A Principal Odor Map Unifies Diverse Tasks in Human Olfactory Perception

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

[1]Google Research, Brain Team; Cambridge, MA, USA.
[2]Department of Food Science and Human Nutrition, Michigan State University; East Lansing, MI, USA.
[3]Department of Computer Science, University of Illinois; Urbana-Champaign, IL, USA.
[4]Monell Chemical Senses Center; Philadelphia, PA, USA.
[5]Department of Food and Nutritional Sciences, University of Reading; Reading, Berkshire, UK.
[6]School of Life Sciences, Arizona State University; Tempe, AZ, USA.
[7]Department of Neuroscience, University of Pennsylvania; Philadelphia, PA, USA.

[†]These authors contributed equally.
*Corresponding authors. Email: jmainland@monell.org, alex.bw@googlemail.com

Abstract

Mapping molecular structure to odor perception is a key challenge in olfaction. Here, we use graph neural networks (GNN) to generate a Principal Odor Map (POM) that preserves perceptual relationships and enables odor quality prediction for novel odorants. The model is 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.
(119 words)

One-Sentence Summary

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

Introduction

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.

Results

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 is represented as a graph, with each atom described by its valence, degree, hydrogen count, hybridization, formal charge, and atomic number. Each bond is 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 use 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 represents known perceptual hierarchies and distances, 2) extends to novel odorants, 3) is robust to discontinuities in structure-odor distances, and 4) generalizes to other olfactory tasks.

To train the model, we curated a reference dataset of approximately 5000 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). The model (Fig. 1C) achieved strong cross-validation predictive performance of AUROC=0.89 (*15*).

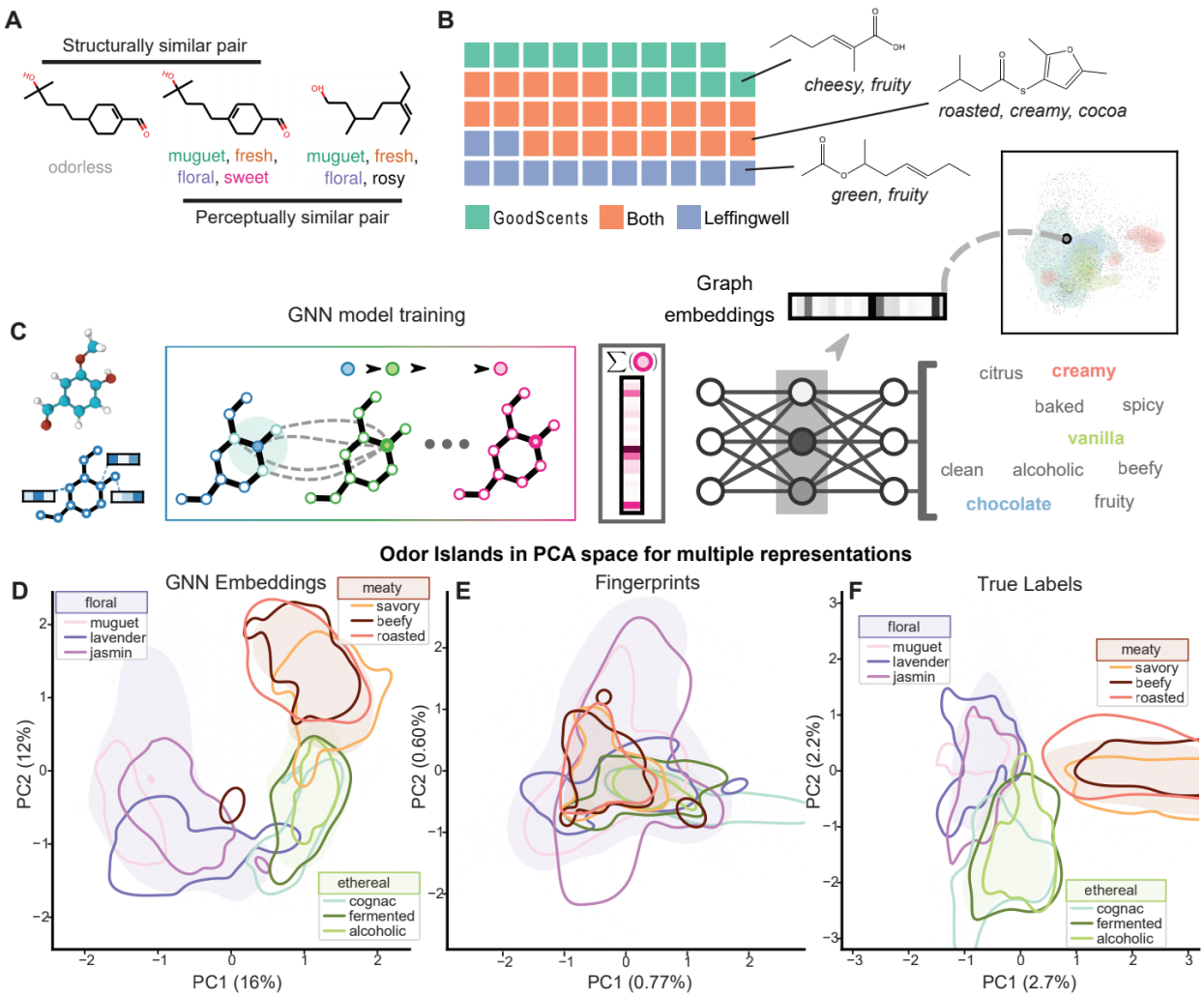

Fig. 1. POM preserves the structure of odor perceptual space. (A) Example triplet of molecules in which the structurally similar pair is not the perceptually similar pair. **(B)** The GNN was trained on a curated dataset of ~5000 semantically labeled molecules drawn from GoodScents (*13*) and Leffingwell (*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.

To test how well the POM represents known perceptual relationships, we compared both the POM and a map built with standard chemoinformatic features - Morgan fingerprints (FP) - to empirical perceptual space (Fig. 1D-F). We measured the fidelity of the maps in representing true relative 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). The POM better represents relative distances: distances in the perceptual map (Fig. 1D) are more significantly correlated to distances in the POM (R=0.73, Fig. S1A) than to distances in the FP map (R=-0.12, $p < 0.001$, Fig. S1B). The POM better represents perceptual hierarchies: molecules with a shared odor label have significantly tighter cluster density (CD) in the POM (CD = 0.51± 0.19) than in the FP map (CD = 0.68 ± 0.23, $p < 0.001$, Fig. S2), where smaller CD values denote more dense clusters.

To test if the model extends to novel odorants, we designed a prospective validation challenge (*16*) 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 (*17*). Like other sensory modalities, odor perception is variable across individuals (*18*, *19*), but group-averaged odor ratings have been shown to be stable across repeated measurements (*20*) 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. S3). 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. S4), 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. S5-7. Our panel's mean ratings were highly stable (panel test-retest: R = 0.80, n = 15; Fig. S8) and more consistent than the DREAM cohort's ratings (*6*) (Fig. S9-10).

Of the 400 molecules characterized, 80 were dropped from the final prospective validation set due to low intensity (42) (Fig. S11), redundancy (1), mistaken inclusion (1), or with confirmed or potential contamination (26) (Data S1). Model performance was evaluated on the remaining 320 molecules without model retraining.

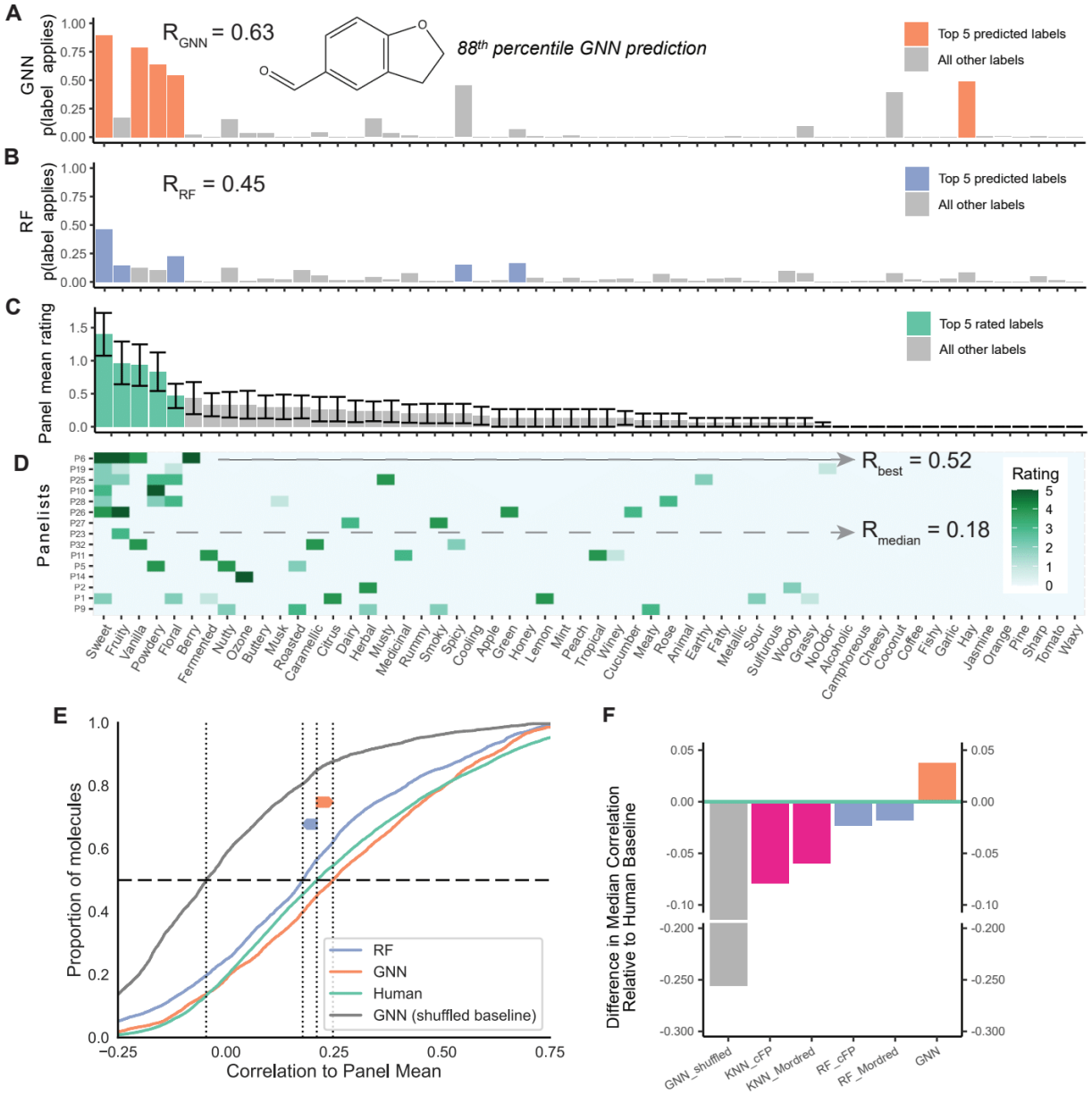

Fig. 2: GNN model displays human-level odor description performance. (A) GNN model label predictions, **(B)** random forest (RF) model label predictions, **(C)** panel mean ratings with standard error bars, and **(D)** individual panelist ratings, averaged over 2 replicates, for the molecule 2,3-dihydrobenzofuran-5-carboxaldehyde. 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.

