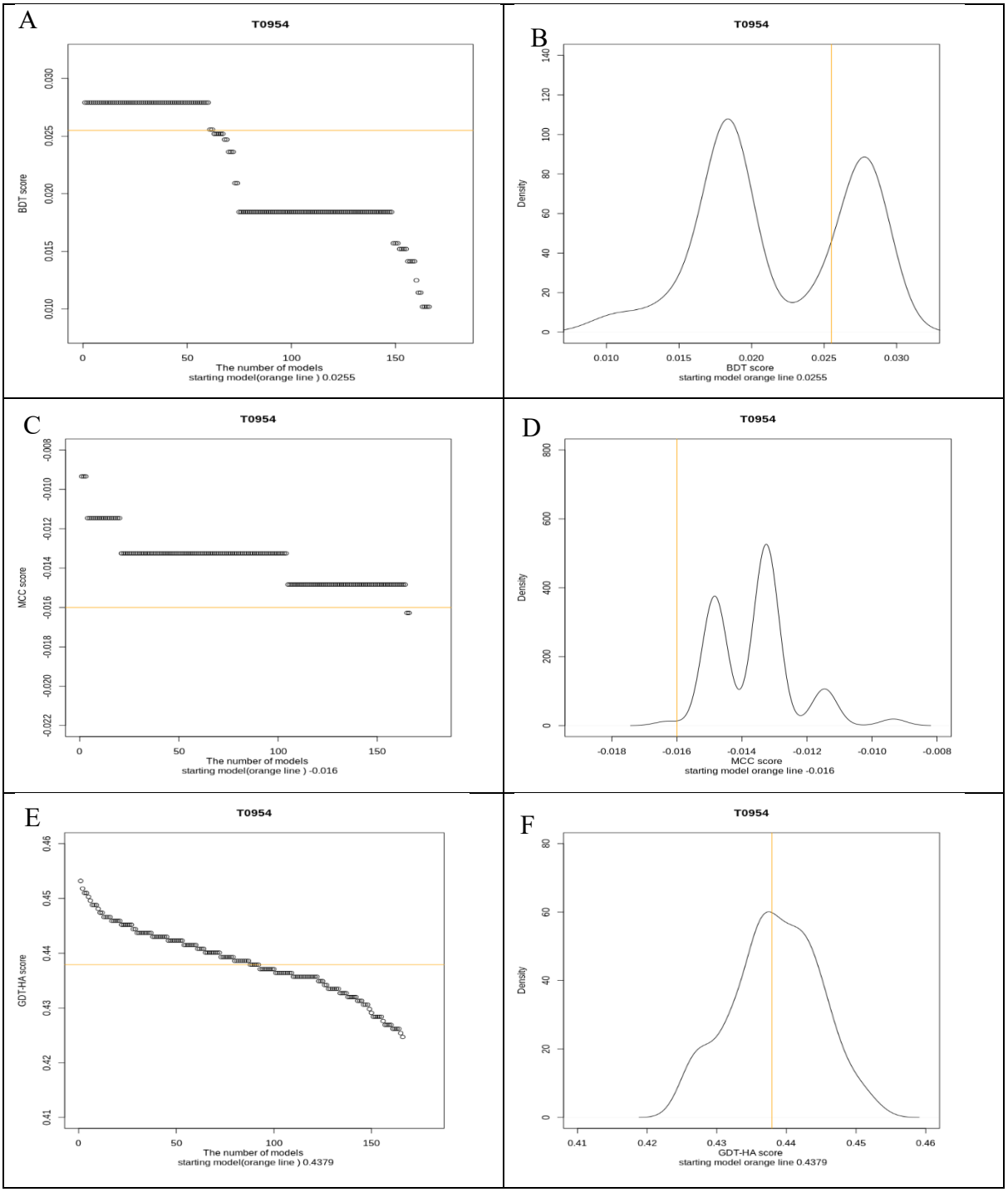

Figure S. 15 Figure 4. 6 The performance of the binding site-focused MD-based protocol for T0954 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

Appendix 36

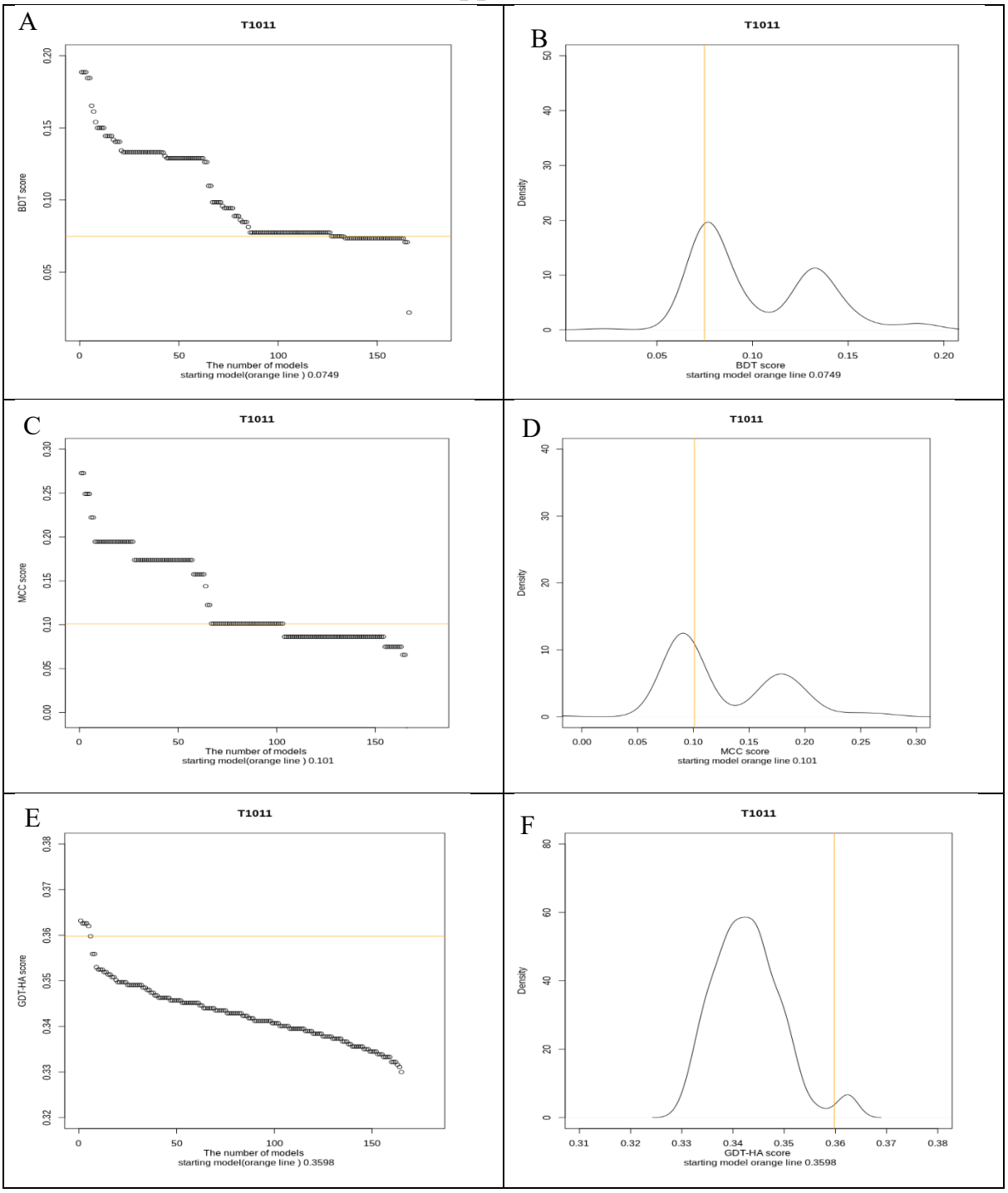

Figure S. 16 The performance of the binding site-focused MD-based protocol for T1011 models.

