Corrected 27 October 2023; see below

Supplementary Materials for

A principal odor map unifies diverse tasks in olfactory perception

Brian K. Lee *et al.*

Corresponding authors: Joel D. Mainland, jmainland@monell.org; Alexander B. Wiltschko, alex@osmo.ai

Science **381**, 999 (2023)
DOI: 10.1126/science.ade4401

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Correction: Identifier codes have been added to data S1 to S7 that will allow those who have made the MTA request to link the materials that we send directly to them to the corresponding supplementary data files.

Materials and Methods

Training dataset

The Good Scents (6) 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

Our model architecture is based on Message Passing Neural Networks (*7*). For the message passing layer, we used 2 layers of edge-conditioned matrix multiplication with 63 hidden units and GRU updates, on top of a 45-dimensional atom featurization. In the readout layer, each atom’s embedding is folded into its adjacent bond embeddings into a 152-dimensional embedding, and then summed to generate a molecule embedding. This molecule embedding was transformed through a 4-layer fully connected network, with decreasing layer sizes from 1024 to 256 and a final sigmoid function to make label predictions for each of the 138 curated descriptors from GS-LF dataset listed above.

All references to the GNN “embedding space” or the Principal Odor Map (POM) 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 ~5,000 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 training/test/validation splits (*37*). 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.

The GNN descriptor predictions are the average predictions of an ensemble of 50 GNNs trained with the same hyperparameters but with different random seeds. The POM is extracted from a single one of these models.

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 (*38*) 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 training/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] (*17*).

Model- and label-space visualizations and comparisons

To visualize the odor space constructed by our model, GS-LF odorants are plotted by the first two principal components of their POM coordinates. In main text Fig. 1F, the POM is annotated with perceptual labels from the floral, meaty, and ethereal perceptual classes. Locations of high density for all 138 perceptual labels are visualized by perceptual cluster in Fig. S1.

