A principal odor map unifies diverse tasks in olfactory perception

Article

Accepted Version

A Principal Odor Map Unifies Diverse Tasks in Olfactory Perception

Brian K. Lee[1]†, Emily J. Mayhew[2,3]†, Benjamin Sanchez-Lengeling[1], Jennifer N. Wei[1], Wesley W. Qian[1,4,5], Kelsie A. Little[2], Matthew Andres[2], Britney B. Nguyen[2], Theresa Moloy[2], Jake Yasonik[1,4], Jane K. Parker[6], Richard C. Gerkin[1,4,7], Joel D. Mainland[2,8*], Alexander B. Wiltschko[1,4*]

[1]Google Research, Brain Team; Cambridge, MA, USA.

[2]Monell Chemical Senses Center; Philadelphia, PA, USA.

[3]Department of Food Science and Human Nutrition, Michigan State University; East Lansing, MI, USA.

[4]Osmo; Cambridge, MA, USA.

[5]Department of Computer Science, University of Illinois; Urbana-Champaign, IL, USA.

[6]Department of Food and Nutritional Sciences, University of Reading; Reading, Berkshire, England.

[7]School of Life Sciences, Arizona State University; Tempe, AZ, USA.

[8]Department of Neuroscience, University of Pennsylvania; Philadelphia, PA, USA.

*Corresponding author. Email: jmainland@monell.org, alex@osmo.ai

†These authors contributed equally to this work.

Mapping molecular structure to odor perception is a key challenge in olfaction. We used graph neural networks to generate a Principal Odor Map (POM) that preserves perceptual relationships and enables odor quality prediction for novel odorants. The model was as reliable as a human in describing odor quality: on a prospective validation set of 400 novel odorants, the model-generated odor profile more closely matched the trained panel mean (n=15) than did the median panelist. Applying simple, interpretable, theoretically-rooted transformations, the POM outperformed chemoinformatic models on several other odor prediction tasks, indicating that the POM successfully encoded a generalized map of structure-odor relationships. This approach broadly enables odor prediction and paves the way toward digitizing odors.

One-Sentence Summary: An odor map achieves human-level odor description performance and generalizes to diverse odor-prediction tasks.

A fundamental problem in neuroscience is mapping the physical properties of a stimulus to perceptual characteristics. In vision, wavelength maps to color; in audition, frequency maps to pitch. By contrast, the mapping from chemical structures to olfactory percepts is poorly understood. Detailed and modality-specific maps like the CIE color space (1), and Fourier space (2) led to a better understanding of visual and auditory coding. Similarly, to better understand olfactory coding, olfaction needs a better map.

Pitch increases monotonically with frequency; in contrast, the relationship between odor percept and odorant structure is riddled with discontinuities, exemplified by Sell's triplets (3), trios of molecules in which the structurally similar pair is not the perceptually similar pair (Fig. 1A). These discontinuities in the structure-odor relationship suggest that standard chemoinformatic representations of molecules—functional group counts, physical properties, molecular fingerprints, etc.— used in recent odor modeling work (4–6) are inadequate to map odor space.

The principal odor map represents perceptual distances and hierarchies

To generate odor-relevant representations of molecules, we constructed a Message Passing Neural Network (MPNN) (7), a specific type of graph neural network (GNN) (8), to map chemical structures to odor percepts. Each molecule was represented as a graph, with each atom described by its valence, degree, hydrogen count, hybridization, formal charge, and atomic number. Each bond was described by its degree, aromaticity, and whether it is in a ring. Unlike traditional fingerprinting techniques (9), which assign equal weight to all molecular fragments within a set bond radius, a GNN can optimize fragment weights for odor-specific applications. Neural networks have unlocked predictive modeling breakthroughs in diverse perceptual domains (e.g., natural images (10), faces (11), and sounds (12)) and naturally produce intermediate representations of their input data that are functionally high-dimensional, data-driven maps. We used the final layer of the GNN (henceforth, "our model") to directly predict odor qualities, and the penultimate layer of the model as a principal odor map (POM). The POM 1) faithfully represented known perceptual hierarchies and distances, 2) extended to novel odorants, 3) was robust to discontinuities in structure-odor distances, and 4) generalized to other olfactory tasks.

We curated a reference dataset of approximately 5,000 molecules, each described by multiple odor labels (e.g. creamy, grassy), by combining the GoodScents (13) and Leffingwell (14) (GS/LF) flavor and fragrance databases (Fig. 1B). To train the model, we optimized model parameters with a weighted-cross entropy loss over 150 epochs using Adam (15) with a learning

rate decaying from 5e-4 to 1e-5, with a batch size of 128. The GS/LF dataset was split 80/20 train/test, and the 80% train set further subdivided into 5 cross-validation splits. These CV splits were used to optimize hyperparameters using Vizier (16), a Bayesian optimization algorithm, tuning over 1,000 trials. Details about model architecture and hyperparameters are given in the supplementary methods. When properly hyperparameter-tuned, performance was found to be robust across many model architectures. We present results for the model with the highest mean AUROC on the cross-validation set (AUROC=0.89) (17).

How does the POM compare to perceptual odor maps and conventional structure-based maps of odorants? Empirical perceptual space (Fig. 1D) intuitively represents perceptual distances, (e.g. two molecules that smell of jasmine should be nearer to each other than to a beefy molecule) and hierarchies (e.g. jasmine and lavender are subtypes of the floral odor family). We show that while this structure is lost in Morgan fingerprint-based maps of odorant space (Fig. 1E), the POM preserves relative perceptual distances and hierarchies (Fig. 1F, Fig. S1-3).