(A) The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

Appendix 37

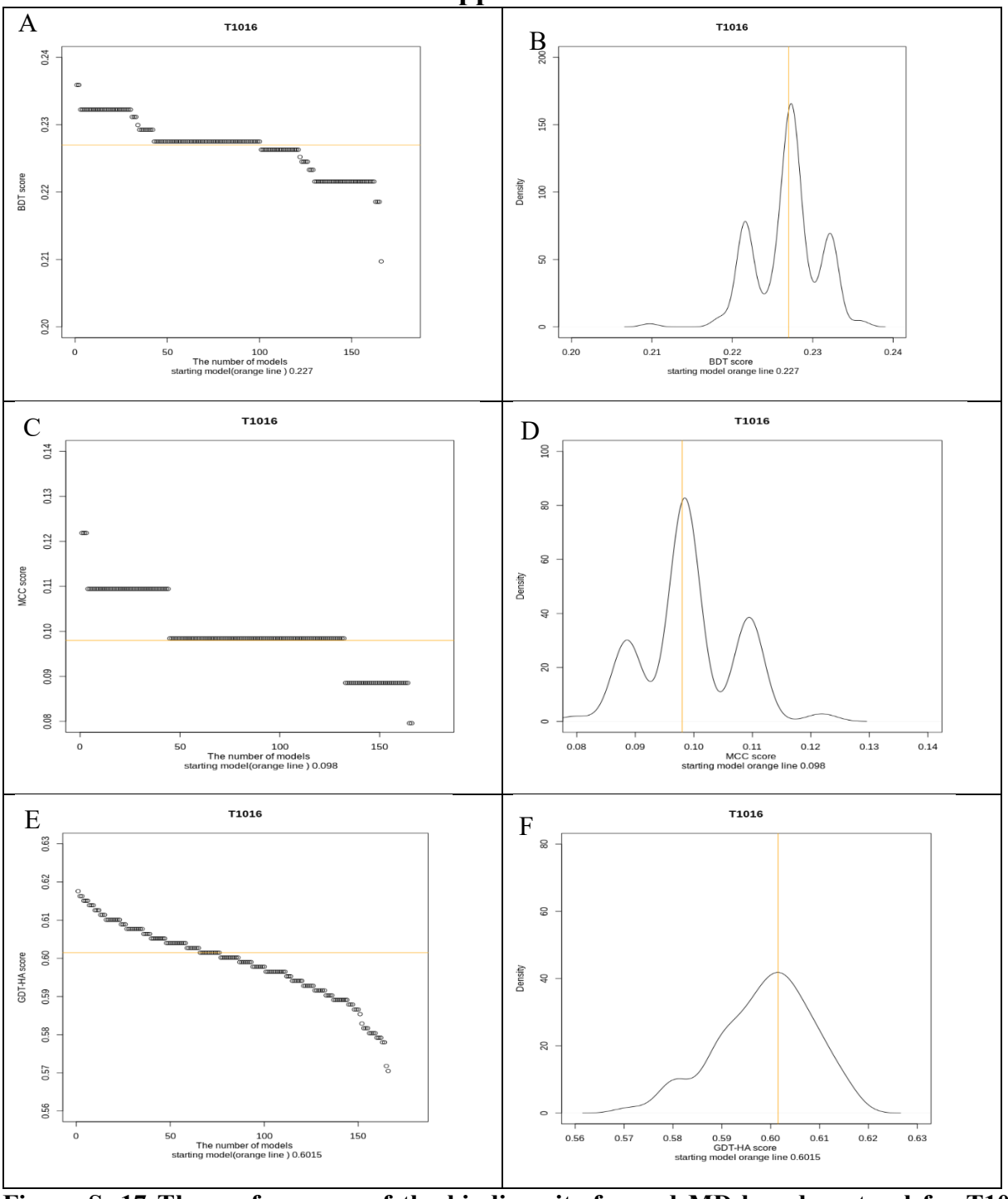

Figure S. 17 The performance of the binding site-focused MD-based protocol for T1016 models.

(A)The black points represent the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (B) the black line represents the BDT scores of 3D models generated by the binding site-focused MD-based protocol and the orange line represents the starting model score. (C) and (D) ditto but according to the MCC score. (E) and (F) ditto according to GDT-HA score (higher scores are better)

Appendix 38

	GDT-HA score													Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			The percentage of the improved models			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact	Local	ReFOLD	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0954	0.4379	0.4254	0.4382612	0.4518	0.4196	0.4389604	0.4569	0.3969	0.428642	0.4547	47.56	48.17	18.29	n.s.	***	***
T0959	0.4405	0.4312	0.449549	0.4722	0.4299	0.449423	0.4749	0.4087	0.458736	0.5278	84.14	81.098	79.26	n.s.	***	***
T0964	0.5947	0.5421	0.574515	0.6079	0.5474	0.575708	0.6263	0.5105	0.56555	0.6342	6.703	4.87	10.97	n.s.	***	***
T0965	0.4073	0.3826	0.399764	0.4161	0.385	0.399831	0.4161	0.3339	0.367456	0.4026	9.75	17.07	0	n.s.	***	***
T0966	0.3821	0.3653	0.378603	0.3892	0.3674	0.379289	0.3892	0.3171	0.359113	0.3963	24.39	26.21	6.09	n.s.	**	***
T1011	0.3598	0.3266	0.342546	0.3643	0.326	0.341817	0.3637	0.29	0.318816	0.3643	2.439	2.43	2.43	*	***	***
T1015s2	0.4632	0.4128	0.443472	0.4787	0.4089	0.445853	0.4729	0.4167	0.46261	0.5039	7.92	9.75	46.34	n.s.	**	***
T1021s2	0.4678	0.4484	0.460195	0.4756	0.4427	0.459936	0.4721	0.4019	0.452318	0.4928	11.58	6.70	15.85	n.s.	***	***
T1022s2	0.3626	0.3522	0.364468	0.377	0.3537	0.363856	0.374	0.3135	0.348521	0.3765	64.63	53.65	15.85	*	***	***
T0973	0.3067	0.2872	0.307886	0.328	0.2872	0.307517	0.3245	0.2713	0.304115	0.3369	51.21	48.17	39.63	n.s.	**	**
T0974s1	0.6558	0.6486	0.679824	0.721	0.6413	0.676316	0.7283	0.5507	0.682384	0.7754	89.02	87.19	75.60	n.s.	n.s.	n.s.
T0993s1	0.4601	0.4306	0.456712	0.4762	0.4344	0.454698	0.4715	0.4192	0.4603	0.5057	34.75	21.34	46.95	n.s.	*	**
T0993s2	0.4796	0.4362	0.460664	0.4898	0.4337	0.458144	0.4847	0.3827	0.459719	0.5153	6.70	2.43	25.60	**	n.s.	n.s.
T0995	0.5442	0.5347	0.559795	0.5828	0.5331	0.559068	0.582	0.5	0.535818	0.5686	98.17	96.95	23.17	**	***	**
T1004	0.4087	0.3778	0.395595	0.4181	0.3791	0.394914	0.4125	0.3249	0.367283	0.4162	5.48	4.26	3.65	n.s.	***	***
T1006	0.8701	0.789	0.833411	0.8831	0.7857	0.835857	0.8831	0.7078	0.810686	0.8961	4.87	6.70	6.09	n.s.	***	***
T1013	0.6422	0.6052	0.629155	0.6491	0.6078	0.631066	0.6534	0.5241	0.58507	0.65	10.36	13.41	2.43	n.s.	***	***
T1016	0.6015	0.5743	0.596896	0.625	0.5743	0.598407	0.63	0.5743	0.597623	0.6262	29.87	35.97	31.70	n.s.	n.s.	n.s.
T1017s1	0.55	0.4818	0.504881	0.5568	0.4727	0.510317	0.5568	0.4295	0.497176	0.5682	2.43	1.82	4.26	*	***	***
T1020	0.4625	0.4585	0.475152	0.4936	0.4561	0.4738976	0.4928	0.3994	0.4504315	0.508	96.95	94.51	32.92	*	***	***
The Cumulative scores	9.8973	9.3105	9.7513442	10.2563	9.286	9.754875	10.2657	8.4731	9.5123675	10.5197						