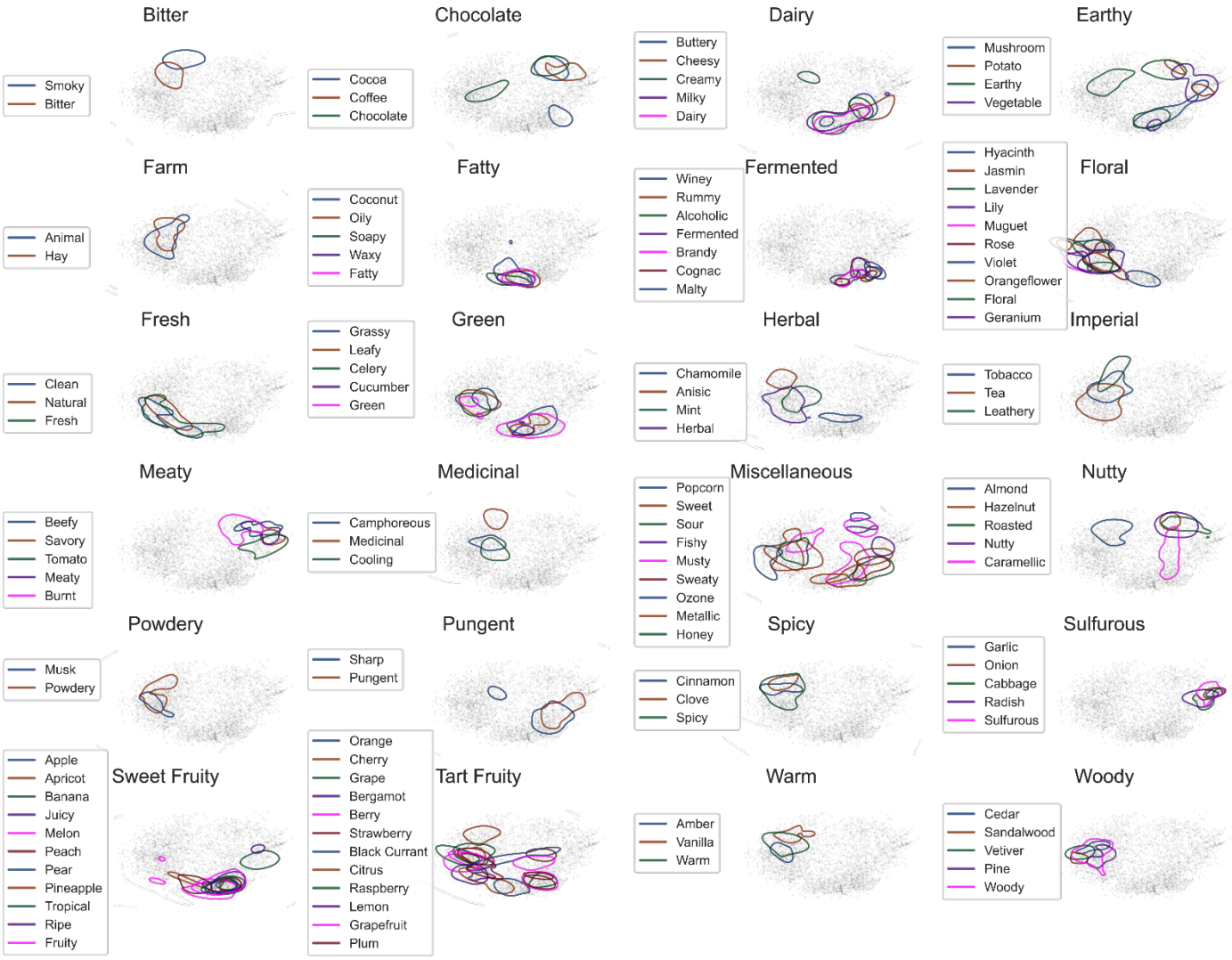


Fig S1. Odorants plotted by the first and second principal components (PC) of their POM coordinates (256 dimensions). Each subplot shows a subset of related odor labels; the category label of each perceptual cluster was subjectively determined. Areas dense with molecules that are described by each label are outlined in the plots.

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. After applying a top-2 principal component reduction to the GNN embedding, Morgan fingerprints, and perceptual label spaces, we found that the POM better represents relative distances: distances in the perceptual map are more significantly correlated to distances in the POM (R=0.73, Fig. S2) than to distances in the FP map (R= -0.12, $p < 0.001$, Fig. S2). Inter-cluster distances are computed as the mean Euclidean distance of all |Cluster 1| * |Cluster 2| pairwise molecule distances. Both panels in Fig. S2 show all (138^2 - 138) pairwise comparisons of odor clusters.

The POM also 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. S3), where smaller CD values denote more dense clusters.