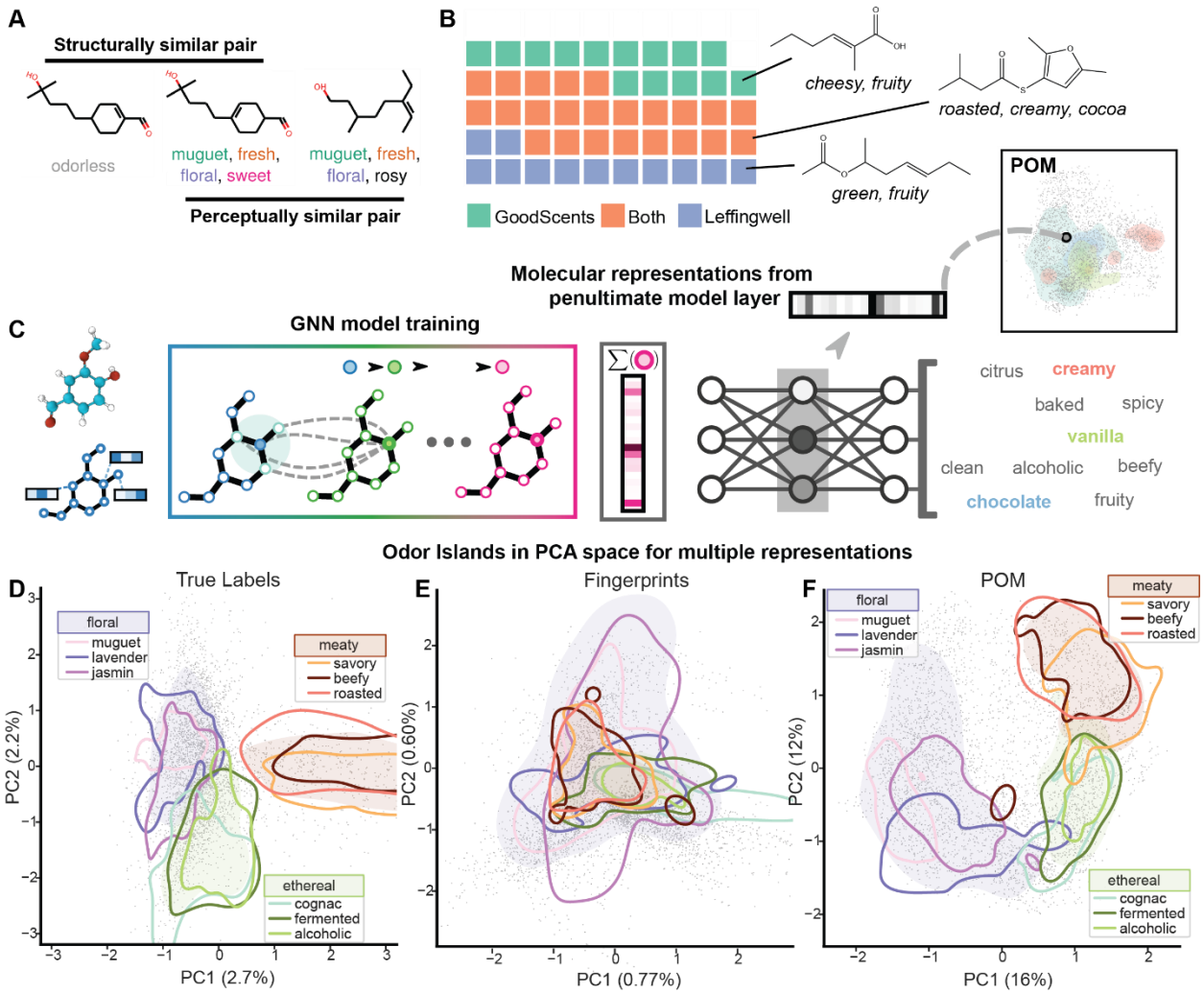

Fig. 1. POM preserves the structure of odor perceptual space. (A) Example triplet of molecules in which the structurally similar pair is not the perceptually similar pair. **(B)** The GNN was trained on a curated dataset of ~5000 semantically labeled molecules drawn from GoodScents (13) and Leffingwell (14) flavor and fragrance databases; one square represents 100 molecules; three example training set molecules and their odor descriptions are shown: 2-methyl-2-hexenoic acid (top), 2,5-dimethyl-3-

thioisovalerylfuran (middle), 1-methyl-3-hexenyl acetate (bottom). **(C)** Schematic illustrating the process of training a GNN to generate the POM. **(D-F)** Odorants plotted by the first and second principal components (PC) of their **(D)** perceptual labels from GS/LF training dataset (138 labels), **(E)** cFP structural fingerprints (radius 4, 2048-bit), and **(F)** POM coordinates (256 dimensions). Areas dense with molecules having the broad category labels floral, meaty, or alcoholic are shaded; areas dense with narrow category labels are outlined. The POM recapitulates the true perceptual map, but the FP map does not; note that only relative (not absolute) coordinates matter. Additional labels are visualized for POM in Fig. S1.

Model outperformed the median panelist on prospective validation task

To test if the model extends to novel odorants, we designed a prospective validation challenge (18) in which we benchmarked model predictive performance against individual human raters. In olfaction, no reliable instrumental method of measuring odor perception exists, and trained human sensory panels are the gold standard for odor characterization (19). Odor perception is variable across individuals (20, 21), but group-averaged odor ratings are stable across repeated measurements (22) and represent our best avenue to establish the ground-truth odor character for novel odorants. We trained a cohort of subjects to describe their perception of odorants using the Rate-All-That-Apply method (RATA) and a 55-word odor lexicon. During training sessions, each term in the lexicon was paired with visual and odor references (Table S1; Fig. S4). Only subjects that met performance standards on the pretest of 20 common odorants (Data S2; individual test-retest correlation $R > 0.35$; reasonable label selection for common odorants) were invited to join the panel.

To avoid trivial test cases, we applied the following selection criteria for the set of 400 novel odorants: 1) molecules must be structurally distinct from each other (Fig. S5), 2) molecules should cover the widest gamut of odor labels (Data S1), and 3) molecules must be structurally or perceptually distinct from any training example (e.g. Fig. 1A, Data S1). Our prospective validation set consists of 55-odor label RATA data for 400 novel, intensity-balanced odorants generated by our cohort of ≥15 panelists (2 replicates). Summary statistics and correlation structure of the human perceptual data is presented in Fig. S6-8. Our panel's mean ratings were highly stable (panel test-retest: $R = 0.80$, $n = 15$; Fig. S9) and more consistent than the DREAM cohort's ratings (6) (Fig. S10-11).

Of the 400 molecules characterized, 77 were dropped from the final prospective validation set due to low intensity (42) (Fig. S12), redundancy (2), mistaken inclusion (1), or with confirmed (19) or potential contamination (13) (Data S1). Model performance was evaluated on the remaining 323 molecules without model retraining.

To measure the model's performance, we compared its normalized predictions with the normalized panel mean rating (Fig. 2A and 2C). One example of raw ratings and predictions for a single molecule, representative of relative GNN and RF performance and panel ratings trends, is given in Fig. 2; additional examples are provided (Fig S14). While there is considerable variation across molecules in the ability of both individual raters and the model to match the panel mean ratings, the model output comes closer to the panel mean than does the median panelist for 53% of molecules (Fig. 2E and 2F). Notably, panelists were able to smell each odorant as they rated it, while the model's predictions were based solely on nominal molecular structure.

As a baseline comparison, we trained a cFP-based random forest (RF) model, the previous state-of-the-art (6), on the same dataset (Fig. 2B). This baseline model surpassed the median panelist for only 41% of molecules.