Table S. 21 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 39

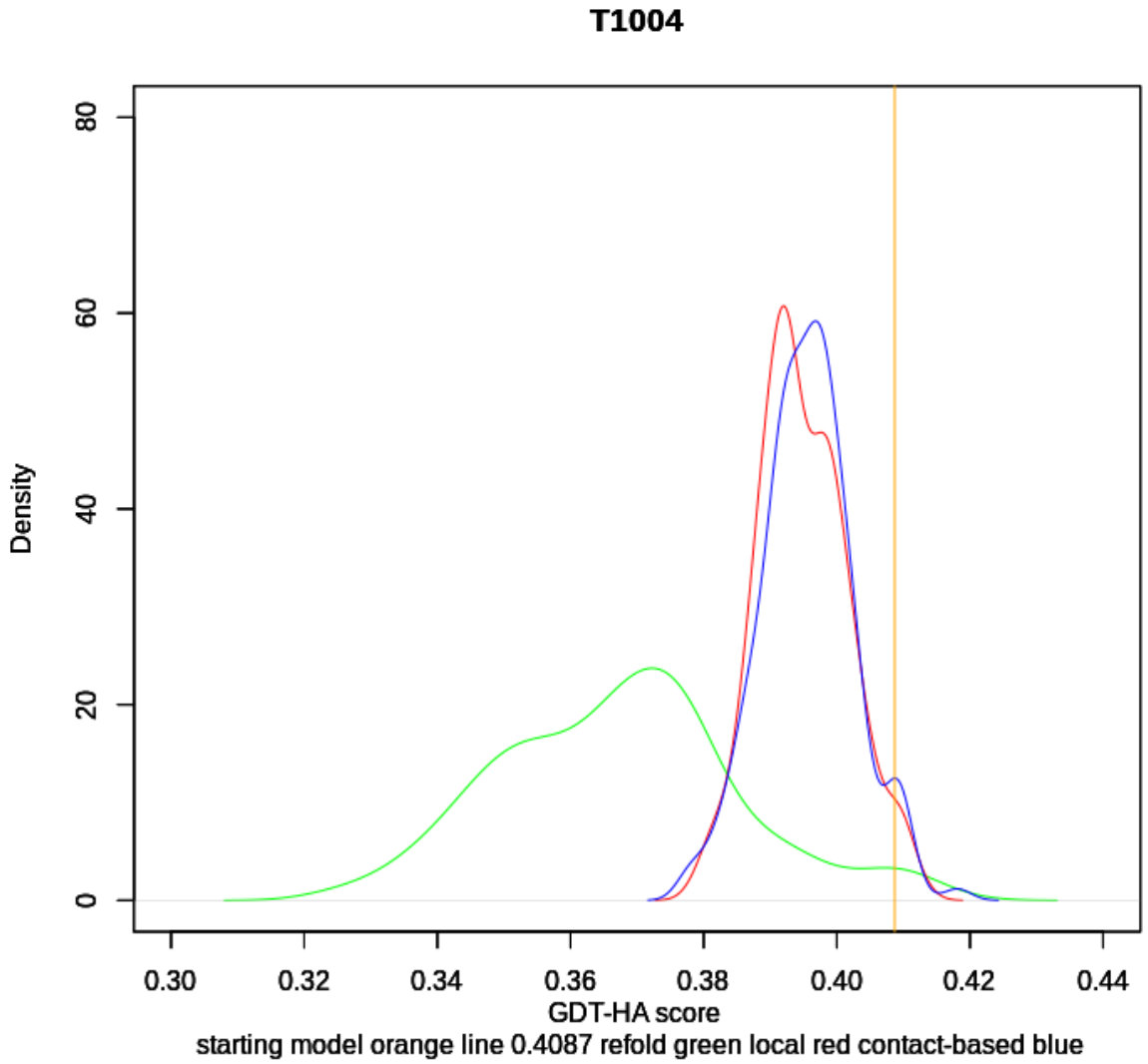

Figure S. 18 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on a TBM target.

Performance of methods on T1004 (a TBM-easy CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.4087 and higher GDT HA scores are better).