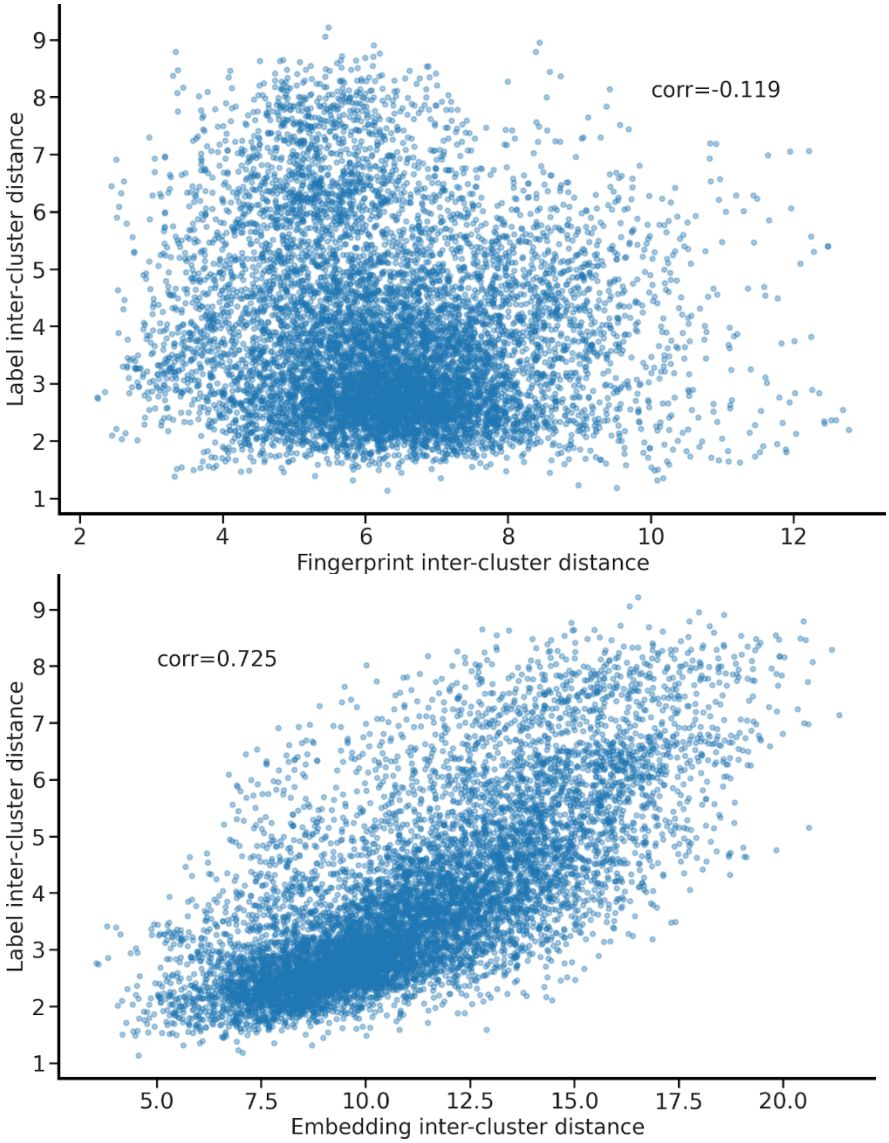


Fig. S2. Inter-cluster distance correlation. Each point represents a pair of odor labels (e.g. [fruity, sweet]). All label distances are calculated as the mean distance of all pairwise molecules. If there are 1500 fruity molecules and 2000 sweet molecules, the [fruity sweet] point represents the mean of 3,000,000 pairwise distances (**top**). Cluster fingerprint distance does not correlate strongly with cluster label distance, whereas (**bottom**) cluster embedding distance correlates well with cluster label distance.

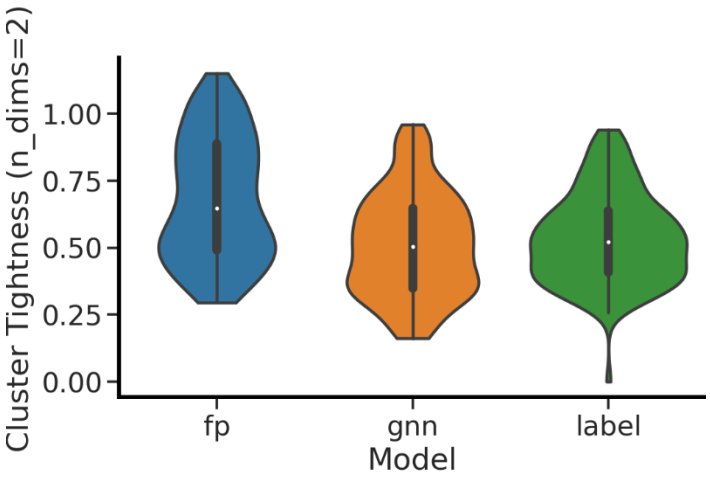


Fig. S3. Distribution of cluster tightness across 138 odor classes. Each violin plot represents a histogram over 138 odor class tightness metrics. Cluster tightness for an odor class is defined as the ratio of {mean Euclidean distance of all in-group pairwise distances} to {mean Euclidean distance of 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 a limited set of 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 at least one aroma reference to facilitate rater training. Woody, floral, and spicy each had a second aroma reference.