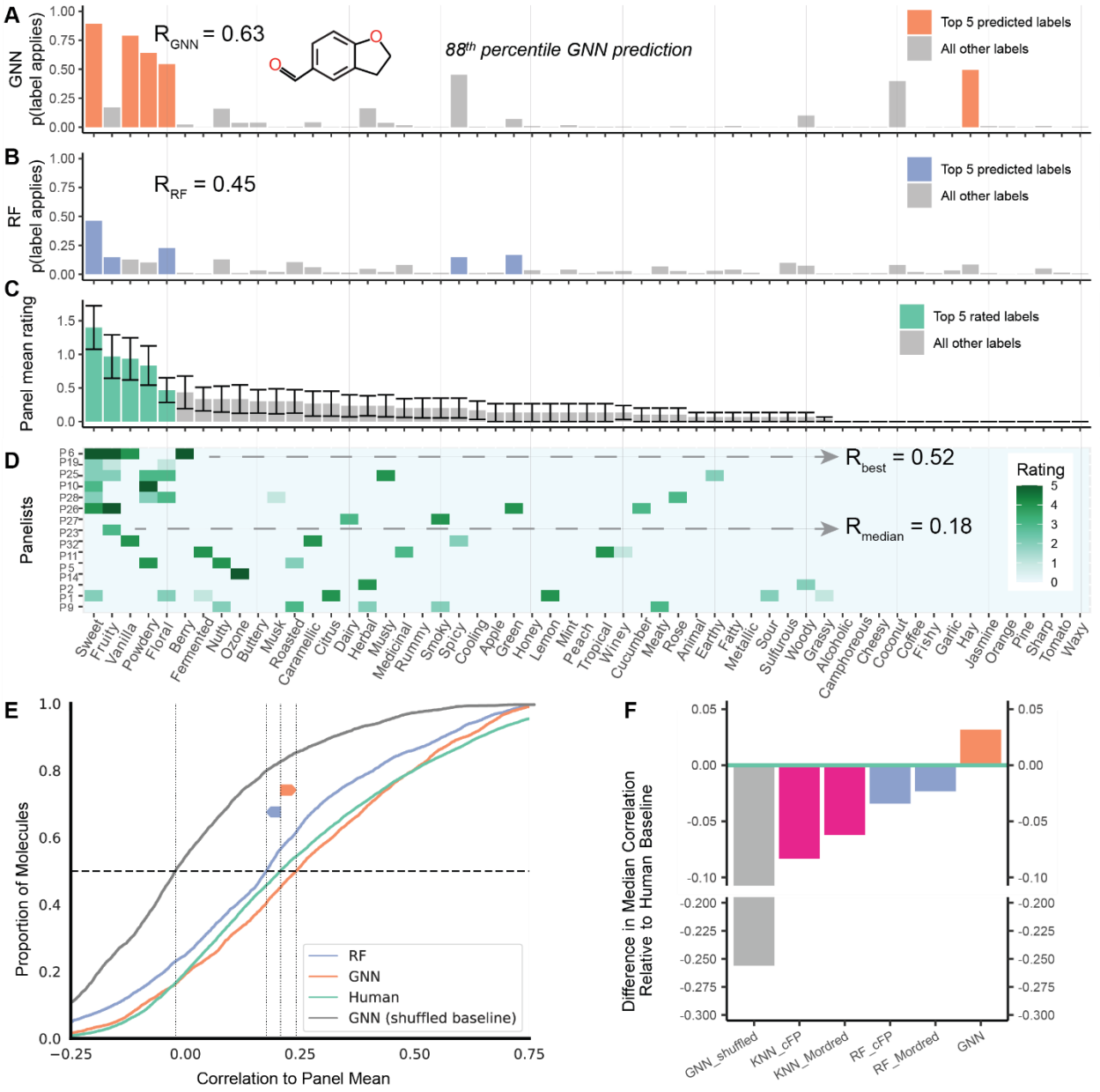

Fig. 2: GNN model displays human-level odor description performance. (A) GNN model label predictions, **(B)** random forest (RF) model label predictions, **(C)** panel mean ratings with standard error bars, and **(D)** individual panelist ratings, averaged over 2 replicates, for the molecule 2,3-dihydrobenzofuran-5-carboxaldehyde. In panels **A-C**, the top 5 ranked descriptors are in orange (GNN), purple (RF), or green (panel). Descriptors in panels **A-D** are ordered by panel mean ratings. Panels **A**, **B**, and **D** are annotated with the Pearson correlation coefficient of their data to the panel mean rating shown in panel **C**. Panel **D** includes panelist/panel correlation coefficients for the panelist that best matches the panel mean and for the panelist with the median match. **(E)** Cumulative density plot showing the distribution of correlations between human panelists and the panel mean (in green) and between the GNN, RF, and GNN shuffled model predictions and the panel mean on a per molecule basis. Curves shifted to

the right are more strongly correlated to the panel mean. **(F)** Difference in the median correlation to the panel mean relative to the median human subject's correlation to the panel mean for models trained using k-nearest neighbor (KNN) and RF, trained on cFPs or Mordred features, and the GNN model. Only the GNN model has a median correlation to the panel mean that is higher than that of the median panelist.

The GNN model shows human-level performance in aggregate, but how does it perform across perceptual and chemical classes? When we disaggregated performance by odor label, the model was within the distribution of human raters for all labels except musk and surpasses the median panelist for 30/55 labels (55%, Fig. 3A). This per-label view indicates that the GNN model is superior to the previous state of the art model trained on the same data (paired 2-tailed t-test p=3.3e-7).

Predictive performance for a given label depends on the complexity of the structure-odor mapping for that label. It is thus unsurprising that it performs best for labels like garlic and fishy that have clear structural determinants (sulfur-containing for garlic; amines for fishy), and worst for the label musk, which includes at least 5 distinct structural classes (macrocyclic, polycyclic, nitro, steroid-type, and straight-chain) (23, 24). In contrast, a panelist's performance for a given label depends on their familiarity with the label in the context of smell; consequently, we see strong panelist-panel agreement for labels describing common food smells like nutty, garlic, and cheesy and weak agreement for labels like musk and hay.

Model performance also depends on the number of training examples for a given label; with enough examples, models can learn even complex structure-percept relationships. In general, our model's performance was high for labels with many training examples (e.g, fruity, sweet, floral) (Fig. 3B), but performance for labels with few training examples was either high (e.g., fishy, camphoreous, cooling) or low (e.g, ozone, sharp, fermented). Likewise, model performance was bounded above by panel test-retest correlation (Fig. S15). When we disaggregated by chemical classes (e.g. esters, phenols, amines), both panelist and model performance was relatively uniform (Fig. 3C), with sulfur-containing molecules showing strongest performance from panelists and the model (R = 0.52).