Appendix 40

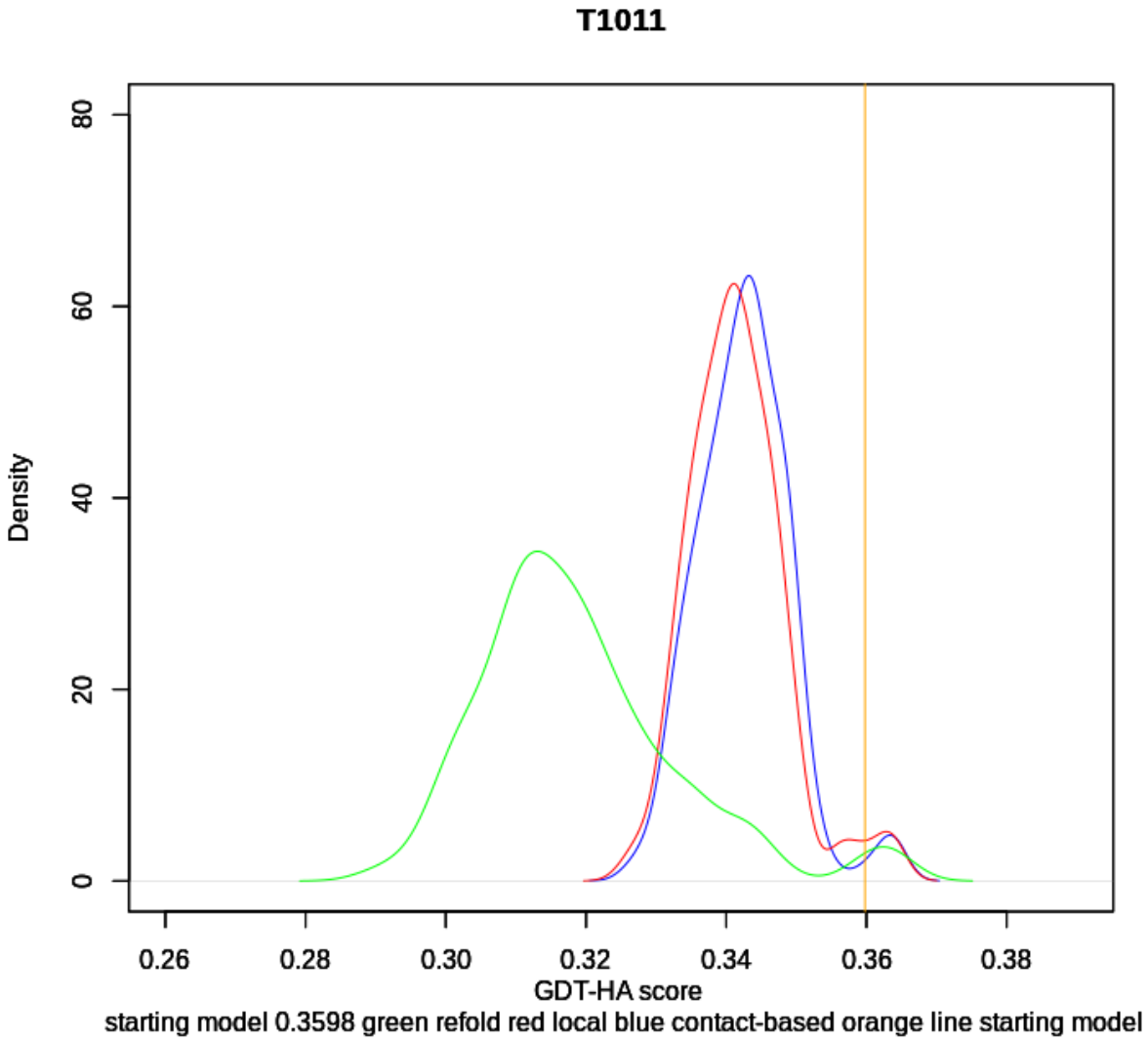

Figure S. 19 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on a TBM target.

Performance of methods on T1011 (a TBM-hard CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.3598 and higher GDT HA scores are better).

Appendix 41

	GDT-HA score													Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			The percentage of the improved models			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact	Local	ReFOLD	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0953s2	0.1321	0.129	0.135504	0.1442	0.128	0.136705	0.1452	0.1149	0.127758	0.1442	79.26	19.51	23.78	*	**	***
T0955	0.6707	0.622	0.663787	0.7012	0.622	0.665286	0.6951	0.5488	0.62531	0.6951	25	23.78	6.70	n.s.	**	***
T0957s1	0.2191	0.2037	0.214382	0.2269	0.2037	0.214349	0.2299	0.2022	0.225586	0.2562	10.97	7.92	61.58	n.s.	***	***
T0958	0.4481	0.4416	0.452806	0.487	0.4286	0.452396	0.4773	0.3994	0.452514	0.5065	61.58	6.70	53.04	n.s.	n.s.	n.s.
T0960	0.1457	0.1357	0.145476	0.1524	0.137	0.145227	0.1531	0.1229	0.135504	0.1559	39.63	35.97	57.92	n.s.	***	***
T0963	0.1511	0.1408	0.147979	0.1566	0.1401	0.146997	0.1566	0.1229	0.138704	0.1559	15.85	4.26	3.65	***	***	***
T0970	0.3118	0.2706	0.294296	0.3176	0.2735	0.296382	0.3176	0.2559	0.289021	0.3206	2.43	2.43	2.43	n.s.	***	***
T0981	0.11	0.1046	0.109555	0.1149	0.1058	0.109509	0.1129	0.0935	0.102878	0.1182	35.97	33.53	10.97	n.s.	***	***
T0984	0.319	0.3012	0.311407	0.3194	0.3008	0.311644	0.3222	0.2726	0.295018	0.3226	3.65	3.048	4.26	n.s.	***	***
T0992	0.6355	0.5864	0.624764	0.6589	0.6005	0.627206	0.6542	0.5327	0.591867	0.6519	17.07	16.46	5.48	n.s.	***	***
T1005	0.2991	0.2753	0.29181	0.3021	0.2745	0.29365	0.3037	0.2339	0.268402	0.3137	6.70	13.41	6.70	n.s.	***	***
T1019s1	0.2284	0.2026	0.22145	0.2457	0.1983	0.222785	0.2371	0.2026	0.217466	0.2543	8.53	16.46	12.19	n.s.	***	***
T1022s1	0.2814	0.2601	0.274406	0.2948	0.2601	0.275921	0.2892	0.2309	0.273495	0.3072	10.976	10.36	29.87	*	***	n.s.
The Cumulative scores	3.952	3.6736	3.887622	4.1217	3.6729	3.898057	4.0941	3.3332	3.743523	4.2023						

Table S. 22 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 FM/TBM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 42