To measure the model's performance, we compared the concordance of its normalized predictions with the normalized panel mean rating (Fig. 2A and 2C). 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). The model's superiority at the task is even more impressive given that panelists are able to smell each odorant as they rate it, while the model's predictions are 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, showing that our GNN model's performance increase comes not only from the volume and quality of the data, but importantly from the model architecture.

The GNN model shows human-level performance in aggregate, but how does it perform across perceptual and chemical classes? When we disaggregate performance by odor label, the model is within the distribution of human raters for all labels except musk and surpasses the median panelist for 32/55 labels (58%, Fig. 3A). This per-label view supports the view 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=1.0e-7).

Predictive performance for a given label depends on the complexity of the structure-odor mapping for that label, so it is 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) (*21*, *22*). 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. Weak agreement for musk may also be due to genetic variability in perception, a well-documented phenomenon (*23*).

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 is high for labels with many training examples (e.g, fruity, sweet, floral) (Fig. 3B), but performance for labels with few training examples can be either high (e.g., fishy, camphoreous, cooling) or low (e.g, ozone, sharp, fermented). In other words, collecting more training data raises the floor for model performance. Likewise, model performance is

bounded above by panel test-retest correlation (Fig. S13). When we disaggregate by chemical classes (e.g. esters, phenols, amines), both panelist and model performance is 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(*24*). 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. Our QC led to diverse conclusions: the nominal compound caused the odor (12/50), the nominal compound and contaminants contribute to the odor (16/50), contaminants caused the odor (18/50), or the cause of the odor could not be determined (4/50) (Fig. 3D). 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, through QC we determined that the nominal compound was floral only and that 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 26 molecules known or suspected to have high degrees of odorous contamination (Data S1).

The prevalence of odorous contamination that we found demonstrates that it is not safe to assume that the odor percept of a purchased chemical is due to the nominal compound. The Flavor & Fragrance (F&F) industry is motivated to minimize odorous contaminants for commercially valued odorants, but there is no such incentive for non-F&F commodity chemicals. We stress the need for caution and diligence in expanding odor stimulus space.

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.

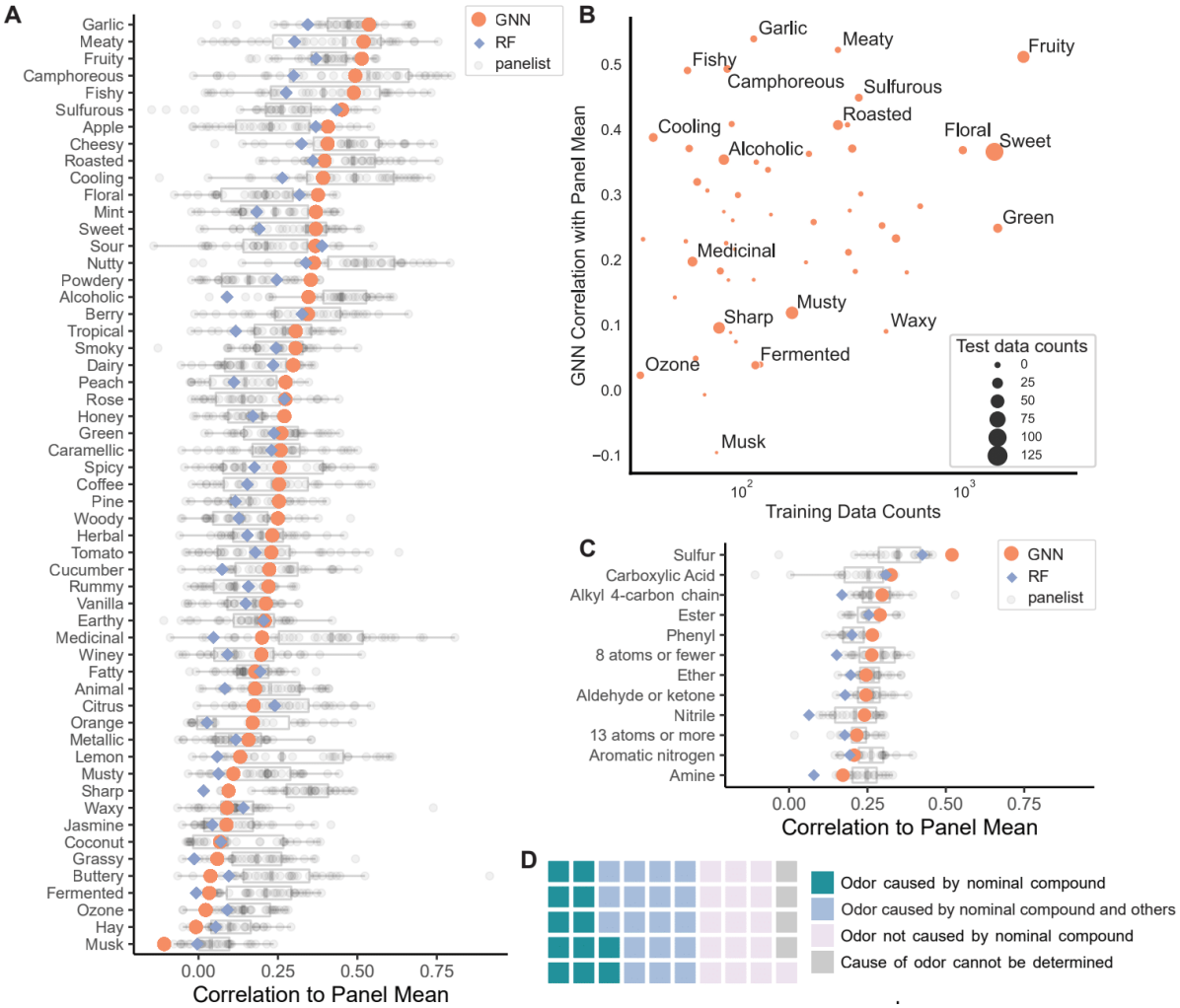

Fig. 3. Model performance is robust across structural and perceptual classes. (A) Correlation of GNN (in orange) and RF (in purple) model predictions and panelist ratings (in gray) to the panel mean for each of the 55 odor labels. **(B)** GNN model correlation to panel mean for each of the 55 odor labels 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 12 common chemical classes. **(D)** Categorization of gas chromatography-olfactometry quality control results for 50 validation set stimuli.

To test if the model is robust to discontinuities in structure-odor distances, we designed an additional challenge in which 41 new triplets (example in Fig. 4A) were constructed and validated by the panel (as in Fig. 1A). In each triplet, the anchor molecule is a known odorant, and is matched with one structurally similar and one structurally dissimilar novel odorant, and in which the more *structurally dissimilar* odorant is predicted to be the more *perceptually similar* of the two to the anchor. Our trained panelists were presented with the three odorants as a set and rated the perceptual distance between each of the molecules in the triplet (Fig. 4B). Confirming the model's predictions -- counterintuitive under simpler structural models of odor -- our panelists generally rated the *structurally dissimilar* molecules

as being more *perceptually similar* to an anchor molecule than the anchor's structural neighbor ($p < 2.2e-16$, Fig. 4C). This significant result is further evidence that the POM overcomes discontinuities in the structure-odor relationship.

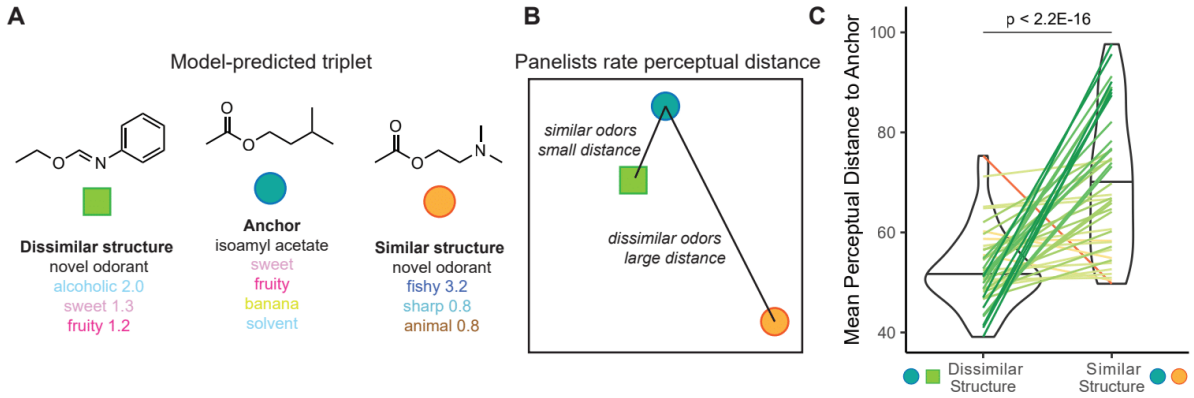

Fig. 4. POM solves discontinuities in structure-odor mapping. (A) Example triplet of molecules identified by the GNN model in which the structurally similar pair is not the perceptually similar pair. We used the model to select 41 such triplets. **(B)** Diagram of the psychophysical task in which panelists rated perceptual distances between molecules in predicted triplets. **(C)** Mean perceptual distance rating for molecules that are structurally dissimilar (left) or structurally similar (right) to the same anchor molecule. Lines connect each pair of molecules compared to the same anchor molecule; line color corresponds to the relative difference in perceptual similarity. Perceptual similarity followed model predictions rather than structural similarity.

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). These molecules would take approximately 70 person-years of continuous smelling time to collect using our trained human panel.