Chemical materials are impure - a fact too often unaccounted for in olfactory research (26). To measure the contribution of impurities to the odor percept of our stimuli, we applied a gas chromatography-mass spectrometry (GC-MS) and gas chromatography-olfactometry (GC-O) quality control (QC) procedure to 50 stimuli (Data S1). This QC procedure matches an odor percept to its causal molecule, allowing us to identify stimuli for which the primary odor character was not due to the nominal compound. We selected the 50 molecules to represent 3 quadrants of intrapanel agreement and model-panel agreement (high/high, high/low, low/high), anticipating that oodorous contaminants may explain cases of poor model-panel agreement. Our QC led to diverse conclusions: the nominal compound caused the odor (11/50), contaminants contribute to the odor in a minor way (15/50) or major way (5/50), contaminants caused the odor (15/50), or the cause of the odor could not be determined (4/50) (Fig. 3D). Fishy, garlic, and sulfurous were the most prevalent contaminant odor qualities (Fig. S17); these labels were overrepresented in the QC set, so we expect that the rate of severe contamination is likely lower in the full test set than the QC set. In some cases, while we purchased a novel odorant, the dominant odorant was not novel; for example, the stimulus 4,5-dimethyl-1,3-thiazol-2-amine was described by the panel as buttery, sweet, and dairy, but this odor percept was attributed

through QC to the contaminant diacetyl, a well-known buttery odorant. In another case, the purchased odorant, isobornyl methylacrylate, was described by the panel and the model as both piney and floral; however, the nominal compound was floral only and the piney aroma was due to the closely related compound, borneol, which was detected as a contaminant in the sample. Based on QC results, we removed 32 molecules known or suspected to have high degrees of odorous contamination (Data S1).

Implications of each QC result on model performance are unique (Data S1). In some cases, the model performed well despite the presence of odorous contaminants. We estimate that, if these contaminants were removed from the rated samples, model performance improves in 6 of 50 scenarios, degrades in another 6 of 50 scenarios, remains neutral in 21 of 50 scenarios, and cannot be determined in 17 of 50 scenarios. We estimate the overall rate of significant odorous contamination in our stimulus set at 31.5% (95% CI: 27.4- 35.6%) (Fig. S18).

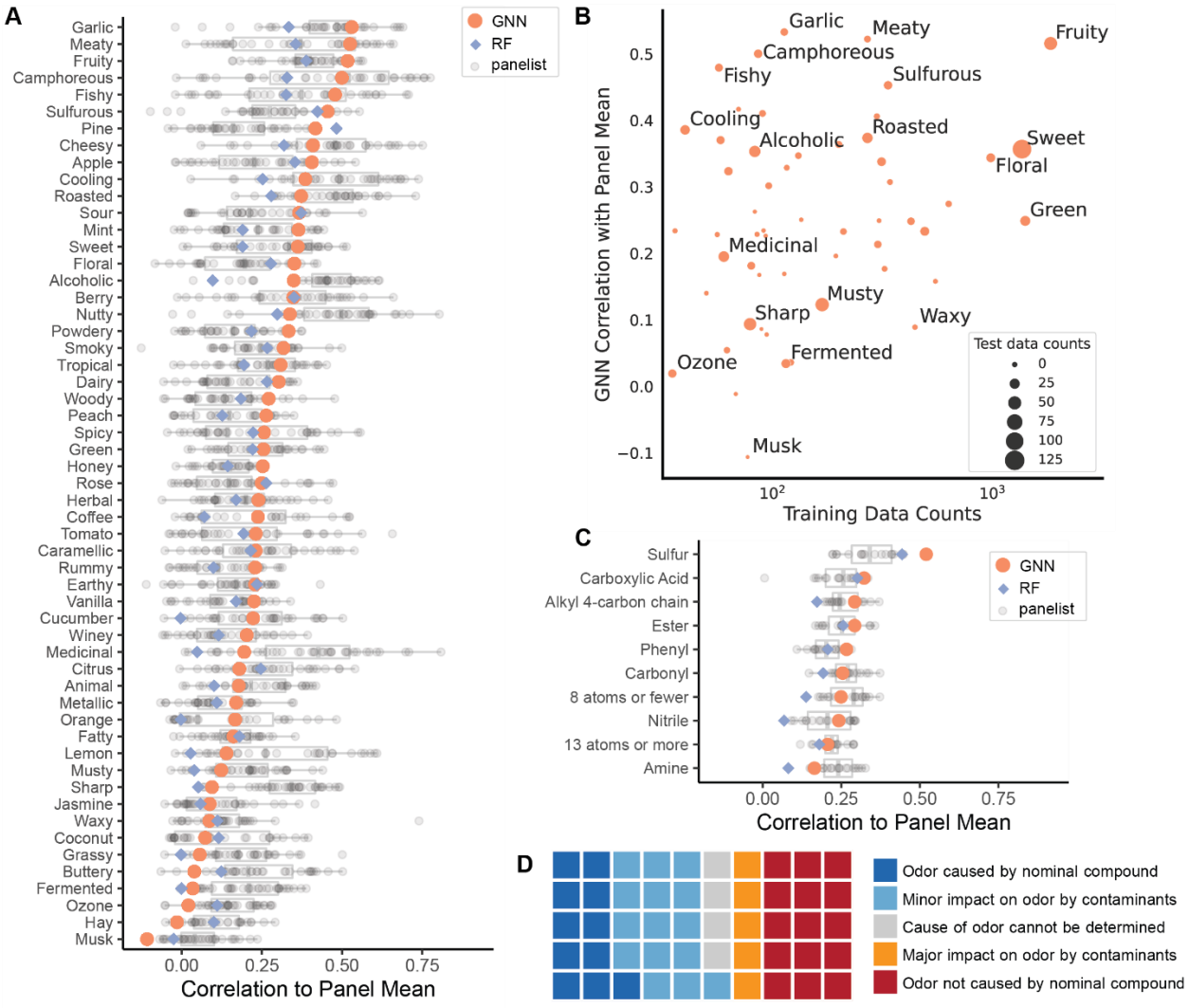

Fig 3. Model performance is robust across structural and perceptual classes. (A) Correlation of GNN (in orange) and RF (in purple) model predictions and panelist ratings (in gray) to the panel mean for each of the 55 odor labels. **(B)** GNN model correlation to panel mean for each of the 55 odor labels plotted against the number of molecules in the training data for which the label applies. Circle size is proportional to the number of test set molecules for which the label applies. Selected data points are annotated. **(C)**

Mean correlation of GNN (in orange) and RF (in purple) model predictions and panelist ratings (in gray) to the panel mean for molecules belonging to 10 common chemical classes. **(D)** Categorization of gas chromatography-olfactometry quality control results for 50 test set stimuli.