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 (#843955), 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. S4) 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 S4. Initial odorant reference/training kit, containing 58 distinct 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). Some odor labels were not precisely captured by a single reference and have multiple references.

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 (*39*); 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 (Gradient, Boston, MA), and approved by the University of Pennsylvania IRB), and removed

likely odorless molecules according to water-solubility (cLogP < 0) and volatility (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. S5). Upon receipt of purchased molecules, we manually removed 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.

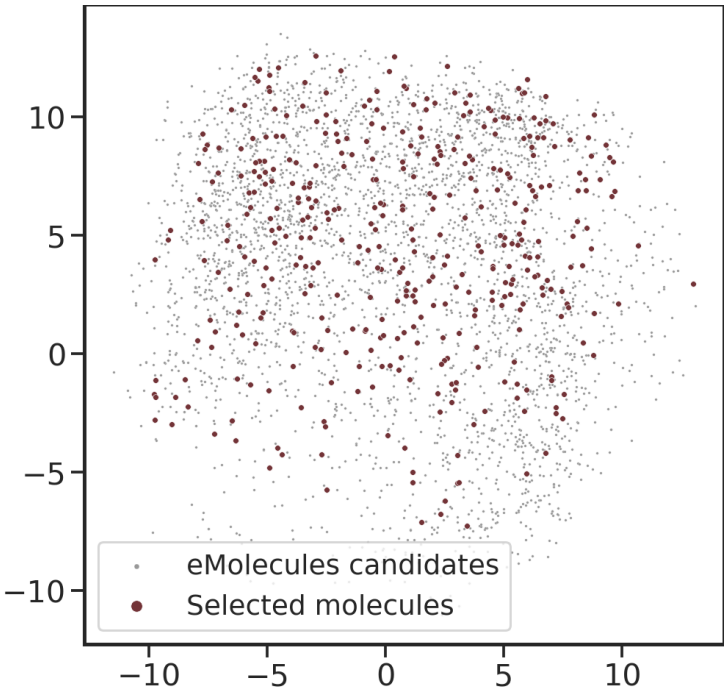


Fig. S5. Candidate and selected molecules from the eMolecules catalog displayed in Morgan fingerprint PCA space. Selected molecules span the space of potential candidate molecules from eMolecules.

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). Odorants were blinded with random 3-digit codes, and the order in which subjects evaluated the odorants was determined following a Williams Latin Square design. 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 $n \geq 15$ 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. S6A, and the distribution of the number of descriptors applied to each molecule is shown in Fig. S6B. Each molecule is typically assigned between 1-6 descriptor ratings by each rater. Most descriptors are used at least once by every rater. Fig. S6C 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. S7). 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. S8). 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.