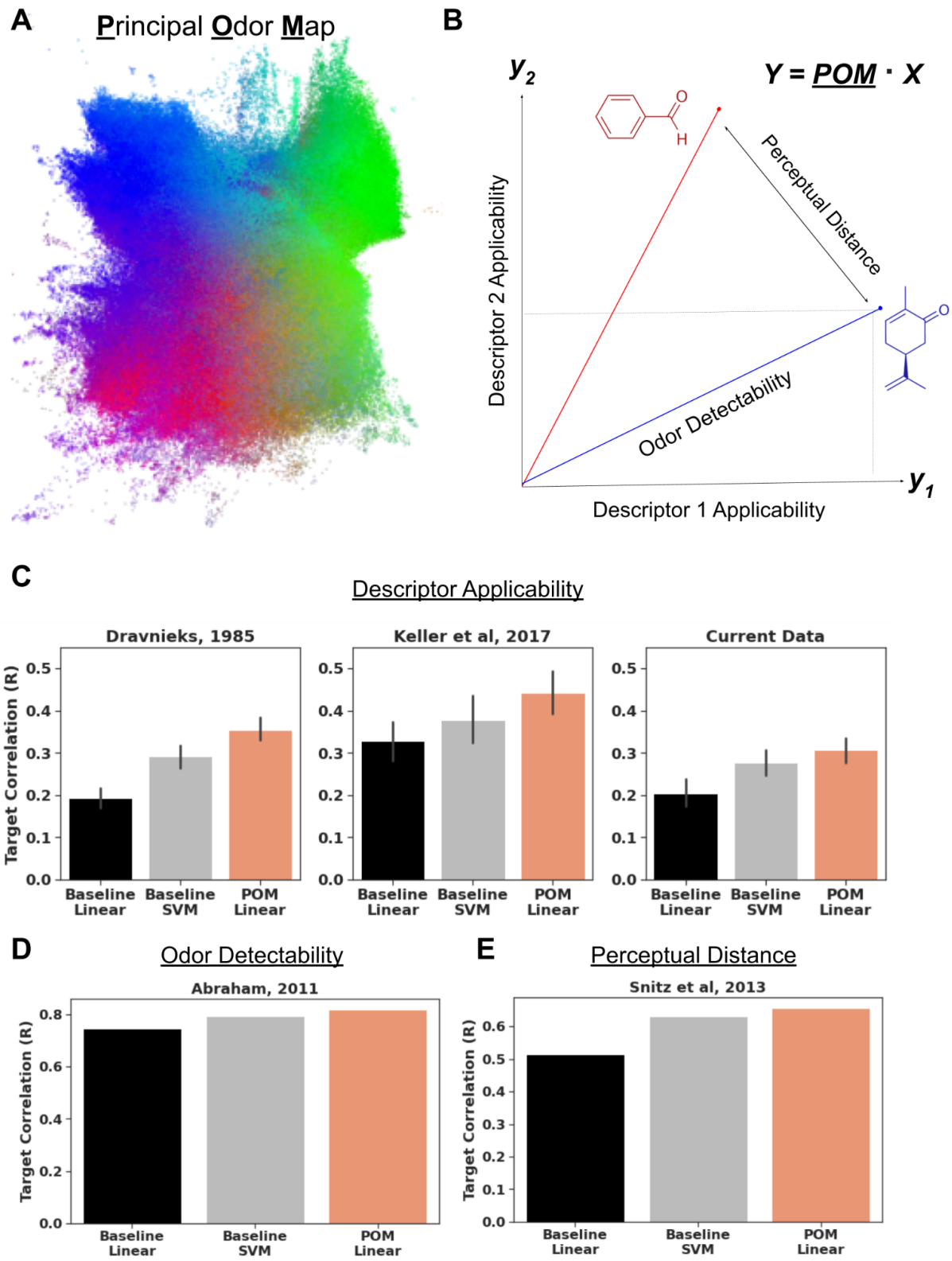

Fig. 5. POM solves a fundamental set of olfactory prediction tasks. (A) 2D trimap embedding of 500,000 unique likely odorants previously uncharacterized. The position of each point (molecule) is determined by POM coordinates, and the RGB values of each point correspond to their coordinates in the first 3 dimensions of a non-negative matrix factorization of the predicted odor labels. **(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 (*25*) 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 (*26*). **(E)** A linear model atop POM outperforms a chemoinformatic SVM baseline at predicting perceptual similarity on Snitz et al, 2013 (*4*).

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. S12), and odor similarity corresponds to the distance between such projects for different molecules. We find that a simple linear model applied to POM and using these geometric interpretations has 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.

Discussion

There is no universally accepted method for quantifying and categorizing an odor percept. In other words, olfaction has been a sense without a map. Systems of odor classification have been proposed: first intuitive categorizations (*28*), then empirically-supported universal spaces (*29*, *30*), and later attempts to incorporate receptor mechanisms (*31*, *32*). However, these systems do not tie stimulus properties to perception, and none have reached broad acceptance. Here we propose and validate a novel, data-driven, high-dimensional map of human olfaction. We have shown that this map recapitulates the structure and relationships of odor perceptual categories evoked by single molecules, that it can be used to achieve prospective predictive accuracy in odor description that exceeds that of the typical individual human, and that it is broadly transferrable to arbitrary olfactory perceptual tasks using natural and interpretable transformations. This map represents for odor what the CIE color space represents for vision.

Nearly all published chemosensory models were fit to the data used in their construction. Even using cross-validation, the opportunity for over-fitting is high, because the data comes from a single distribution, task, or experimental source. Prospective validation on new data from a new source with no adjustments, as we performed, 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. This process is called 'online learning' (*27*), and is a central capability of many real-world ML systems. Fig 5C demonstrates that a linear model atop POM reaches an even higher level of performance when the POM is tuned to the new dataset.

The success of this model is not merely an advance in predictive modeling. It offers a simple, intuitive, 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.

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. So 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 (*33*), 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 (*24*), particularly those that have not been developed for use in fragrance applications. Neural networks are known to 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. Finally, datasets in real-world settings are not static, but grow in size, and shift in distribution — models should be periodically retrained to incorporate new data. We showed that model performance tends to improve with increased training data (Fig. 3B) and data quality (Fig. S13), consistent with ML applications in other areas (*34*, *35*). Indeed, the most important future work -- work which will increase the accuracy and resolution of the map and any model that uses it -- will be scaling the volume and quality of training data.

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, faces in the superior temporal sulcus, 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. We have had no such map for odor, but 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.

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: 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; 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: BKL, JNW, BSL, WWQ, RCG, and ABW are employees of Google. We declare no competing financial interests.

Data and materials availability: Human psychophysics data, model predictions, and model embeddings for all 400 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 and scripts will be shared via https://github.com/google-research/google-research.

References:

1. C. R. des Séances, Commission Internationale de l'Eclairage (1931).
2. E. F. Evans, Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus.

Psychophysics and physiology of hearing (1977).
3. C. S. Sell, On the Unpredictability of Odor. *Angew. Chem. Int. Ed.* **45**, 6254–6261 (2006).
4. K. Snitz, A. Yablonka, T. Weiss, I. Frumin, R. M. Khan, N. Sobel, Predicting Odor Perceptual Similarity from Odor Structure. *PLoS Comput. Biol.* **9** (2013), doi:10.1371/journal.pcbi.1003184.
5. A. Ravia, K. Snitz, D. Honigstein, M. Finkel, R. Zirler, O. Perl, L. Secundo, C. Laudamiel, D. Harel, N. Sobel, A measure of smell enables the creation of olfactory metamers. *Nature*. **588**, 118–123 (2020).
6. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, P. Meyer, Predicting human olfactory perception from chemical features of odor molecules. *Science*. **355** (2017), doi:10.1126/science.aal2014.
7. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in *Machine Learning Meets Quantum Physics*, K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller, Eds. (Springer International Publishing, Cham, 2020), 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. *Communications of the ACM*. **60** (2017), pp. 84–90.
11. Schroff, Kalenichenko, Philbin, Facenet: A unified embedding for face recognition and clustering. *Proc. IEEE* (available at https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html).
12. N. Jaitly, P. Nguyen, A. Senior, V. Vanhoucke, Application of pretrained deep neural networks to large vocabulary speech recognition (2012) (available at https://research.google/pubs/pub38130/).
13. W. Luebke, The Good Scents Company Information System. *Online Access: http://www. thegoodscentscompany. com* (2019).
14. J. C. Leffingwell, Leffingwell & Associates (2015), (available at https://www.leffingwell.com/cooler_than_menthol.htm).
15. B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *arXiv preprint arXiv* (2019) (available at https://arxiv.org/abs/1910.10685).
16. S. Kearnes, Pursuing a Prospective Perspective. *TRECHEM*. **3**, 77–79 (2021).
17. H. T. Lawless, H. Heymann, in *Sensory Evaluation of Food: Principles and Practices*, H. T. Lawless, H. Heymann, Eds. (Springer New York, New York, NY, 2010), pp. 227–257.
18. C. Trimmer, A. Keller, N. R. Murphy, L. L. Snyder, J. R. Willer, M. H. Nagai, N. Katsanis, L. B. Vosshall, H. Matsunami, J. D. Mainland, Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proc. Natl. Acad. Sci. U. S. A.* (2019), doi:10.1073/pnas.1804106115.
19. A. Keller, M. Hempstead, I. A. Gomez, A. N. Gilbert, L. B. Vosshall, An olfactory demography of a diverse metropolitan population. *BMC Neurosci.* **13**, 122 (2012).
20. A. Dravnieks, Odor quality: semantically generated multidimensional profiles are stable. *Science*. **218**, 799–801 (1982).
21. K. J. Rossiter, Structure-odor relationships. *Chem. Rev.* **96**, 3201–3240 (1996).
22. O. R. P. David, A Chemical History of Polycyclic Musks. *Chemistry*. **26**, 7537–7555 (2020).
23. B. Li, M. L. Kamarck, Q. Peng, F.-L. Lim, A. Keller, M. A. M. Smeets, J. D. Mainland, S. Wang, From musk to body odor: decoding olfaction through genetic variation. *PLOS Genetics* (2022), doi:10.1101/2021.04.27.441177.
24. M. Paoli, D. Münch, A. Haase, E. Skoulakis, L. Turin, C. G. Galizia, Minute Impurities Contribute Significantly to

Olfactory Receptor Ligand Studies: Tales from Testing the Vibration Theory. *eneuro*. **4**, ENEURO.0070–17.2017 (2017).
25. A. Dravnieks, Atlas of odor character profiles (1985).
26. 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).
27. G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review. *Neural Netw.* **113**, 54–71 (2019).
28. H. Zwaardemaker, *Die Physiologie Des Geruchs* (Рипол Классик, 1895).
29. H. Henning, *Der geruch* (JA Barth, 1916).
30. 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).
31. M. Guillot, Anosmies partielles et odeurs fondamentales. *C. R. Acad. Sci.* . **226**, 1307–1309 (1948).
32. J. E. Amoore, C. Pfaffmann, A plan to identify most of the primary odors. *Olfaction & Taste III*, 158–171 (1969).
33. E. J. Mayhew, C. J. Arayata, R. C. Gerkin, B. K. Lee, J. M. Magill, L. L. Snyder, K. A. Little, C. W. Yu, J. D. Mainland, Transport features predict if a molecule is odorous. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2116576119 (2022).
34. Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever, Amodei, Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* (available at https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).
35. Gwern, The Scaling Hypothesis. *gwern.net*, (available at https://www.gwern.net/Scaling-hypothesis).
36. RDKit: Open-source cheminformatics, (available at http://www.rdkit.org).
37. M. Meyners, S. R. Jaeger, G. Ares, On the analysis of Rate-All-That-Apply (RATA) data. *Food Qual. Prefer.* **49**, 1–10 (2016).
38. N. Hauser, P. Kraft, E. M. Carreira, The Serendipitous Discovery of a Rose Odorant. *Chimia* . **74**, 247–251 (2020).
39. M. Stoll, *Drug Cosmet. Ind.* **38**, 334–337 (1936).
40. G. Ohloff, W. Pickenhagen, P. Kraft, *Scent and Chemistry: The Molecular World of Odors* (Wiley-VCH, ed. 1, 2012).