POM generalizes to diverse olfactory tasks

To test if the POM was robust to discontinuities in structure-odor distances, we designed an additional challenge in which 41 new triplets (Fig. 4A-B) were constructed and validated by the panel (as in Fig. 1A, Fig. 4C). In each triplet, the anchor molecule was a known odorant, and was matched with one structurally similar and one structurally dissimilar novel odorant, and in which the more *structurally dissimilar* odorant was the more *perceptually similar* of the two to the anchor. To visualize model logic, we made small changes to each node in a molecule, and observed which perturbations had large effects on model predictions; perturbations in nodes with a darker red highlight had a larger impact (Fig. 4A, Fig. S19). Explicit similarity ratings agreed with odor profile distances in 90% of the triplets (Fig. 4D). The model correctly predicted this counterintuitive structure-odor relationship in 50% of cases (Fig. 4E), while the random forest model failed in 81% of cases ($p<0.01$, binomial test of proportions, Fig. 4F).

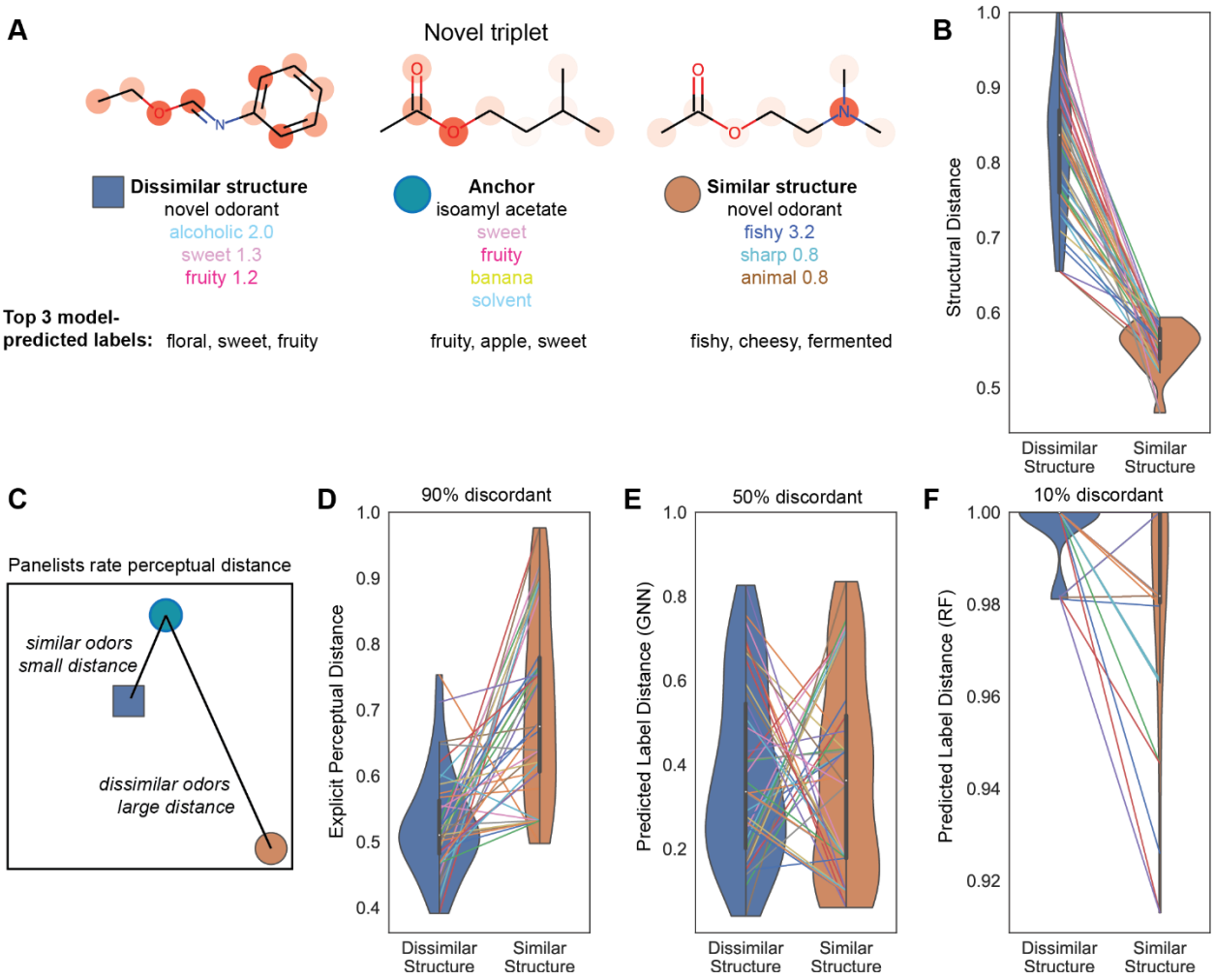

Fig. 4. POM is robust to discontinuities in structure-odor mapping. (A) Example triplet of molecules in which the structurally similar pair is not the perceptually similar pair (i.e. "discordant"), according to the empirical odor labels of each molecule. Training set descriptors (anchor) and mean panel ratings (novel odorants) are shown beneath the molecular structure in colored text; model-predicted labels are listed in black text. Structural nodes highlighted in darker red are more important to model predictions. **(B)** We selected 41 such triplets from the empirical label data, without consulting the model; by design, 100% of these are discordant, and thus represent a difficult test for a predictive perceptual model based on molecular structure. Each colored line connects molecules in a triplet that share the same anchor, as in **(C)**. **(C)** Diagram of the psychophysical task in which panelists rated explicit perceptual distances between molecules in triplets. **(D)** Experimentally-measured explicit perceptual distance ratings in the same triplets also show high discordance with structural distance, i.e. the molecule more structurally similar to the anchor is usually (90%) less perceptually similar. **(E)** The GNN model-predicted labels agree with the counter-intuitive-but-correct perceptual relationship 50% of the time, i.e. they correctly predict the empirical discordance half of the time, as measured by the cosine distance of the predicted, binarized labels. **(F)** A baseline model correctly predicts the empirical discordance only 19% of the time. The models in **(E)** and **(F)** are the same as those from Figures 2 and 3.