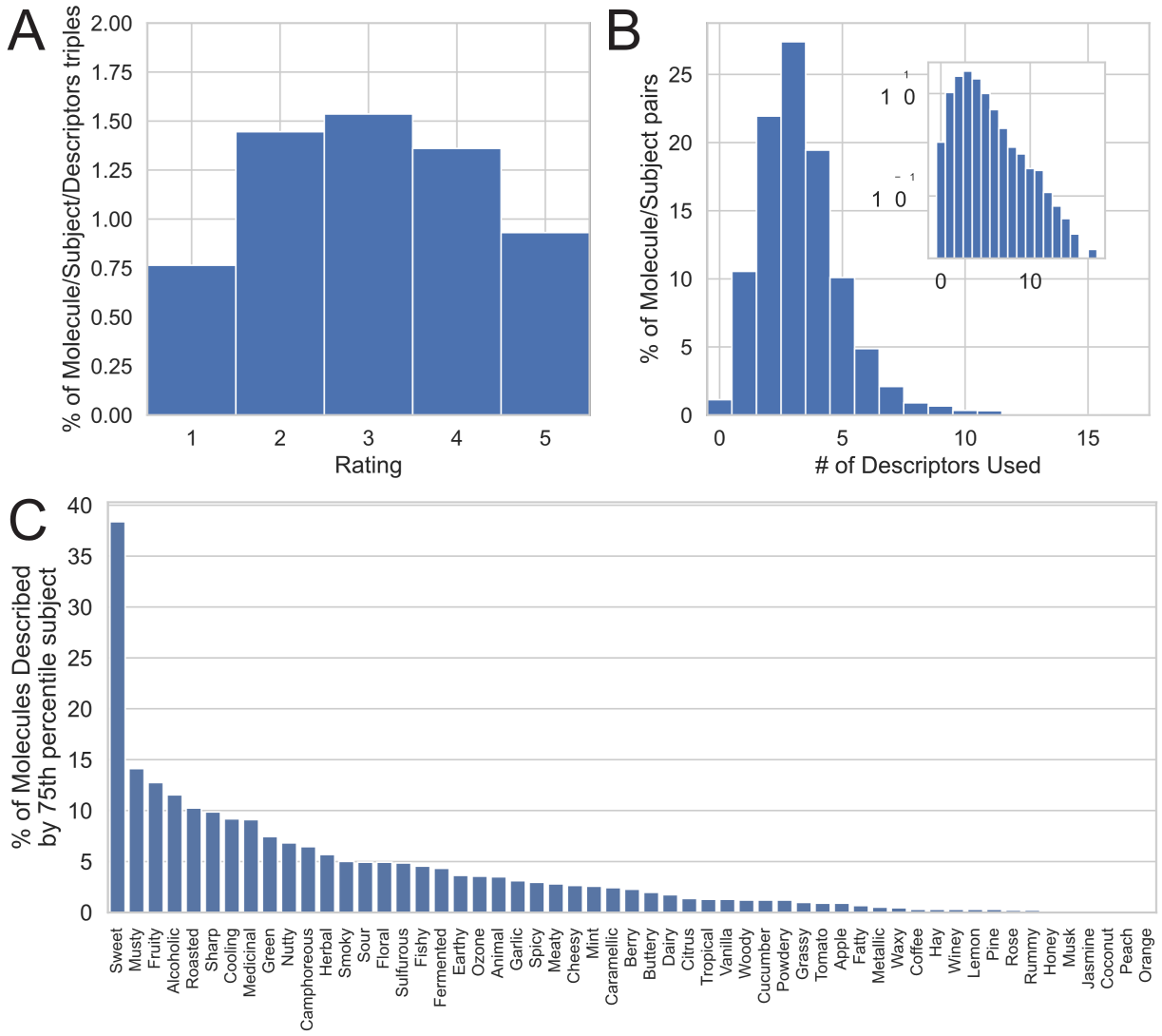


Fig. S6. Human psychophysics prospective validation set summary statistics. (A) Distribution of non-zero descriptor ratings. 95% of all ratings are 0; the remaining 5% of ratings are roughly evenly distributed between ratings of 1-5. (B) Distribution of the number of descriptors applied to each molecule. On average, each molecule-rater encounter generated a set of 3-4 nonzero descriptors (inset shows the same distribution with a logarithmic y-axis). (C) Percent of molecules described by each of the 55 odor descriptors according to the 75th percentile panelist's ratings. In other words, for 38 molecules, at least 25% of raters ascribed a sweet label.