Supplemental Methods

Training dataset

The GoodScents (http://www.thegoodscentscompany.com/) and Leffingwell PMP 2001 (https://zenodo.org/record/4085098#.YqoYk8jMIUE) datasets each contain odorant molecules and corresponding odor descriptors. Variations and misspellings of odor descriptors were merged, and any odor descriptor with <=30 occurrences in the dataset were discarded. The remaining list of odor descriptors is: [

'alcoholic', 'aldehydic', 'alliaceous', 'almond', 'amber', 'animal',
'anisic', 'apple', 'apricot', 'aromatic', 'balsamic', 'banana', 'beefy',
'bergamot', 'berry', 'bitter', 'black currant', 'brandy', 'burnt',
'buttery', 'cabbage', 'camphoreous', 'caramellic', 'cedar', 'celery',
'chamomile', 'cheesy', 'cherry', 'chocolate', 'cinnamon', 'citrus', 'clean',
'clove', 'cocoa', 'coconut', 'coffee', 'cognac', 'cooked', 'cooling',
'cortex', 'coumarinic', 'creamy', 'cucumber', 'dairy', 'dry', 'earthy',
'ethereal', 'fatty', 'fermented', 'fishy', 'floral', 'fresh', 'fruit skin',
'fruity', 'garlic', 'gassy', 'geranium', 'grape', 'grapefruit', 'grassy',
'green', 'hawthorn', 'hay', 'hazelnut', 'herbal', 'honey', 'hyacinth',
'jasmin', 'juicy', 'ketonic', 'lactonic', 'lavender', 'leafy', 'leathery',
'lemon', 'lily', 'malty', 'meaty', 'medicinal', 'melon', 'metallic',
'milky', 'mint', 'muguet', 'mushroom', 'musk', 'musty', 'natural', 'nutty',
'odorless', 'oily', 'onion', 'orange', 'orangeflower', 'orris', 'ozone',
'peach', 'pear', 'phenolic', 'pine', 'pineapple', 'plum', 'popcorn',
'potato', 'powdery', 'pungent', 'radish', 'raspberry', 'ripe', 'roasted',
'rose', 'rummy', 'sandalwood', 'savory', 'sharp', 'smoky', 'soapy',
'solvent', 'sour', 'spicy', 'strawberry', 'sulfurous', 'sweaty', 'sweet',
'tea', 'terpenic', 'tobacco', 'tomato', 'tropical', 'vanilla', 'vegetable',
'vetiver', 'violet', 'warm', 'waxy', 'weedy', 'winey', 'woody'

]

These datasets were merged and are subsequently referred to as “GS/LF”.

Model Training and Tuning

The network consists of several message-passing layers, followed by a radius 0 combination to fold atom and bond embeddings together, followed by a reduce-sum across atoms, followed by several fully connected layers and a final sigmoid function to make label predictions for each of the 138 curated descriptors described above.

All references to the GNN “embedding space” refer to the 256-dimensional activation of the final dense neural network layer. These embeddings are mapped to the final 138-dimensional prediction by one final dense layer of 138 neurons followed by a sigmoid function.

Hyperparameters of the neural network were optimized using 5-fold cross validation in our training set of ~4,000 molecules, using 500 trials of random search. Each model fit took less than 1 hour on a Tesla P100. We present results for the model with the highest mean AUROC on the cross-validation set. Since our multi-label problem had highly unbalanced labels, we used second-order iterative stratification to build our train/test/validation splits [47]. Iterative stratification is a procedure for stratified sampling that attempts to preserve many-order label ratios, prioritizing more unbalanced combinations. For second order, this means preserving ratios of pairs of labels in each split.

The objective function for training was a summed cross-entropy loss over all 138 descriptors, with each descriptor's contribution to the loss being weighted by a factor of log(1+ class_imbalance_ratio), such that rarer descriptors were given a higher weighting. l1 and l2-norm losses were also utilized.

For our random forest (RF) baseline methods, we tuned an exhaustive space of configurations of fingerprinting methods (bits, radius, counted/binary, RDKit/Morgan), and RF hyperparameters. The RDKit software(*36*) was used to calculate all features. We found a radius-4, 2048-bit Morgan fingerprint to perform most strongly in predicting odor labels.

Using an 80-20 stratified train/test split, we found that a trained GNN achieved an AUROC of 0.894 [CI 0.888 - 0.902] on the combined GS/LF dataset, whereas RF on Morgan fingerprints was the strongest baseline method with an AUROC of 0.850 [CI 0.838 - 0.860]. (*15*)

Model- and label-space comparisons

After applying a top-2 principal component reduction to GNN embedding, Morgan fingerprints, and perceptual label spaces, we found that GNN embedding inter-cluster distances correlate strongly (r = 0.725) with label inter-cluster distances, whereas Morgan fingerprints correlate weakly (r = -0.119). Inter-cluster distances are computed as the mean Euclidean distance of all |Cluster 1| * |Cluster 2| pairwise molecule distances. Fig S1A and S1B show all (138^2 - 138) pairwise comparisons of odor clusters.

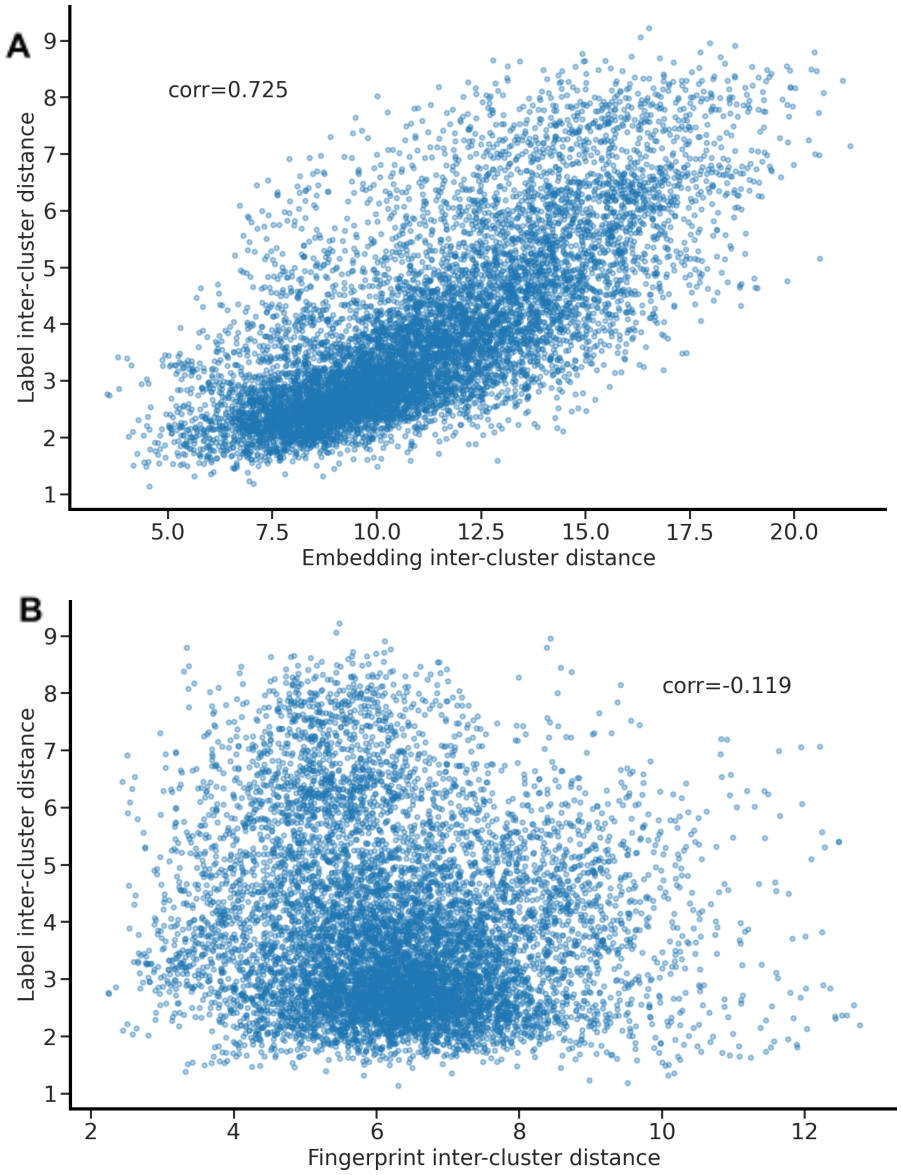

Fig. S1. Inter-cluster distance correlation. Each point represents a pair of odors (e.g. [fruity, sweet]. **(A)** Cluster distance, as measured by Euclidean distance of embeddings, correlates strongly with Jaccard overlap of clusters. **(B)** Cluster distance, as measured by Euclidean distance of fingerprints, does not correlate with Jaccard overlap of clusters.

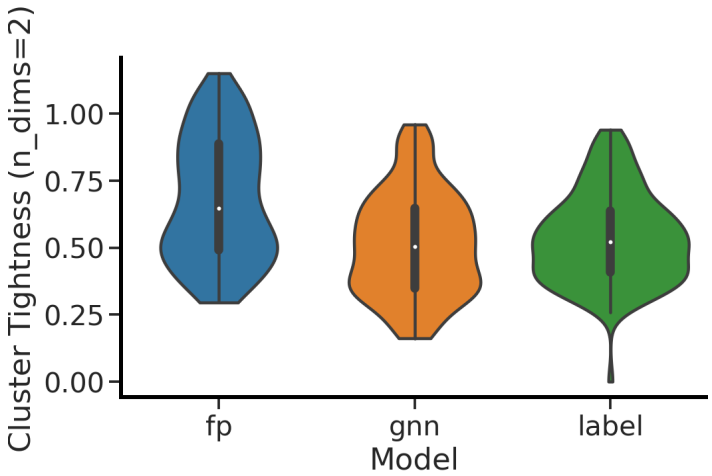

Fig. S2. Distribution of cluster tightness across 138 odor classes. Cluster tightness is defined as the ratio of {mean Euclidean distance of all in-group pairwise distances} to {mean Euclidean distance of all in/out-group pairwise distances}. 50th percentile cluster tightness for GNN embedding space was 0.513, similar to cluster tightness for label space (0.536), and tighter than fingerprint space (0.679).

Collecting a Prospective Validation Set

Aroma lexicon

We selected 55 of the 138 common labels from the GS/LF dataset to form our lexicon. We prioritized the selection of broad category terms (e.g., fruity, floral) to span the range of possible odor percepts, but also included specific terms from within an aroma category to measure model precision (e.g., fruity/citrus/lemon). Hierarchical clustering of GS/LF data supported our selections. Subjects used this lexicon exclusively to describe their odor quality perception of the odorants. The final list of odor descriptors used during human labeling is presented in Data S1.

Table S1. Lexicon and associated aroma references. The 55 descriptors in the lexicon were organized so that perceptually similar terms were close together in the list; terms are listed in the order they appeared on rater's ballots. Each term was paired with one or more aroma references to facilitate rater training.