A reliable structure-odor map allows us to explore odor space at scale. We compiled a list of ~500,000 potential odorants whose empirical properties are currently unknown to science or industry; most have never been synthesized before. Because a molecule's coordinates in the POM are directly computable from the model, we can plot these potential odorants in the POM (Fig. 5A), revealing a potential space of odorous molecules that is much larger than the much smaller space covered by current fragrance catalogs (~5,000 purchasable, characterized odorants; Fig. 5A inset). These molecules would take approximately 70 person-years of continuous smelling time to collect using our trained human panel.

We show that the POM has a meaningful interpretation by extracting intuitive, geometric measures and mapping them to several olfactory prediction tasks (Fig. 5B). The applicability of any set of odor descriptors corresponds to a projection of the POM coordinates onto axes corresponding to those descriptors; odor strength (detectability) corresponds to the magnitude of this projection (Fig. S13); odor similarity corresponds to the distance between such projections for different molecules. A simple linear model applied to POM and using these geometric interpretations had comparable or superior performance to a chemoinformatic support vector machine (SVM) model across multiple published datasets (Fig 5C, D, E), collectively representing some of the most thorough previous public efforts to characterize these features of odor.

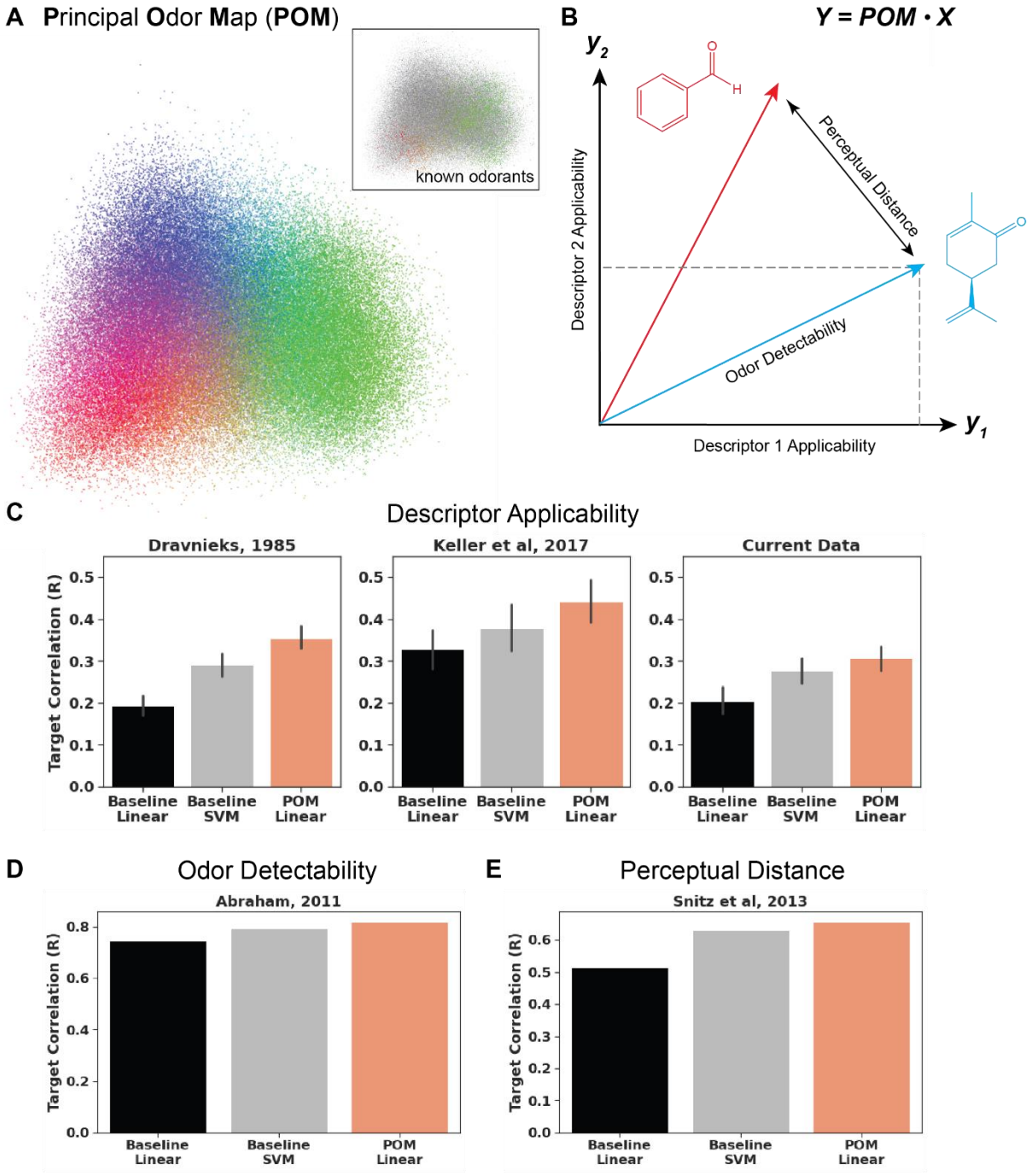

Fig. 5. POM solves a fundamental set of olfactory prediction tasks. (A) 2D trimap embedding of 500,000 unique likely odorants previously uncharacterized. The position of each point (molecule) is determined by POM coordinates, and the RGB values of each point correspond to their coordinates in the first 3 dimensions of a non-negative matrix factorization of the predicted odor labels. The inset plot shows the known odorants from the GS/LF training set (~5,000) in color superimposed over the likely odorants in gray. **(B)** Intuitive geometric measures like vector length, vector distance, and vector projection correspond to the odor prediction tasks of odor detectability, similarity, and descriptor applicability. Equation shows that the projected space *Y* represents the dot product between POM and a task-specific projection matrix *X*. **(C)** A linear model atop POM outperforms a chemoinformatic SVM baseline at predicting odor applicability on two extant datasets, Dravnieks (27) and DREAM (6), as well as the

current data. **(D)** A linear model atop POM outperforms a chemoinformatic SVM baseline at predicting odor detection threshold using data from Abraham et al, 2011 (28). **(E)** A linear model atop POM outperforms a chemoinformatic SVM baseline at predicting perceptual similarity on Snitz et al, 2013 (4).

Discussion

There is no universally accepted method for quantifying and categorizing an odor percept Systems of odor classification have been proposed: first intuitive categorizations (29), then empirically-supported universal spaces (30, 31), and later attempts to incorporate receptor mechanisms (32, 33). However, these systems do not tie stimulus properties to perception, and none have reached broad acceptance. Here we propose and validate a data-driven, high-dimensional map of human olfaction. This map recapitulates the structure and relationships of odor perceptual categories evoked by single molecules. It achieves prospective predictive accuracy in odor description that exceeds that of the typical individual human, and it is broadly transferrable to arbitrary olfactory perceptual tasks using natural and interpretable transformations.