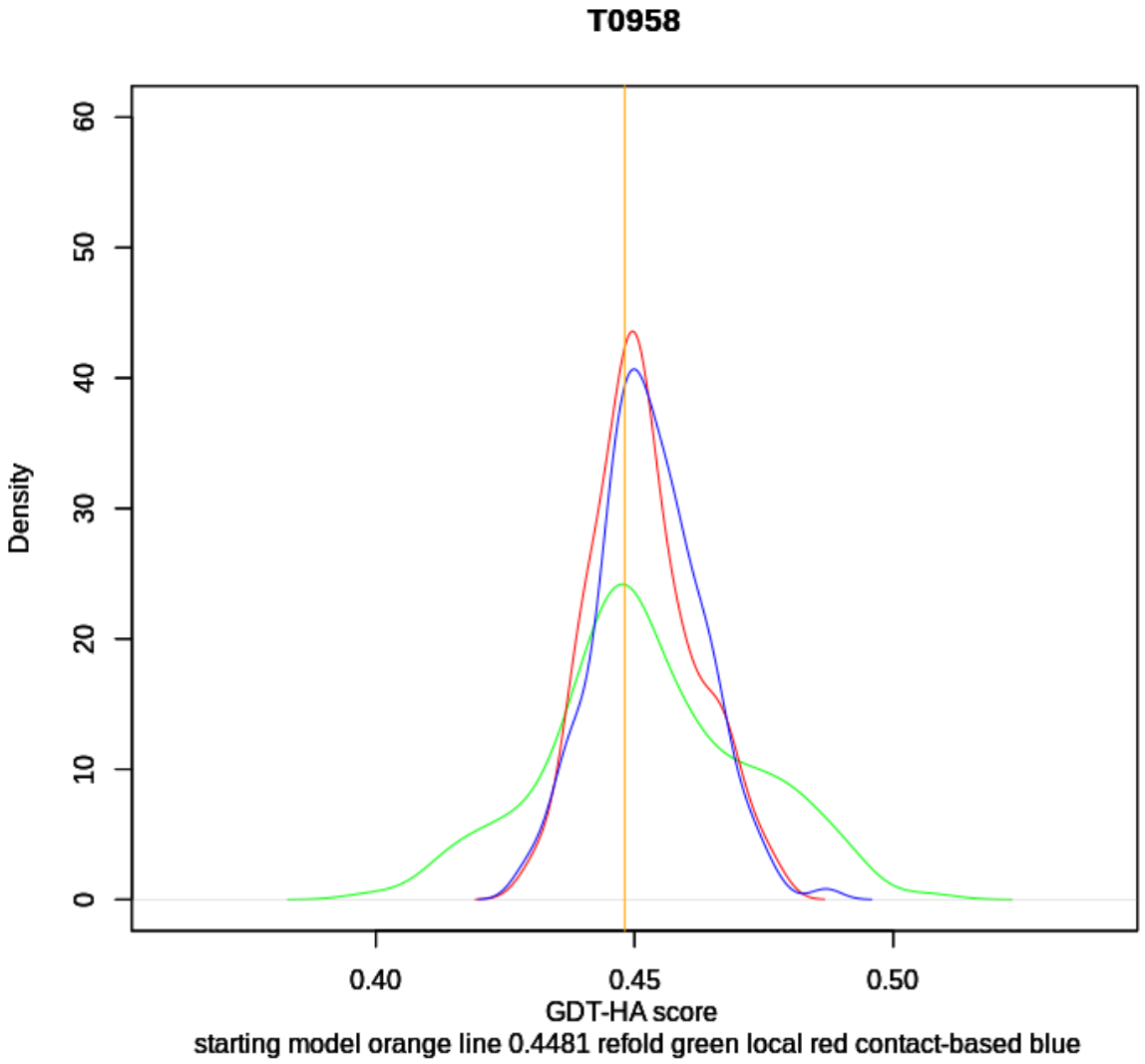

Figure S. 20 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on an FM/TBM target.

Performance of methods on T0958 (an FM/TBM-hard CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.4481 and higher GDT HA scores are better).

Appendix 43

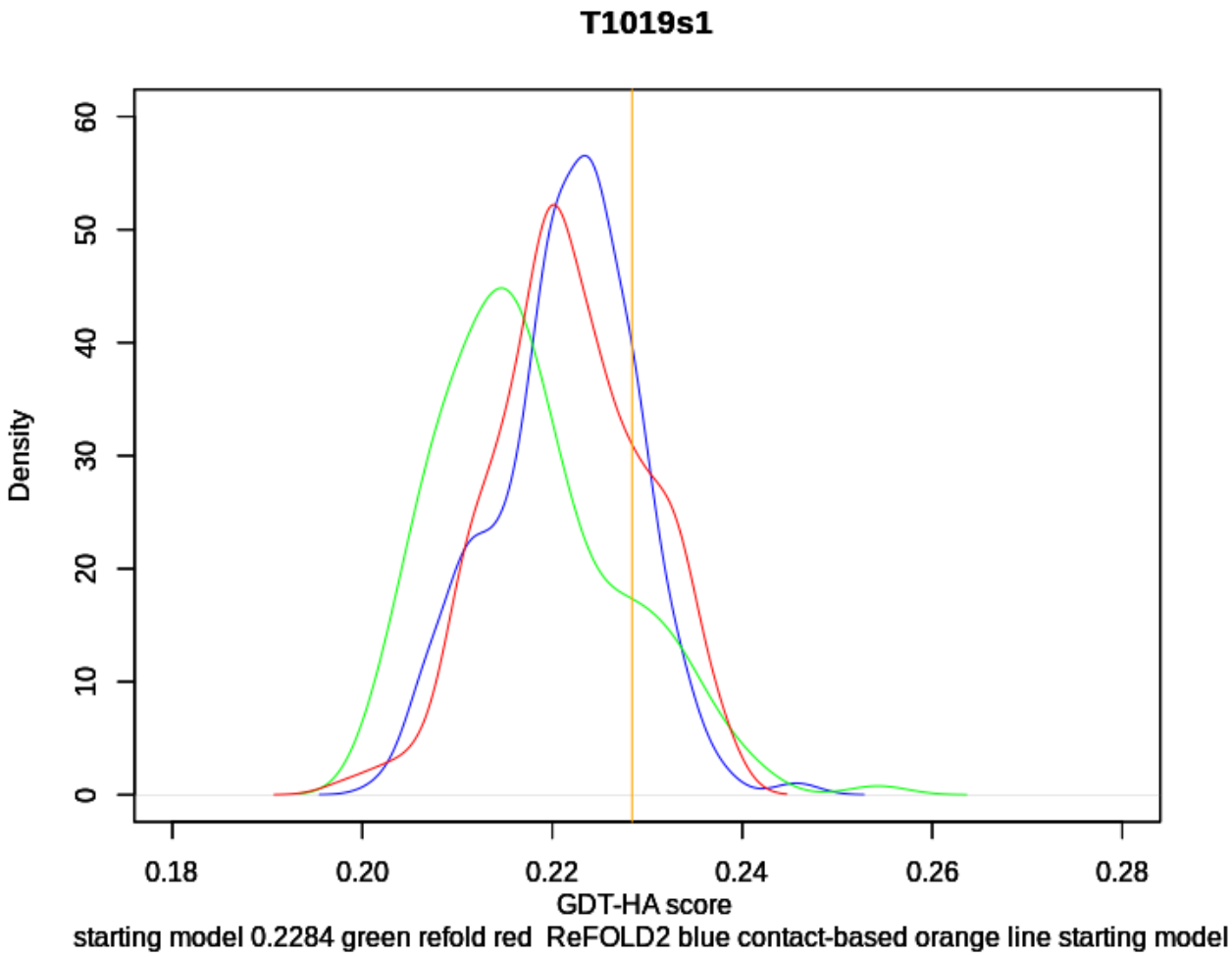

Figure S. 21 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on an FM/TBM target.

Performance of methods on T1019s1 (an FM/TBM-hard CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.2284 and higher GDT HA scores are better).