Lexicon Term	Aroma References
green	Le Nez Du Vin aroma standard - Vegetal
grassy	2% cis-3-hexenol solution
cucumber	1% nona-2,6-dienal solution
tomato	AromaMasters aroma standard - Tomato
hay	Animal bedding
herbal	Herbs de Provence; Dried dill
mint	Dried mint
woody	AromaMasters aroma standard - Cedar; Pine shavings
pine	5% alpha-pinene solution
floral	AromaMasters aroma standard - Linden; AromaMasters aroma standard - Honeysuckle
jasmine	Jasmine essential oil
rose	Rose water
honey	Honey
fruity	Good & Gather flavor fusion fruit strips
citrus	Combined essential oils of orange, grapefruit, lime, and lemon
Lemon	Lemon essential oil

orange	Orange essential oil
tropical	Trident tropical twist gum
berry	Le Nez Du Vin aroma standard - Bilberry
peach	Le Nez Du Vin aroma standard - Peach
apple	Jolly Rancher - green apple
sour	White vinegar
fermented	GT's original kombucha
alcoholic	200 proof ethanol
winey	Sutter Home red wine blend
rummy	Oakheart spiced rum
caramellic	Caramel flavor extract
vanilla	Vanilla extract
spicy	Blend of ground cinnamon, nutmeg, cloves, and allspice; Ground black pepper
coffee	Folgers medium roast ground coffee
smoky	5% guaiacol solution
roasted	Le Nez Du Vin aroma standard - Toasted
meaty	Le Nez Du Vin aroma standard - Cooked beef
nutty	Roasted mixed nuts
fatty	10% (E,E)-2,4-decadienal solution
coconut	AromaMasters aroma standard - Coconut
waxy	Crayola crayon
dairy	Carnation half & half pods
buttery	Butter extract
cheesy	Kernel Seasons white cheddar powder
sulfurous	Le Nez Du Vin aroma standard - Rotten egg
garlic	Garlic powder
earthy	Peat moss
musty	0.1% 2,4,6-tribromoanisole solution
animal	AromaMasters aroma standard - Horse sweat
musk	Solution of galaxolide, ethylene brassylate, and tonalide
powdery	Johnson & Johnson baby powder
sweet	Charms cotton candy
cooling	Menthol crystals
sharp	10% acetic acid solution
medicinal	Vicks VapoRub
camphoreous	1% camphor solution
metallic	Pennies to be rubbed against skin
ozone	Adoxal
fishy	0.5% trimethylamine solution

Panelist training and screening

A pool of 26 prospective panelists between the ages of 18 and 55 and with a normal sense of smell were recruited from the Philadelphia area to participate in a 5-session series of training and screening exercises. The research protocol was approved by the University of Pennsylvania IRB, and all subjects gave informed consent prior to enrolling in the study. Subjects received odorant kits shortly before the start of the experiment and participated in sessions from home, facilitated by an experimenter over a Zoom video call (Zoom Meetings, https://zoom.us). The initial odorant kit (Fig. S3) contained 58 odor references (Table S1), 10 blinded odor references used for training quizzes, and 20 common odorants used in screening exercises (Data S2).

Figure S3. Initial odorant kit, containing 58 unique odor references corresponding to odor labels in the lexicon (rows 1-6, 8), 10 blinded odor references (row 7), and 20 blinded common odorants (rows 9-10).

In the first session, subjects were introduced to the study and trained to use the rate-all-that-apply (RATA) method to describe their perception of odorants (*37*); in the RATA method, subjects choose from a list terms that apply to the sample being evaluated, and then rate how strongly the chosen terms apply to the sample from 1 (low/slightly applicable) to 5 (high/very applicable). Subjects were given guidance on how to evaluate the odorants (e.g. take several short sniffs, hold the vial far from the nose to start and gradually bring it closer while sniffing, keep the cap tightly on each vial while not actively evaluating it) and taught to use a standard nasal rinse protocol (wet clean washcloth until just saturated, heat in microwave for 60s to produce steam, breathe in moist air above washcloth through nose for 30 seconds between evaluations, re-warm washcloth as needed; washcloths were provided with odor kits). Subjects then evaluated 20 common odorants using the RATA method as a pre-test.

In the second and third sessions, researchers trained subjects on the meaning of labels in the aroma lexicon. For each label, researchers described the olfactory meaning of the label, showed a related image, and prompted subjects to smell the associated odor reference(s). At the end of each session, subjects participated in a quiz in which they tried to identify blinded odor references and mixtures of references. Researchers then revealed the true identity of the blinded references and led a discussion about the results, prompting subjects to re-smell references and reinforce label meanings.

Following training, subjects evaluated the 20 common odorants in the fourth and again in the fifth session. Researchers reviewed subjects' quiz responses label selections for the 20 common odorants and calculated their test-retest correlation for post-training ratings. We invited 18 subjects who met our test-retest criterion (R>0.35) and made reasonable label selections for common odorants (e.g. mint selected to describe (-)-carvone) to join our panel (12 female, 6 male; 12 Caucasian, 4 African American or Black, 2 Asian; 3 Hispanic or Latino/a).

Virtual screening protocol for molecule selection

We began by filtering molecules listed in the eMolecules catalog -- which contains ~1 million commercially available molecules -- for atom composition (C/N/O/S/H only), price (<$1000 per 10 grams), purity (>95%), and availability (<4 weeks lead time). We developed a toxicity filter to conservatively remove potentially irritating or harmful compounds, (protocol developed by a certified toxicologist, approved by the University of Pennsylvania IRB), and removed likely odorless molecules according to water-soluble (cLogP < 0) and nonvolatile (boiling point > 300 C) criteria. We manually removed molecules that were likely to degrade or react under our experimental conditions. Finally, we compared predicted odor descriptors to the odor descriptors of all structurally similar reference molecules. All selected molecules satisfied one of two criteria:

1. Structurally similar to a molecule in the reference GS/LF dataset, yet with a negative prediction for that molecule's given descriptors. Prediction thresholds for descriptors were set at a threshold according to a geometric mean of training data frequency and test data empirical label frequency.
2. Structurally dissimilar to all molecules in the reference GS/LF dataset having a particular descriptor.

We selected and purchased 580 structurally distinct molecules from these structurally/perceptually divergent candidates (Fig. S4). Upon receipt of purchased molecules, we manually inspected for odorless molecules and diluted with propylene glycol to manually intensity-balanced each sample. 400 molecules were evaluated by human panelists, and the remainder of molecules were not tested further. Selection rationale for the 400 molecules is noted in Data S1.

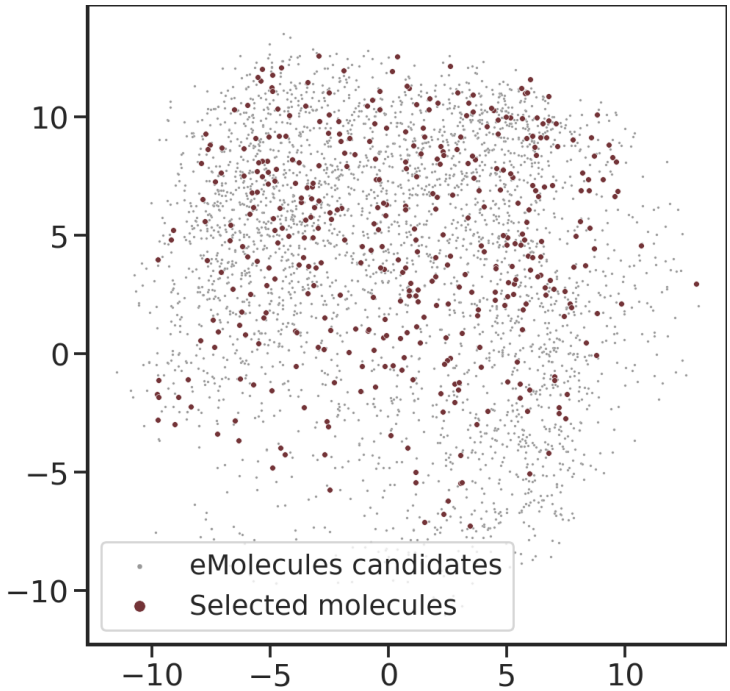

Fig. S4. Candidate and selected molecules from the eMolecules catalog displayed in chemical feature PCA space. Selected molecules span the space of potential candidate molecules from eMolecules.

Odor evaluations

Invited panelists were asked to rate the applicability of the 55 odor labels for each sample using the RATA method, as well as rate the intensity and pleasantness of each sample. Panelists received the odorants in sets of 50 and evaluated each twice over 4 sessions (25 evaluations/session). There was an enforced 30s break between each evaluation, and subjects followed the standard nasal rinse protocol described above during that break. In total, we characterized 400 novel odorants using this approach, and at least 15 of the initial 18 panelists participated in each phase of the study such that n ≥ 15 for each odorant*replicate.

Against the backdrop of a global COVID-19 pandemic, we were wary of COVID-induced anosmia. Each session began with a training warm-up exercise to engage panelists in the rating task, reinforce label meanings, and enable researchers to verify that panelists had a normally functioning sense of smell. No subjects became anosmic during the course of the study.

Overall, we collected 400 molecules X 55 odor classes X 15 panelists X 2 replicates = 660,000 human sensory data points. The raw ratings are provided in Data S3, and summary statistics are described in Fig. S5. The distribution of non-zero descriptor ratings is shown in Fig. S5A, and the distribution of the number of descriptors applied to each molecule is shown in Fig. S5B. Each molecule is typically assigned between 1-6 descriptor ratings by each rater. Most descriptors are used at least once by every rater. Fig. S5C shows the percentage of molecules that are described by each of the 55 terms in the lexicon. Sweet was the most commonly applied descriptor.
Descriptor ratings show a clear correlation structure (Fig. S6). For example, fruity descriptors are more likely to co-occur with each other and less likely to co-occur with other descriptors including meaty, sulfurous, and roasted. Descriptor ratings are also related to odorant chemical class (Fig. S7). For example, molecules containing a sulfur atom are more likely to be described as meaty, molecules containing an amine group are more likely to be described as fishy, and molecules containing a carboxylic acid group are more likely to be described as sour.