Nearly all published chemosensory models used are fit to the data in their construction. Even using cross-validation, the opportunity for over-fitting is high, because the data come from a single distribution, task, or experimental source. Prospective validation on new data from a new source with no adjustments represents a much more stringent test of real-world utility. In this prospective context, we found that our model performs roughly on par with the median human panelist, beating a chemoinformatic baseline. However, in a real-world setting, models can and should be updated as new data becomes available (34). A linear model atop POM reaches an even higher level of performance when the POM is tuned to the new dataset (Fig. 5C).

The success of this model is not merely an advance in predictive modeling. It offers a simple, contiguous, hierarchical, parseable map of molecular space in terms of odor, much as color spaces represents wavelengths of light in terms of colors and color components. It enables human-level performance not only for odor description but also generalizes to a gamut of other olfactory tasks. It offers the opportunity to reason, intuitively and computationally, about the relationships within and between molecular and odor spaces. Unlike well-known color spaces, it does not provide clear guidance about how stimuli can be mixed to produce new percepts, nor does it use a biologically-plausible architecture. Its closest analogy in vision is the Munsell color system, being a principled way to describe a stimulus in terms of coordinates, but lacking any specific guarantees about mixture behavior. Nonetheless, the Munsell system (and we hope the POM) is still considered to be a useful representation of sensory perception. Further work can aim for a CIE-like representation of odor, one that specifically predicts what odors can be made from mixing what components.

There are some practical considerations to keep in mind when using this map. First, the concentration of an odor influences odor character, but is not explicitly included in the map. While it can predict detection thresholds, a property of the odorant molecule, it cannot predict suprathreshold intensity, a function of the odorant and its concentration. Many molecules have no odor, which we addressed by pre-screening with a separate, simpler model (35), and we diluted odorants to standardize intensity. Second, predictive performance is strong for organic molecules, the vast majority of odorants we encounter, but we could not extend the predictions into halides or molecules that include novel elements due to the lack of safety data for those

molecules. Given uniformly strong performance across broad chemical classes tested in our prospective validation set (Fig. 3C), we expect high accuracy on novel chemicals within these chemical classes, but we would not expect high performance for molecules that have chemical motifs not represented in our training set. For instance, if our training dataset did not contain any molecules with carbon macrocycles, we would not expect the model to accurately predict the odor of an unseen macrocyclic musk (Fig. 3A). Third, many chemical stimuli have odorous contaminants (*26*), particularly those that have not been developed for use in fragrance applications. Neural networks perform well, even with substantial noise in the training and test sets, which we see in the present work. Nonetheless, we recommend isolating the compound of interest from odorous contaminants, and/or characterizing the perceptual quality of contaminants. Fourth, the model was designed to predict the population average, much as color maps predict average perception. It does not yet account for individual differences in perception. Finally, datasets in real-world settings are not static, but grow and shift in distribution — models should be periodically retrained to incorporate new data. Model performance tends to improve with increased training data (Fig. 3B) and data quality (Fig. S15), consistent with ML applications in other areas (36, 37).

Progress in neuroscience is often measured by the creation and discovery of new maps of the world supported by neural circuitry—maps of space in hippocampus, tonotopy in auditory cortex, and retinotopy and Gabor filters in V1 visual cortex, among others. Each is only possible because scientists first possessed a map of the external world, and then measured how responses in the brain varied with stimulus position on the map. This study proposes and validates a novel data-driven map of human olfaction. We hope this map will be useful to researchers in chemistry, olfactory neuroscience, and psychophysics: first, as a drop-in replacement for chemoinformatic descriptors, and more broadly as a new tool for investigating the nature of olfactory sensation.

References and Notes

1. T. Smith, J. Guild, The C.I.E. colorimetric standards and their use. *Trans. Opt. Soc.* **33**, 73–134 (1931).

2. E. F. Evans, Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus. *Psychophys. Physiol. Hear.*, 185–192 (1977).

3. C. S. Sell, On the Unpredictability of Odor. *Angew. Chem. Int. Ed.* **45**, 6254–6261 (2006).

4. K. Snitz *et al.*, Predicting Odor Perceptual Similarity from Odor Structure. *PLOS Comput. Biol.* **9**, e1003184 (2013).

5. A. Ravia *et al.*, A measure of smell enables the creation of olfactory metamers. *Nature* (2020), doi:10/ghmtvk.

6. A. Keller *et al.*, Predicting human olfactory perception from chemical features of odor molecules. *Science*. **355** (2017), doi:10.1126/science.aal2014.

7. J. Gilmer *et al.*, "Message Passing Neural Networks" in *Machine Learning Meets Quantum Physics*, K. T. Schütt et al., Eds. (Springer International Publishing, Cham, 2020; http://link.springer.com/10.1007/978-3-030-40245-7_10), vol. 968 of *Lecture Notes in Physics*, pp. 199–214.

8. B. Sanchez-Lengeling, E. Reif, A. Pearce, A. Wiltschko, A Gentle Introduction to Graph Neural Networks. *Distill*. **6** (2021), doi:10.23915/distill.00033.

9. H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).

10. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM*. **60**, 84–90 (2017).

11. F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering" in (2015; https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html), pp. 815–823.

12. N. Jaitly, P. Nguyen, A. Senior, V. Vanhoucke, "Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition" in *Proceedings of Interspeech 2012* (2012).

13. W. Luebke, The Good Scents Company Information System, (available at http://www.thegoodscentscompany.com/).

14. J. C. Leffingwell, *PMP 2001, database of perfumery materials and performance* (http://www.leffingwell.com/bacispmp.htm).

15. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization (2014), doi:10.48550/ARXIV.1412.6980.

16. D. Golovin *et al.*, "Google Vizier: A Service for Black-Box Optimization" in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Halifax NS Canada, 2017; https://dl.acm.org/doi/10.1145/3097983.3098043), pp. 1487–1495.

17. B. Sanchez-Lengeling *et al.*, Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *ArXiv191010685 Phys. Stat* (2019) (available at http://arxiv.org/abs/1910.10685).

18. S. Kearnes, Pursuing a Prospective Perspective. *Trends Chem.* **3**, 77–79 (2021).

19. H. T. Lawless, H. Heymann, *Sensory Evaluation of Food: Principles and Practices* (Springer, New York, NY, 2010; http://link.springer.com/10.1007/978-1-4419-6488-5), *Food Science Text Series*.