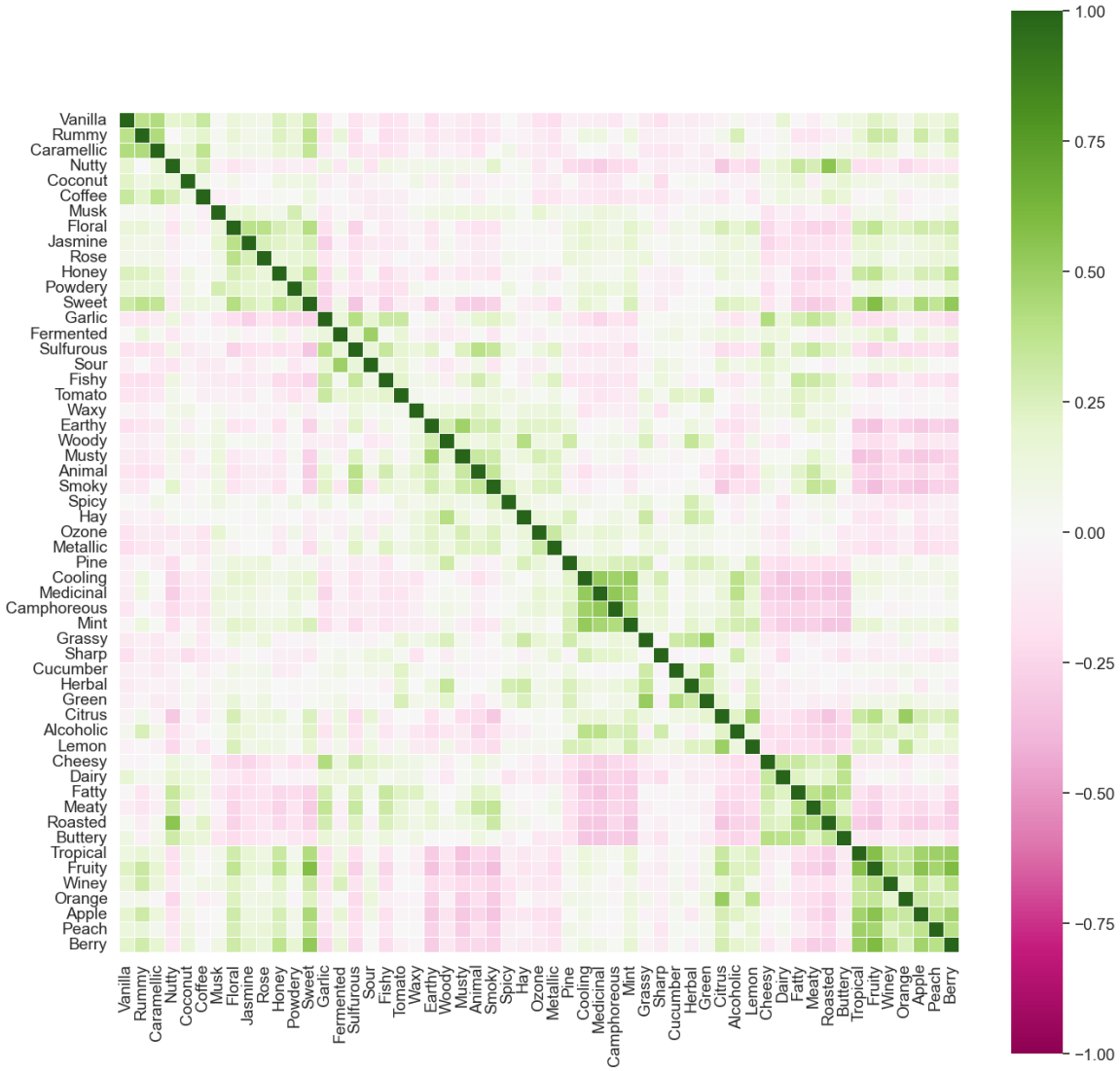


Fig S7. Correlation matrix of panelist ratings for the 55 odor lexicon descriptors. Descriptors with strong positive correlations are dark green; descriptors with strong negative correlations are in dark pink. Odor descriptors show a clear correlation structure, when ordered by Spectral co-clustering algorithm (*40*).

	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 S8. 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. See supporting code (Cluster correlations.py) for precise structural cluster SMARTS queries, and for odor cluster definitions, which are near-exact copies of the clusters derived in figure S7.

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. S9). Compared to a prior large-scale human psychophysical study (*6*), our collected dataset has more descriptors and higher panel mean test-retest reliability (Fig. S10 and Fig. S11), 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. S12). 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.