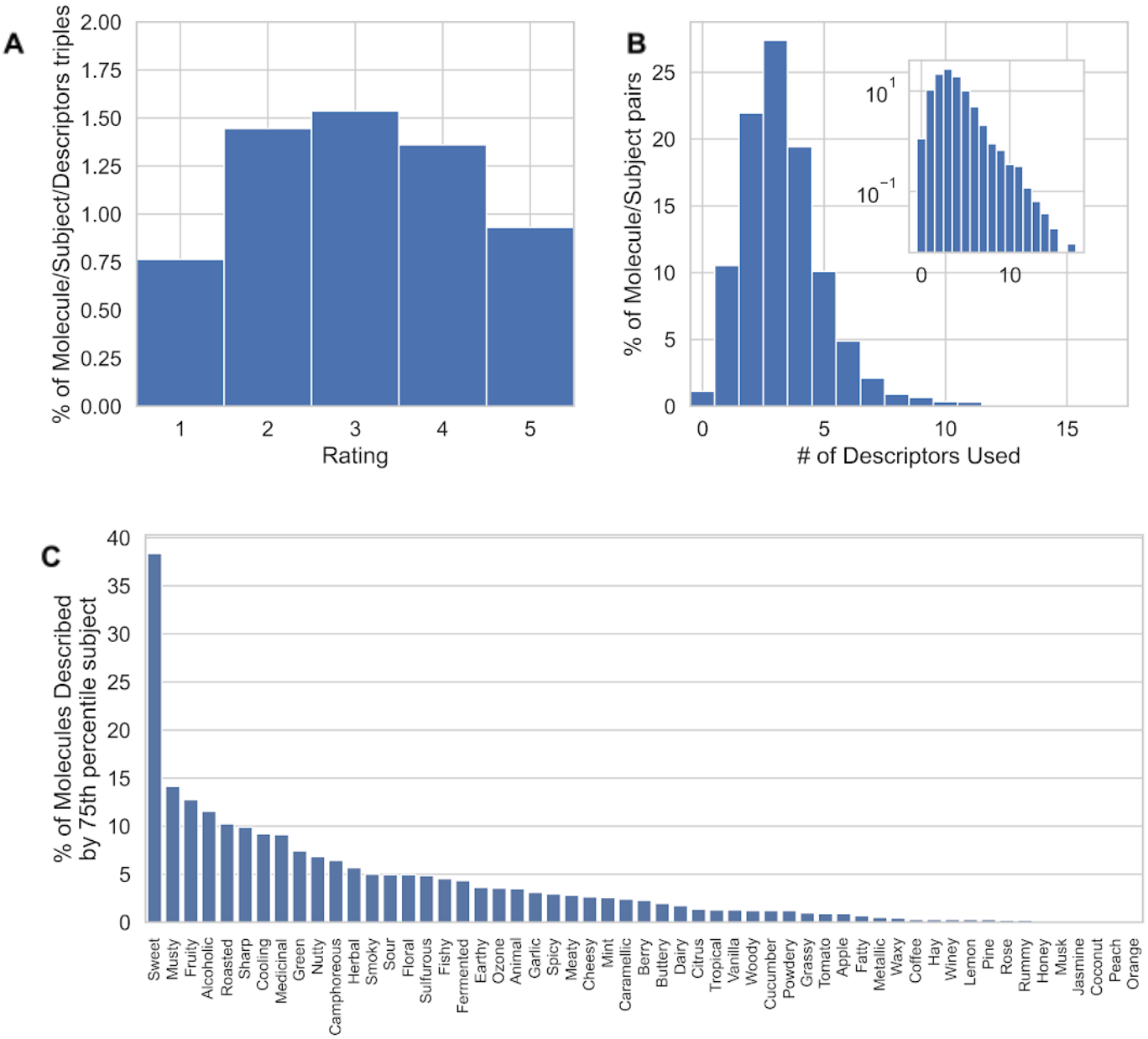

Fig. S5. Human psychophysics prospective validation set summary statistics. (A) Distribution of non-zero descriptor ratings, **(B)** distribution of the number of descriptors applied to each molecule, and **(C)** percent of molecules described by each of the 55 odor descriptors according to the 75th percentile panelist's ratings.

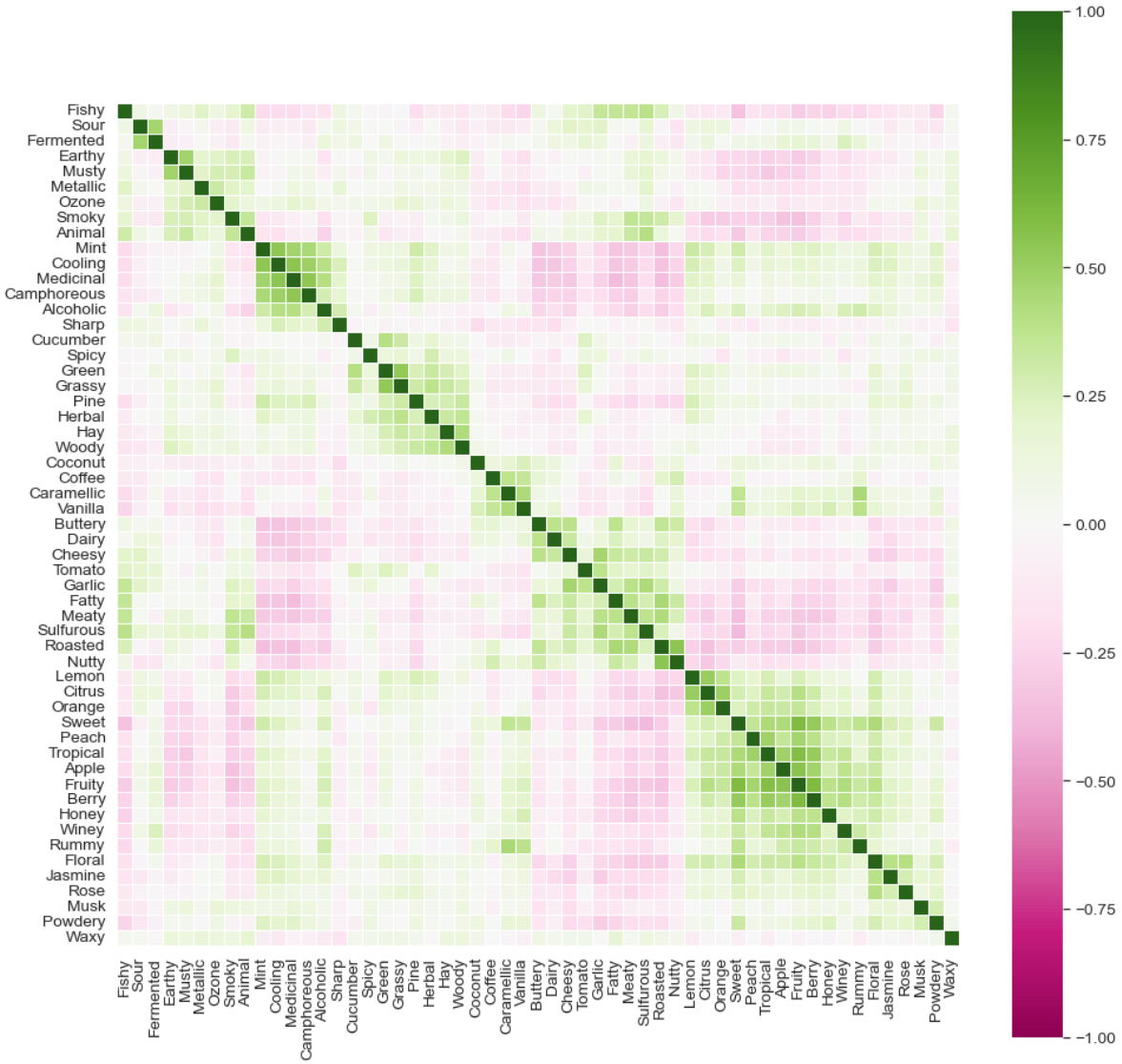

Fig. S6. Correlation matrix of panelist ratings for the 55 odor lexicon descriptors. Descriptors with strong positive correlations are dark green; descriptors with strong negative correlations are in dark pink. Odor descriptors show a clear correlation structure.

	Sulfur	Amine	Phenyl	Nitrile	Ester	Carbonyl	Carboxylic Acid	Alkyl	<=8 atoms (<25%ile)	>=13 atoms (>75%ile)
roasted_cluster	0.184	0.187	0.054	0.064	0.080	0.064	0.061	0.045	0.212	0.113
meaty_cluster	0.244	0.177	0.083	0.039	0.153	0.088	0.068	0.069	0.171	0.134
fishy_cluster	0.020	0.301	0.022	0.033	0.037	0.068	0.000	0.162	0.065	0.059
primeval_cluster	0.071	0.203	0.280	0.031	0.056	0.088	0.054	0.130	0.105	0.082
gourmand_cluster	0.000	0.018	0.050	0.209	0.043	0.078	0.054	0.000	0.072	0.026
herbal_cluster	0.010	0.062	0.224	0.088	0.048	0.067	0.021	0.113	0.072	0.112
fruity_cluster	0.006	0.107	0.234	0.113	0.204	0.178	0.054	0.088	0.108	0.181
floral_cluster	0.020	0.017	0.161	0.107	0.107	0.039	0.021	0.082	0.010	0.111
sour_cluster	0.037	0.059	0.037	0.015	0.061	0.064	0.262	0.064	0.096	0.068
cooling_cluster	0.008	0.148	0.148	0.055	0.114	0.160	0.043	0.118	0.154	0.065

Fig S7. Correlation of structural and perceptual categories. Each of the 400 molecules in the human validation set is non-exclusively classified according to structural and perceptual categories, and each table entry represents the Jaccard overlap (intersection over union) of molecule sets.

Rater performance

Aggregating all molecules and descriptors, our panel exhibited a test-retest correlation of 0.8. The panel test-retest correlation was high for most descriptors (Fig. S8). Compared to a prior large-scale human psychophysical study (6), our collected dataset has more descriptors and higher panel mean test-retest reliability (Fig. S9 and Fig. S10), even with fewer panelists.

Molecules with lower rated intensity were found to have the weakest panel test-retest correlations, indicating that the panel was not able to get a consistent evaluation (Fig. S11). We therefore excluded any molecules with intensity <3 (on a 0-10 scale) from the study. Of the 400 molecules evaluated by the panel, 42 were dropped from the validation set due to low odor intensity.

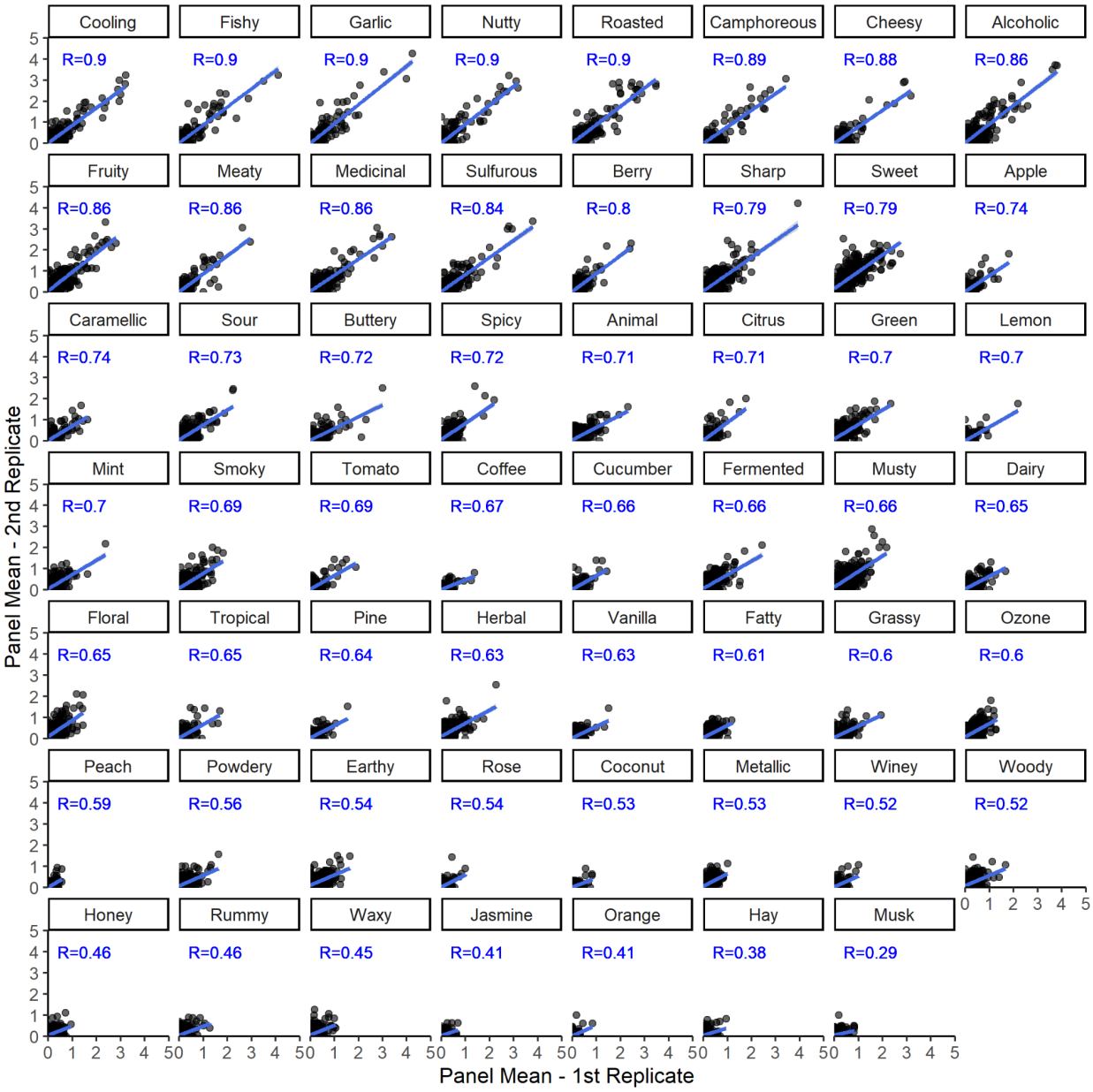

Fig. S8. Panel mean (n≥15 subjects) test-retest correlation (R) for the 55 descriptors in the lexicon applied to the 400 novel odorants in the prospective validation set. Descriptors are ordered by descending correlation.