20. C. Trimmer *et al.*, Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proc. Natl. Acad. Sci.* **116**, 9475–9480 (2019).

21. A. Keller *et al.*, An olfactory demography of a diverse metropolitan population. *BMC Neurosci.* **13**, 122 (2012).

22. A. Dravnieks, Odor quality: semantically generated multidimensional profiles are stable. *Sci. N. Y. NY.* **218**, 799 – 801 (1982).

23. K. J. Rossiter, Structure–Odor Relationships. *Chem. Rev.* **96**, 3201–3240 (1996).

24. O. R. P. David, A Chemical History of Polycyclic Musks. *Chem. – Eur. J.* **26**, 7537–7555 (2020).

25. B. Li *et al.*, From musk to body odor: Decoding olfaction through genetic variation. *PLOS Genet.* **18**, e1009564 (2022).

26. M. Paoli *et al.*, Minute Impurities Contribute Significantly to Olfactory Receptor Ligand Studies: Tales from Testing the Vibration Theory. *eNeuro*. **4** (2017), doi:10.1523/ENEURO.0070-17.2017.

27. A. Dravnieks, *Atlas of odor character profiles* (ASTM, Philadelphia, PA, 1985), *ASTM data series*.

28. M. H. Abraham, R. Sánchez-Moreno, J. E. Cometto-Muñiz, W. S. Cain, An Algorithm for 353 Odor Detection Thresholds in Humans. *Chem. Senses*. **37**, 207–218 (2012).

29. H. Zwaardemaker, *Die Physiologie Des Geruchs* (Рипол Классик, 1895).

30. H. Henning, *Der Geruch* (J. A. Barth, 1916).

31. E. C. Crocker, L. F. Henderson, Analysis and classification of odors: an effort to develop a workable method. *Am Perfum Essent Oil Rev*. **22**, 325 (1927).

32. M. Guillot, Physiologie des Sensations-Anosmies Partielles et Odeurs Fondamentales. *Comptes Rendus Hebd. Seances Acad. Sci.* **226**, 1307–1309 (1948).

33. J. E. Amoore, "A plan to identify most of the primary odors" in *Olfaction and Taste III* (Rockefeller University Press, New York, NY, 1969), vol. 158.

34. G. I. Parisi *et al.*, Continual lifelong learning with neural networks: A review. *Neural Netw.* **113**, 54–71 (2019).

35. E. J. Mayhew *et al.*, Transport features predict if a molecule is odorous. *Proc. Natl. Acad. Sci.* **119**, e2116576119 (2022).

36. T. B. Brown *et al.*, Language Models are Few-Shot Learners (2020), doi:10.48550/ARXIV.2005.14165.

37. G. Branwen, The Scaling Hypothesis (2020) (available at https://gwern.net/scaling-hypothesis).

38. RDKit, (available at https://www.rdkit.org/).

39. M. Meyners, S. R. Jaeger, G. Ares, On the analysis of Rate-All-That-Apply (RATA) data. *Food Qual. Prefer.* **49**, 1–10 (2016).

40. I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning" in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, San Francisco California, 2001; https://dl.acm.org/doi/10.1145/502512.502550), pp. 269–274.

41. D. Zügner, A. Akbarnejad, S. Günnemann, "Adversarial Attacks on Neural Networks for Graph Data" in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018; http://arxiv.org/abs/1805.07984), pp. 2847–2856.

42. A. Madry *et al.*, Towards Deep Learning Models Resistant to Adversarial Attacks (2017), doi:10.48550/ARXIV.1706.06083.

43. H. Boelens, J. Heydel, Chemical Composition and Smell-study on Structural Properties of Chemical Compounds with Different Odor Characteristics. *Chem.-Ztg.* **97**, 8–15 (1973).

44. M. Stoll, Many membered rings and musk odor. *Drug Cosmet. Ind.* **38**, 334–337 (1936).

Acknowledgements: The authors wish to acknowledge Zelda Mariet for experimental design, Yoni Halpern, Bob Datta, Steven Kearnes, Christina Zelano, Ari Morcos, Dan Bear, and Alex Koulakov for feedback on draft, contributions to GC-MS/O analysis from Dr. J.S. Elmore, and domain expertise from Christophe Laudemiel.

Funding:
Google Research
National Institutes of Health grant F32 DC019030 (EJM)
National Institutes of Health grant T32 DC000014 (EJM)

Author contributions:
Conceptualization: ABW
Methodology: BKL, EJM, RCG, JDM, BSL, JKP
Software: BKL, BSL, RCG, JNW
Validation: RCG, BKL
Formal analysis: BKL, EJM, RCG, JY
Investigation: EJM, KAL, MA, BBN, TM, JKP, JDM, BSL, WWQ, JNW
Data curation: BKL, BSL, RCG, EJM, MA, KAL, JKP
Writing – original draft: EJM, BKL, ABW, JDM, RCG
Writing – review & editing: EJM, BKL, ABW, JDM, RCG, JKP, BSL, WWQ, JNW
Visualization: EJM, RCG, BKL, BSL
Supervision: ABW, JDM, EJM, JKP
Funding acquisition: ABW, JDM, EJM
Project administration: ABW, JDM

Competing interests: The original work and funding for this manuscript was provided by Google Research. BKL, JNW, BSL, WWQ, RCG, and ABW were employees of Google at the time this study was conducted. During the review process, ABW, RCG and WWQ joined Osmo Labs, PBC, a new venture that is commercializing some of the technologies described in this manuscript. ABW, RCG, and WWQ each have an ownership interest in Osmo Labs, PBC, and receive a salary from the company. ABW is an officer of the company. JDM received funding from Google and serves on the Scientific Advisory Board of Osmo Labs, PBC. EJM, KAL, MA, and BBN received funding from Google.
Google has signed a transfer of ownership of all relevant IP (data, code, models, patents) to this new company. The details of this document are confidential, and unfortunately cannot be shared.

Data and materials availability: Human psychophysics data, model predictions, and model embeddings for tested odorants are included, and will be deposited at the olfactory data repository pyrfume.org upon publication; odorant identities will be released pending legal review upon publication. Lightweight reproduction notebooks, scripts, and data are shared at https://github.com/osmoai/publications. Datasets used for transfer learning analyses (Fig 5C-E), published in cited references (4, 6, 27, 28), are also available from pyrfume.org.
All material needed to critique and replicate this work will be made freely available, with the following exceptions as mandated by our agreement: ~45% of chemical structures used in the validation study.
The code implementing and training the GNN model is not available to share. However, all code and processed data to replicate every figure that does not require internal model state is provided. The model architecture and hyperparameters (often the most difficult properties in neural network modeling to discover) are provided.