Appendix 44

	GDT-HA score													Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			The percentage of the improved models			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact	Local	ReFOLD	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0950	0.1842	0.1806	0.188219	0.1966	0.1798	0.188403	0.1966	0.1864	0.196828	0.2171	85.36	86.58	93.90	n.s.	***	***
T0953s1	0.2639	0.2569	0.280822	0.3056	0.2431	0.284723	0.316	0.2292	0.268832	0.3125	96.95	95.73	52.43	***	***	***
T0968s1	0.3877	0.3559	0.3782509	0.3983	0.3602	0.3774952	0.4025	0.3686	0.424481	0.4915	10.97	7.31	89.02	n.s.	***	***
T0968s2	0.4043	0.3913	0.408798	0.4283	0.3913	0.408709	0.4239	0.3717	0.404225	0.4435	72.56	71.34	47.56	n.s.	***	***
T0975	0.3256	0.2936	0.316117	0.3336	0.2998	0.316833	0.3292	0.2758	0.31969	0.3577	6.70	5.48	39.02	***	***	***
T0989	0.1524	0.1433	0.15147	0.1596	0.1413	0.151349	0.1596	0.1209	0.140116	0.1585	32.31	6.70	1.82	n.s.	***	***
T0991	0.1462	0.1314	0.143726	0.1525	0.1314	0.143451	0.1547	0.1229	0.146705	0.1674	29.87	14.63	47.56	n.s.	***	***
T1000	0.3132	0.306	0.318578	0.3282	0.3088	0.31863	0.3276	0.2877	0.311263	0.3431	95.73	94.51	38.41	n.s.	***	***
T1001	0.5126	0.4964	0.523345	0.5486	0.5018	0.522794	0.545	0.4658	0.521675	0.5701	80.48	87.80	74.39	n.s.	n.s.	n.s.
T1010	0.2083	0.1881	0.199473	0.2119	0.1881	0.200436	0.2119	0.1393	0.161854	0.2048	2.43	1.82	0	n.s.	***	***
T1015s1	0.1989	0.1847	0.200755	0.2159	0.1903	0.201427	0.2216	0.1875	0.203309	0.2216	51.82	48.78	49.39	n.s.	***	***
The cumulative Score	3.0973	2.9282	3.1095539	3.2791	2.9359	3.1142502	3.2886	2.7558	3.098978	3.4878						

Table S. 23 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 FM targets according to GDT-HA score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (higher GDT-HA scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and higher GDT-HA scores are better).

Appendix 45

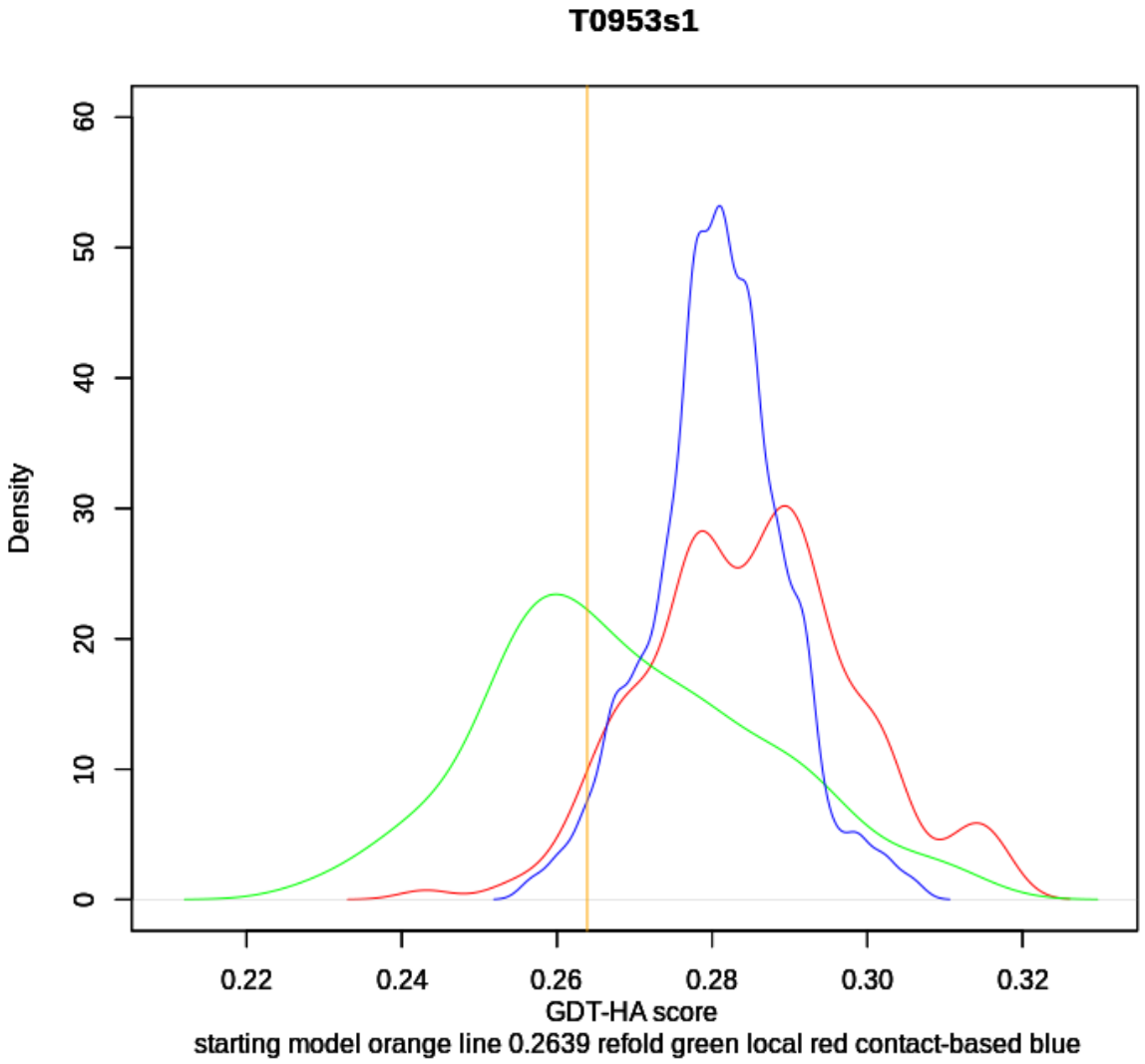

Figure S. 22 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on an FM target.

Performance of methods on T0953s1 (an FM CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.2639 and higher GDT HA scores are better).