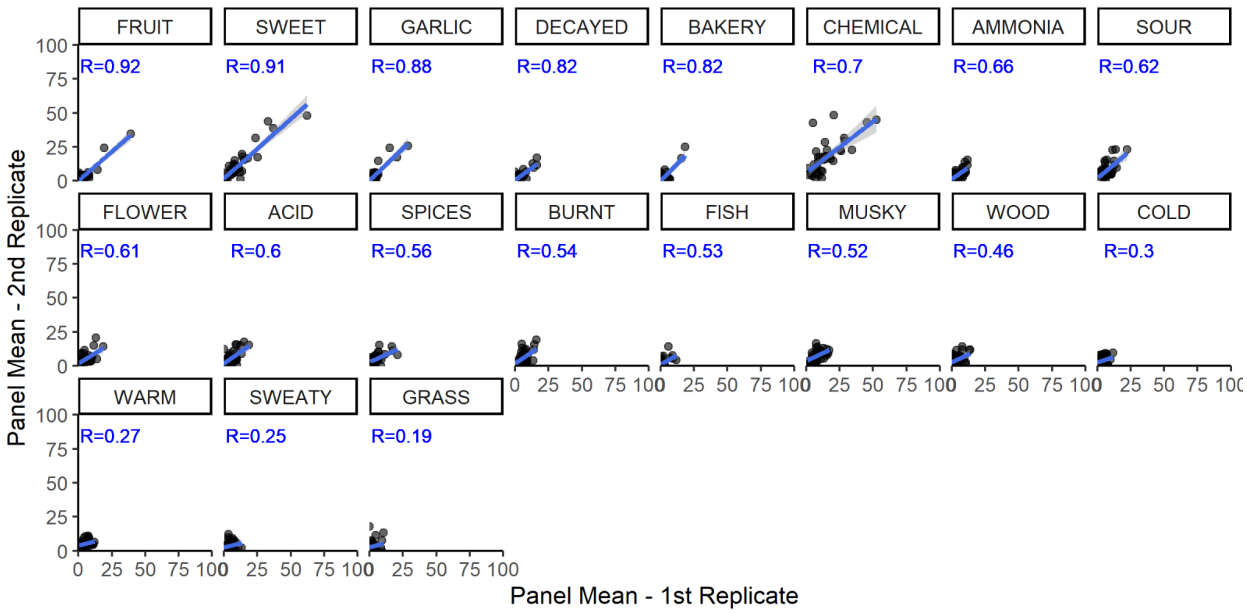

Fig. S9. Panel mean (n=49) test-retest correlation (R) for the 19 descriptors in the DREAM olfaction challenge dataset (6). Descriptors are ordered by descending correlation.

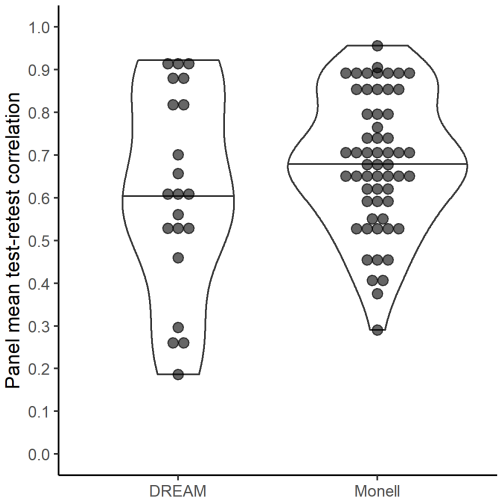

Fig. S10. Panel mean test-retest correlation for the 19 descriptors in the DREAM olfaction challenge (left) (6) and for the 55 descriptors in the present study (right). Each dot represents one odor descriptor.

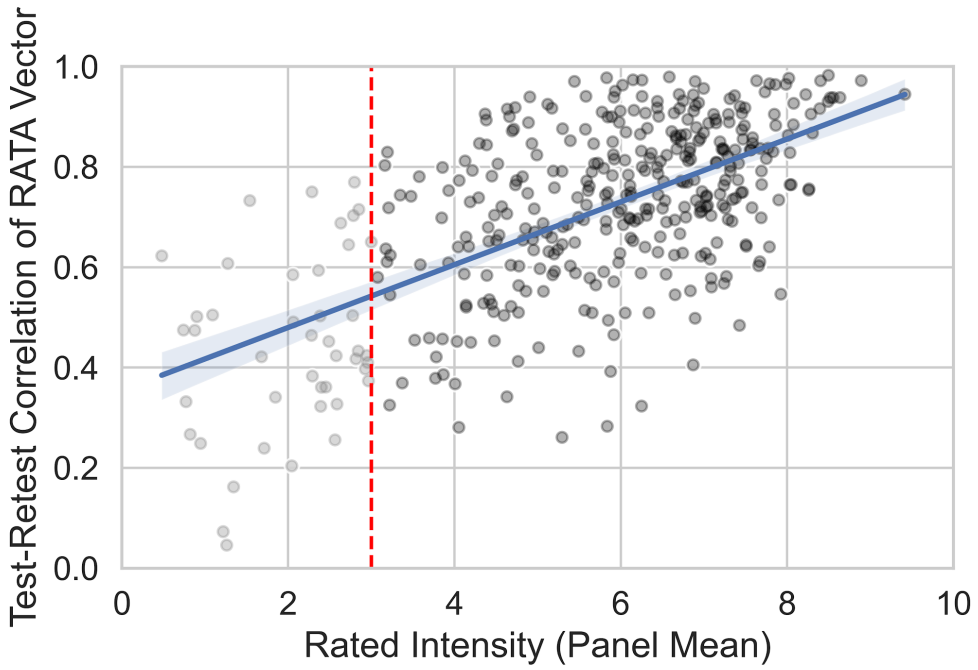

Fig S11. Test-retest correlation for 400 novel odorants as a function of panel mean intensity rating for that odorant. Molecules with lower rated intensity have weaker test-retest correlation of the panel mean.

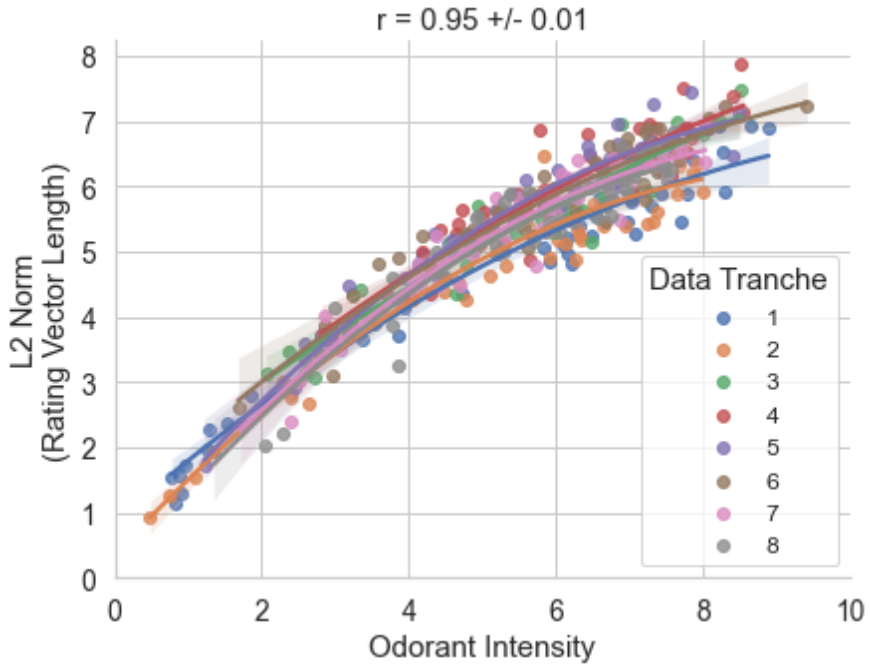

Fig. S12. L2 norm length of the mean RATA rating vector as a function of panel mean rated odorant intensity. Human psychophysical ratings were gathered in 8 data collection waves; differently colored dots and fits come from different tranches of data collection. Panelists use more descriptors and give higher RATA ratings for higher-intensity stimuli.

Evaluating Model Performance on Prospective Validation Set

We chose cosine similarity of a 55-dimensional vector as a metric that would emphasize overall accuracy of the predicted odor profile, rather than a "hit rate" on individual descriptors. This metric encountered some initial difficulties, as we discovered that the model and panelists were not directly comparable. The model makes an independent prediction for all 138 odor classes, resulting in a dense vector, whereas panelists typically rated only the top 3.2 ± 1.7 labels per odorant, resulting in sparse vectors. When shuffled, the model's predictions had a positive nonzero score, indicating a systematic scoring bias in the model's favor. We found that subtracting each individual model or panelists' mean rating across all molecules from the respective predictions had the effect of zeroing out the shuffled baseline's cosine similarity scores. We used this centered prediction or rating as the input to all of our calculations, as it would be a fairer comparison between model and panelist. Mathematically, this is similar to a Pearson correlation calculation, as the ratings are centered, but different due to not rescaling, as this would have destroyed useful information.

One disadvantage of the cosine metric is that it treats all 55 dimensions equally, yet not all mistakes are equally wrong. Descriptors have hierarchical relationships, and as such a "partial credit filter" -- which spreads observed single descriptor ratings across multiple descriptors -- can be learned and indeed can substantially improve performance (data not shown), but complicates the presentation of the results as it goes beyond simple arithmetic operations on raw data.

Model predictive performance is higher when human validation data has greater inter- and intra-subject agreement (Fig. S13). Increasing rater test-retest reliability and agreement is a necessary precursor to increasing measured model accuracy.