Appendix 46

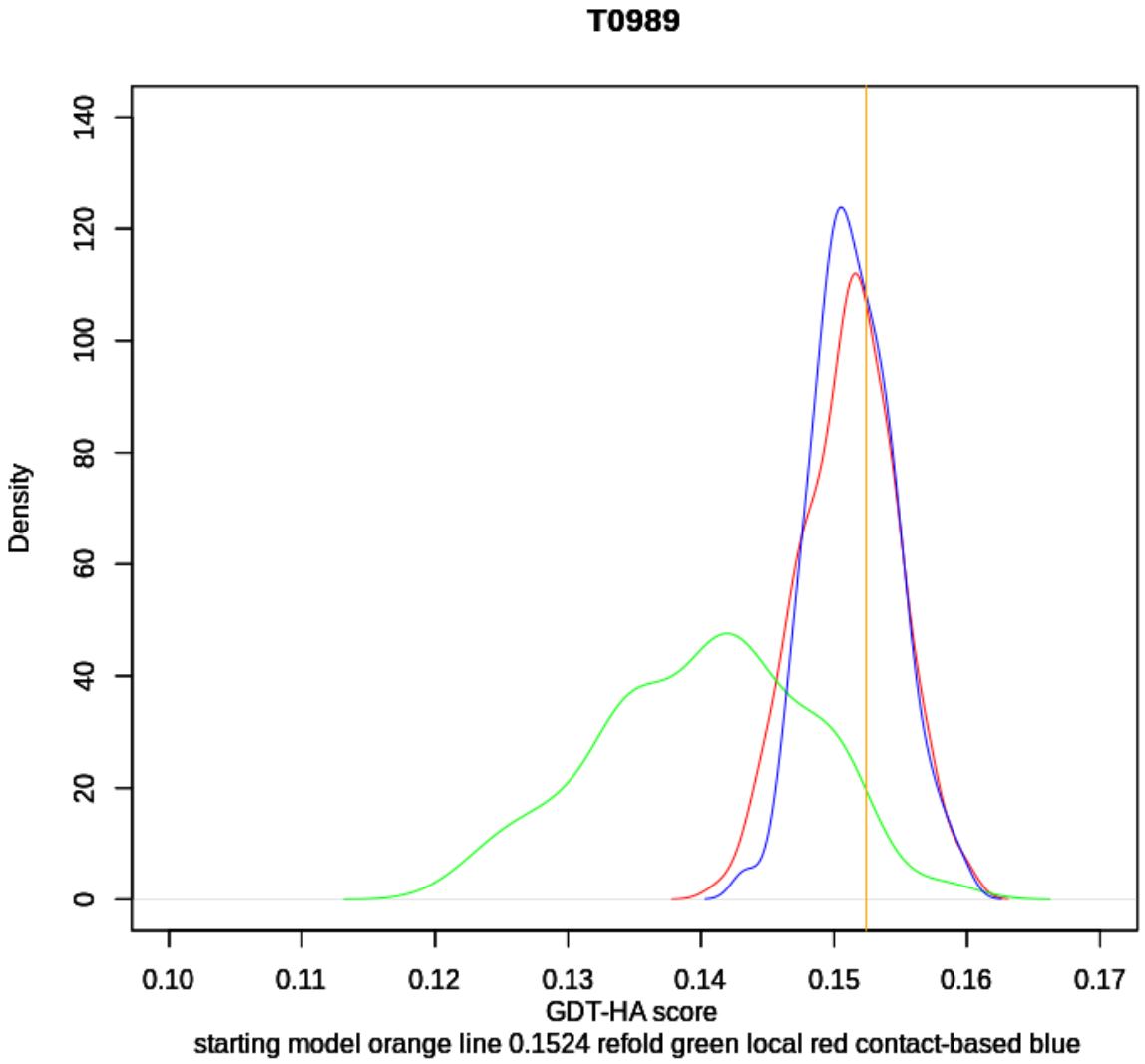

Figure S. 23 A comparison of the contact-assisted MD-based protocol with the original MD-based protocol of ReFOLD and the fixed local quality assessment guided MD-based protocol on an FM target.

Performance of methods on T0989 (an FM CASP13 target) according to GDT-HA score. The blue line represents the contact-assisted MD-based protocol, the red line represents the fixed local quality assessment guided MD-based protocol, the green line represents the MD-based protocol of ReFOLD, and the orange vertical line represents the initial structure (the GDT-HA score of the initial structure was 0.1524 and higher GDT HA scores are better).

Appendix 47

	The contact-assisted guided MD-based protocol versus the fixed local quality assessment guided MD-based protocol				The contact-assisted guided MD-based protocol versus the original MD-based protocol of ReFOLD			The fixed local quality assessment guided MD-based protocol versus the original MD-based protocol of ReFOLD		
CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Maximum vs Starting	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum
FM	0.9037	0.8669	0.8286	**0.0002441**	**0.004639**	0.2119	0.9983	**0.002441**	0.2119	0.9928
TBM	0.05825	0.6711	0.6027	**4.768e-07**	**5.579e-05**	**0.0008001**	0.9985	**5.579e-05**	**0.0009294**	0.9979
FM/TBM	0.4823	0.9877	**0.04584**	**6.104e-05**	**0.0008308**	**0.002014**	0.9873	**0.0008459**	**0.002625**	0.9951
ALL	0.4038	0.9229	0.228	**5.684e-14**	**2.872e-08**	**8.482e-05**	**1**	**2.031e-08**	**0.0001171**	**1**

Table S. 24 Calculated pairwise p-values for comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 targets according to GDT-HA score.

Ho: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. The maximum score of the models generated by the contact-assisted MD-based protocol were also compared with the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface, higher GDT-HA scores are better).

Appendix 48

	Molprobity score										Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0954	1.32	1.05	1.49922	1.68	1.01	1.6417	2.28	1.03	1.42958	1.64	n.s.	n.s.	n.s.
T0959	0.72	0.5	0.999699	1.42	0.5	1.05877	1.48	0.67	1.00807	1.42	n.s.	n.s.	n.s.
T0964	2.59	0.88	1.29355	1.66	0.9	1.44894	2.06	0.91	1.28536	1.65	**	n.s.	n.s.
T0965	2.99	1.54	1.73422	2.06	1.46	1.89144	2.4	1.32	1.59934	2.07	n.s.	n.s.	n.s.
T0966	0.82	0.83	1.2009	1.47	0.83	1.27957	1.71	0.83	1.19819	1.49	n.s.	n.s.	n.s.
T1011	2.79	1.15	1.415	1.62	1.12	1.55383	1.93	0.98	1.33814	1.65	n.s.	n.s.	n.s.
T1015s2	2.75	1.33	1.78205	2.07	1.35	1.97463	2.51	0.8	1.39814	2.06	***	n.s.	n.s.
T1021s2	1.26	1	1.39867	1.67	0.97	1.52043	1.97	0.96	1.33102	1.6	***	n.s.	n.s.
T1022s2	1.27	0.96	1.38	1.61	0.99	1.51787	1.83	0.96	1.35701	1.64	***	n.s.	n.s.
T0973	0.88	0.88	1.26114	1.74	0.92	1.35202	1.78	0.84	1.23102	1.6	n.s.	n.s.	n.s.
T0974s1	0.92	0.5	1.0791	1.73	0.5	1.145	1.66	0.5	1.06251	1.64	n.s.	n.s.	n.s.
T0977	3.46	1.52	1.68422	1.83	1.24	1.74974	2.15	1.19	1.50923	1.74	n.s.	n.s.	n.s.
T0983	0.81	0.57	0.986627	1.38	0.55	1.01404	1.31	0.64	1.21922	1.64	n.s.	***	***
T0993s1	0.73	0.56	1.1582	1.49	0.56	1.2483	1.68	0.56	1.26066	1.6	n.s.	***	***
T0993s2	3.05	1.02	1.53443	1.9	1.07	1.70585	2.16	0.7	1.31904	1.75	**	n.s.	n.s.
T0995	3.18	0.99	1.49569	1.69	1.18	1.67777	2.08	1.02	1.43443	1.71	***	n.s.	n.s.
T1003	0.75	0.69	1.13741	1.43	0.7	1.24261	1.5	0.66	1.1488	1.44	***	n.s.	n.s.
T1004	0.81	0.81	1.10317	1.39	0.85	1.19484	1.7	0.85	1.15048	1.45	n.s.	***	n.s.
T1013	2.76	1.35	1.56193	1.82	1.36	1.71824	1.99	1.02	1.39443	1.81	***	n.s.	n.s.
T1014	2.58	1.39	1.67404	2.5	1.35	1.8309	2.22	0.97	1.38892	1.99	***	1	n.s.
T1016	1.26	0.78	1.07765	1.51	0.81	1.2208	1.55	0.77	1.07874	1.5	***	n.s.	n.s.
T1018	2.6	1.4	1.5997	1.77	1.23	1.74973	1.73	0.92	1.40168	1.7	n.s.	n.s.	n.s.
T1020	3.53	1.82	1.95458	2.21	1.79	2.12963	2.54	1.53	1.72663	2.21	n.s.	n.s.	n.s.
The Cumulative Score	43.83	23.52	32.011196	39.65	23.24	34.86665	44.22	20.63	30.27064	39			

Table S. 25 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 TBM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 49

	Molprobity score										Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0949	3.3	1.59	1.91482	2.27	1.06	1.88595	2.57	1.02	1.54012	2.26	n.s.	n.s.	n.s.
T0953s2	3.36	1.47	1.91096	2.26	1.75	2.13267	2.48	1.22	1.66813	2.26	***	n.s.	n.s.
T0955	2.29	0.86	1.22645	1.73	0.86	1.41527	1.72	0.86	1.18295	1.66	n.s.	n.s.	n.s.
T0957s1	0.64	0.5	0.888563	1.3	0.5	0.978032	1.4	0.5	0.952335	1.38	**	**	n.s.
T0958	0.85	0.53	0.976886	1.44	0.54	1.05612	1.9	0.54	1.10066	1.58	*	**	n.s.
T0960	1.13	0.77	1.29145	1.64	0.83	1.42101	2.05	0.92	1.30693	1.58	***	n.s.	n.s.
T0963	3.16	1.63	1.84042	2.07	1.65	2.02383	2.44	1.27	1.65783	2.1	***	n.s.	n.s.
T0970	1.33	0.88	1.23145	1.73	0.79	1.33553	1.9	0.87	1.18199	1.69	*	n.s.	n.s.
T0981	3.76	1.92	2.05641	2.37	1.95	2.2691	2.84	1.62	1.82928	2.37	***	n.s.	n.s.
T0984	3.51	1.42	1.67331	2.06	1.55	1.88665	2.2	1.09	1.44355	2.08	***	n.s.	n.s.
T0992	1	0.84	1.2644	1.74	0.88	1.29766	1.6	0.81	1.28497	1.73	n.s.	n.s.	n.s.
T1005	3.35	1.68	1.89976	2.2	1.73	2.10399	2.52	1.21	1.61024	2.19	**	n.s.	n.s.
T1019s1	1.4	0.87	1.23012	1.83	0.87	1.3841	2.03	0.87	1.3841	2.03	**	**	*
T1022s1	2.67	1.27	1.5603	1.8	1.12	1.74436	2.33	1.07	1.49934	1.82	*	n.s.	n.s.
The Cumulative Score	31.75	16.23	20.965299	26.44	16.08	22.934272	29.98	13.87	19.642425	26.73			

Table S. 26 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 FM/TBM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 50

	Molprobity score										Wilcoxon Tests		
CASP TARGETS		The contact-assisted MD-based protocol			The fixed local quality assessment guided MD-based protocol			The original the original MD-based protocol of ReFOLD			Significance		
Target ID	Starting model	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Minimum Score	Mean Score	Maximum Score	Contact vs Local	Contact vs ReFOLD	Local vs ReFOLD
T0950	1.06	0.91	1.25545	1.52	0.82	1.32861	1.84	0.88	1.1944	1.52	n.s.	n.s.	n.s.
T0953s1	2.8	0.89	1.41211	1.96	1.01	1.69016	2.57	0.66	1.26801	1.96	***	n.s.	n.s.
T0968s1	1.32	0.65	1.11217	1.54	0.7	1.17973	1.51	0.52	1.11988	1.63	n.s.	n.s.	n.s.
T0968s2	1.04	0.68	1.07952	1.55	0.5	1.25426	2.04	0.53	1.13569	1.64	***	**	n.s.
T0969	3.81	1.9	2.08114	2.38	1.94	2.29176	2.89	1.59	1.79958	2.38	***	n.s.	n.s.
T0975	3.82	1.79	1.99211	2.27	1.84	2.21681	2.89	1.4	1.6882	2.29	**	n.s.	n.s.
T0989	3.31	1.68	1.90313	2.27	1.58	2.10321	2.55	1.31	1.67614	2.29	***	n.s.	n.s.
T0991	2.86	1.18	1.80305	2.17	1.51	1.97899	2.51	0.98	1.50581	2.16	*	n.s.	n.s.
T1000	1.37	0.97	1.4306	1.63	0.97	1.51681	2.1	1	1.34861	1.59	n.s.	n.s.	n.s.
T1001	1.17	0.5	1.00331	1.44	0.5	1.07787	1.43	0.66	1.13401	1.59	n.s.	**	***
T1010	3.49	1.68	1.98006	2.3	1.7	2.17787	2.76	1.19	1.67928	2.29	n.s.	n.s.	n.s.
T1015s1	0.5	0.5	0.845783	1.38	0.57	0.933404	1.55	0.5	0.879341	1.42	*	*	n.s.
T1017s2	2.87	1.07	1.67476	2.08	1.08	1.82686	2.31	0.89	1.35892	2.1	n.s.	n.s.	n.s.
The Cumulative Score	29.42	14.4	19.573193	24.49	14.72	21.576344	28.95	12.11	17.787871	24.86			

Table S. 27 Performance comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 FM targets according to Molprobity score.

One-tailed Wilcoxon tests were also used to compare the MD-based protocols for each target (lower Molprobity scores are better). H_o: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H_1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. P-values ≤ 0.05 indicate significant statistical differences (*, **, *** indicate statistical significance at $p < 0.05$, $p<0.01$ and $p<0.001$, respectively, while n.s. indicates not significant, and lower Molprobity scores are better).

Appendix 51

	The contact-assisted guided MD-based protocol versus the fixed local quality assessment guided MD-based protocol				The contact-assisted guided MD-based protocol versus the original MD-based protocol of ReFOLD			The fixed local quality assessment guided MD-based protocol versus the original MD-based protocol of ReFOLD		
CASP Target Category	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Minimum vs Starting	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum	Minimum vs Minimum	Mean vs Mean	Maximum vs Maximum
FM	0.1448	**6.104e-05**	**0.0003052**	**0.03925**	0.9971	0.9966	0.01806	0.9958	0.9999	0.9997
TBM	0.7494	**5.96e-08**	**0.0002197**	0.06042	0.9958	0.9978	0.8916	0.9984	1	1
FM/TBM	0.4844	**6.104e-05**	**0.0002136**	**0.02063**	0.9962	0.9823	0.1471	0.9853	0.9992	0.999
ALL	0.2542	**4.399e-10**	**1.676e-08**	**0.003338**	1	0.9999	0.2933	1	1	1

Table S. 28 Calculated pairwise p-values for comparison of the contact-assisted, the fixed local quality assessment guided MD-based protocols and the original MD-based protocol of ReFOLD on the CASP13 targets according to Molprobity score.

Ho: The scores of the models generated by the contact-assisted MD-based protocol are equal or lower in quality than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. H1: The scores of the models generated by the contact-assisted MD-based protocol are higher quality models than those generated by the fixed local quality assessment guided and the original MD-based protocol of ReFOLD. The minimum score of the models generated by the contact-assisted MD-based protocol were also compared with the starting models in the Wilcoxon tests. P-values ≤ 0.05 indicate significant statistical differences (in boldface, lower Molprobity scores are better)