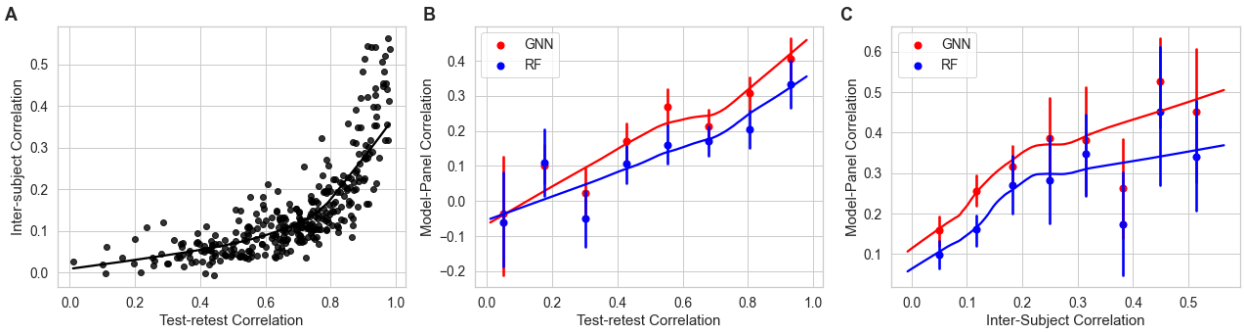

Fig. S13. Relationship between model predictive performance and inter- and intra-subject agreement. **(A)** Inter-subject correlation as a function of intra-subject correlation (test-retest correlation). **(B)** RF and GNN

model-panel correlation as a function of binned test-retest correlation. **(C)** RF and GNN model-panel correlation as a function of binned inter-subject correlation. Model performance is capped by panelists' rating consistency.

Accounting for odorous contaminants

To account for the potential presence of odorous contaminants in the 400 commercial compounds purchased for the human validation study, we developed a gas chromatography-mass spectrometry/olfactometry quality control (QC) procedure. Fifty of the 400 molecules were selected for QC and shipped to the University of Reading for GC-MS/O analysis. By comparing retention indices of recorded odor percepts measured via GC-O to compound identities determined via GC-MS, we were able to identify cases in which contaminants influenced the odor of the material. We classified the molecules into one of 4 verdict categories: 1) Clean - no odorous contaminants found, 2) Mixed - odorous contaminant found but both nominal compound and contaminant contribute to odor, 3) Contaminated - odorous contaminant found, contaminant is the dominant contribution to odor, 4) Inconclusive - the causal odorant was not identified in GC-O, nor was there any detected odor at the expected elution time. This can happen due to thermal or oxidative degradation of the molecule under GC-O conditions, synergistic odorant combinations, or other experimental difficulties. GC-O experimenter notes and classification verdicts for the 50 QC-set molecules are included in Data S1.

In both QC-set cases where a non-sulfur containing molecule was rated sulfurous by the panel, GC-O showed that a sulfur-containing contaminant was the culprit. Additionally, in most QC-set cases where a non-dimethylamino-containing molecule was rated strongly fishy by the panel, GC-O showed that a dimethylamine contaminant was present. On this basis, molecules with an unexpectedly monotonic fishy/garlic/sulfurous profile were excluded from our analysis, including some molecules that had not been confirmed to be contaminated by GC-O analysis. Next, based on anecdotal reports from fragrance chemists that Michael acceptors are aggressive nucleophile scavengers, we excluded Michael acceptors that were reported as garlicky (phosphorous or sulfur impurity), sulfurous (sulfur impurity) or fishy (nitrogen impurity). Acetylene derivatives are also often garlicky due to phosphine (PH3) impurities and we excluded a few molecules fitting this profile. In total, 26 molecules were dropped from the validation set due to confirmed or potential contamination. The rationale for these exclusions are included in table S1. The decision to exclude these molecules had no significant impact on model performance.

GC-O and GC-MS procedures

Extraction of the compound onto the fiber: The 50 compounds destined for gas chromatography-olfactometry (GC-O) and gas chromatography-mass spectrometry (GC-MS) were supplied either diluted in polyethylene glycol or neat, and absorbed onto xxxxx (white polyethylene?) balls (3 mm diameter, purchased from xxxx). For GC-O, approximately 10 balls (or fewer if the compound was very strong, more if it was very weak) were placed in a 20 mL SPME vial and equilibrated in a water bath prior to extraction onto a preconditioned triple phase solid phase microextraction (SPME) fiber (50/30 µm divinylbenzene/carboxen on polydimethylsiloxane (Supelco,

Poole, UK). Generally, the samples were incubated at 45 or 55 °C depending on their volatility for 10 min, and extracted for a further 10 min (details in Table xxx).

Gas Chromatography-Olfactometry (GC-O): After extraction, the SPME device was inserted into the injection port of an HP7890 GC from Agilent Technologies (Santa Clara, CA, USA) coupled to a Series II ODO 2 GC-O system (SGE, Ringwood, Victoria, Australia). The SPME fibre was desorbed in a split/splitless injection port held at 280 °C. The column employed was an Agilent HP-5 MSUi capillary (30 m, 0.25 mm i.d., 1.0 µm df) non-polar column. The temperature gradients was as follows: 40 °C initial temperature with a rise of 8 °C/min up to 200 °C and 15 °C/min from 200 °C to 300 °C and the final temperature held for a further 10 min. Helium was used as carrier gas (2 mL/min). At the end of the column, the flow was split 1:1 between a flame ionisation detector (kept at 250 °C) and a sniffing port using 2 untreated silica-fused capillaries of the same dimensions (1 m, 0.32 mm i.d.).

Odor assessment: Odor assessment was carried out by two flavour experts with 20 years' experience in using the GC-O, who had been familiarised with the standard lexicon. One expert assessed 46, compounds while the second assessed the remaining 4, and confirmed the assessment of a further 24 where clarification was required. Each assessor waited until the solvent had eluted (~5 mins) and sniffed the compounds eluting from the column until 20 min (equivalent to an LRI of 1700). They noted the time, intensity and descriptors for each compound that was detected. Linear retention indices were calculated by comparison with the retention times of C6-C25 n-alkane series analysed on the same day using the same conditions as for sample analyses. Where the LRI matched that of the target compound as determined by GC-MS, this was deemed to be the target compound and any other odors detected were contaminants.

Gas chromatography-mass spectrometry (GC-MS): For identification of the target compound by GC-MS, 2 balls (or more of it was very weak) were placed in a 20 mL SPME vial and equilibrated in a water bath at 30 °C for 10 min prior to a 30 s extraction onto the same SPME fibre type as used for GC-O. For six less volatile samples (133 136 316 728 917), 10 balls were used, the incubation time was increased to 20 min at 55 °C, and extraction time increased to 20 min. A 7890A Gas Chromatograph coupled to a 5975C series GC/MSD from Agilent was used, equipped with the same column as described above. The oven started at 40 °C and increased to 300 °C at a rate of 8 °C/min. Helium was the carrier gas at a flow rate of 0.9 mL/min. Mass spectra were recorded in electron impact mode at an ionization voltage of 70 eV and source temperature of 220 °C. A scan range of m/z 25-450 with a scan time of 0.69 s was employed and the data were controlled and stored by the ChemStation software (Agilent, Santa Clara, CA). Linear retention indices were calculated by comparison with the retention times of C6-C25 n-alkane series analysed on the same day using the same conditions as for sample analyses. Compounds and contaminants were identified by comparison of their mass spectrum with those in the NIST 2020 library and, where available, the LRI was compared to that reported in the online NIST chemistry webbook or PubChem.

Historical explanations of Odor

Historical structure-odor relation models came in the form of empirical rules that are phrased as boolean logic expressions on the presence, absence, or proximity of molecular fragments.For example, Boelens' Rose rule is phrased as "the presence of a 7-9 carbon moiety with a hydroxy or oxy carbonyl or ether group attached to the moiety."(*38*) We also note that the original expression of these

rules are often underspecified, meaning that these plain-language rules cannot be converted directly into e.g. Python code. We show two examples below, including Boelens' 1973 rose rule and Stoll's 1936 musk rule (Fig. S14)(*38–40*).

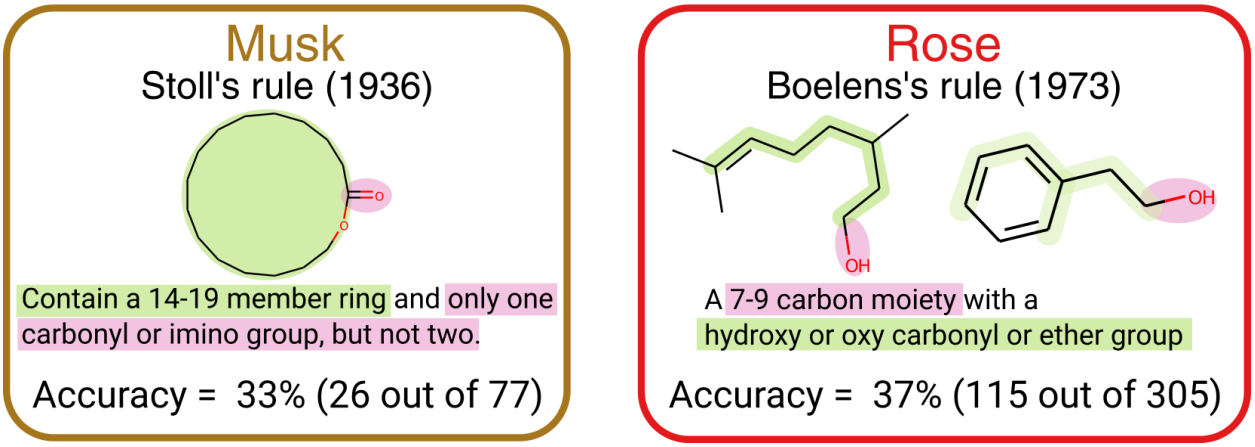

Fig. S14: Example of two historical odor rules with their recall as measured on the Goodscents, Leffingwell datasets.

Supplemental Data

Molecules are indexed by a unique identifier (RedJade Code) allowing for reproduction of most results shown here. Information required to identify chemical structures will be provided upon publication.

Data S1. Metadata for 400 molecules comprising the prospective validation dataset. Columns in the dataset are defined as follows:

RedJade Code	Internal anonymizing tracking number for panelist responses [PRIMARY KEY]
Odor Key	Tracking number from chemical inventory system, ignore.
Category, Kit	Batch number of the molecule. Molecules were tested in 8 waves of 50 molecules.
Solvent	Diluting solvent, if needed for safety or for intensity balancing
Final []	Concentration of molecule (w/w) in final sample
GCO raw commentary	Raw notes from GC-O analyst, if molecule was tested with GC-O
GCO result	Verdict from GC-O analysis
GCO contaminant, if identified	Canonical SMILES of the causal contaminant, if one was successfully identified
Impact on GNN performance	Whether the GCO result had a good, bad, neutral, or unknown effect on GNN's prediction performance.
Disqualification reason	Reason for disqualification. If blank, molecule was retained for analysis
Selection reason	Original selection criteria. Molecules were predicted by the GNN or Random Forest model to have an odor prediction above some threshold despite structural dissimilarity to known instances of that odor class, or to have an odor prediction below some threshold, despite structural similarity to known instances of that odor class.

Data S2. Panelist evaluations of 20 common odorants. Prospective panelists gave RATA

ratings using the 55-word lexicon for 20 common odorants; panelists with a raw test-retest correlation greater than 0.35 were invited to join the panel.

Data S3. Panelist evaluations of 400 novel odorants. Between 15 and 18 panelists rated intensity and pleasantness and gave RATA ratings using the 55-word lexicon for each molecule.

Data S4. Odor attribute predictions on 400 molecules by a random forest model trained on GS/LF datasets.

Data S5. Odor attribute predictions on 400 molecules by a graph neural network model trained on GS/LF datasets. Final layer.

Data S6. Graph neural network embeddings on 400 molecules. Penultimate layer.

Data S7. Correspondence table between internal odorant identifiers and chemical structures.