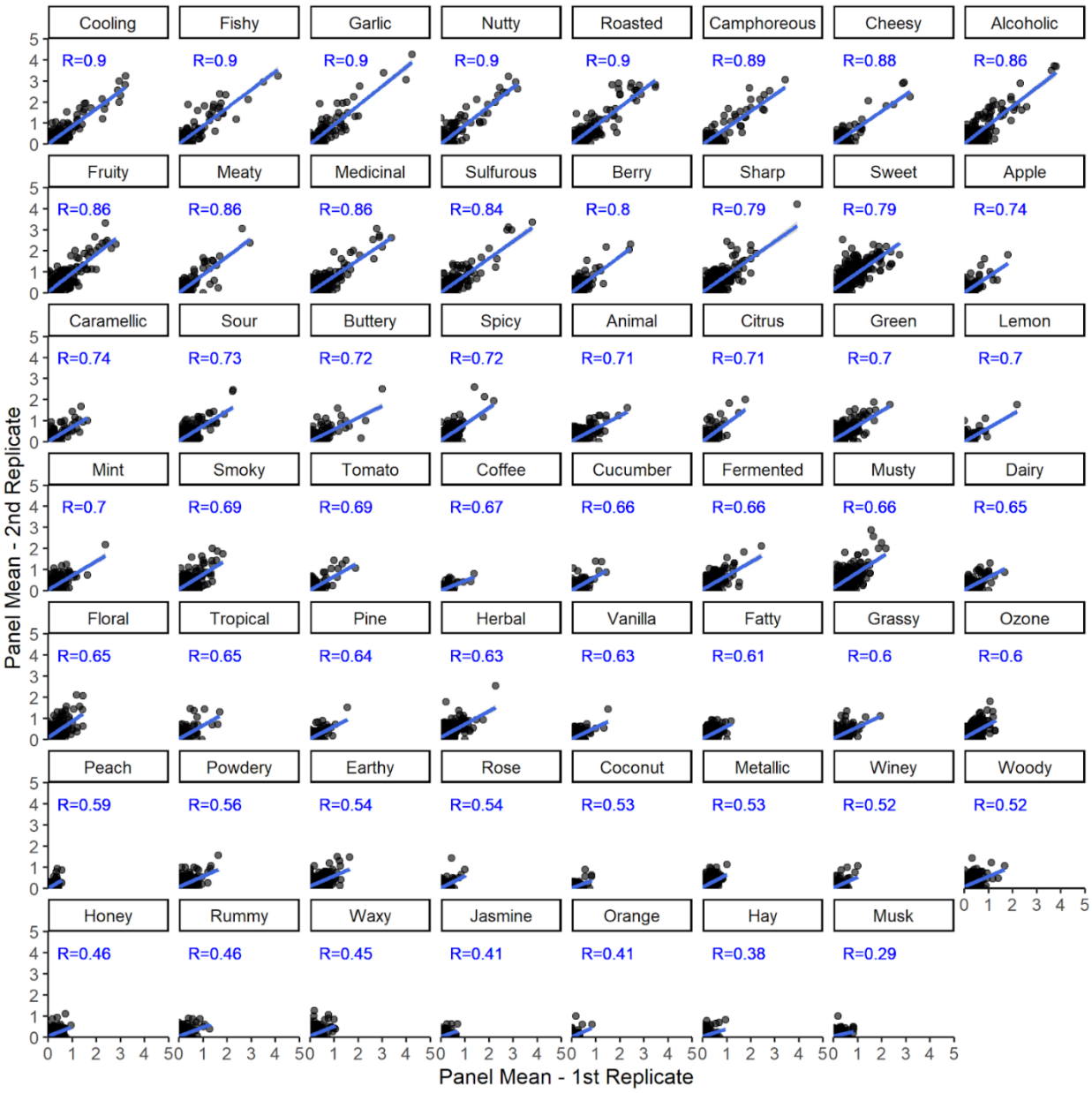


Fig. S9. Panel mean (n≥15 subjects) test-retest correlation (R) for the 55 descriptors in the lexicon applied to the 400 novel odorants in the prospective validation set. Each dot represents 1 molecule. Descriptors are ordered by descending correlation.

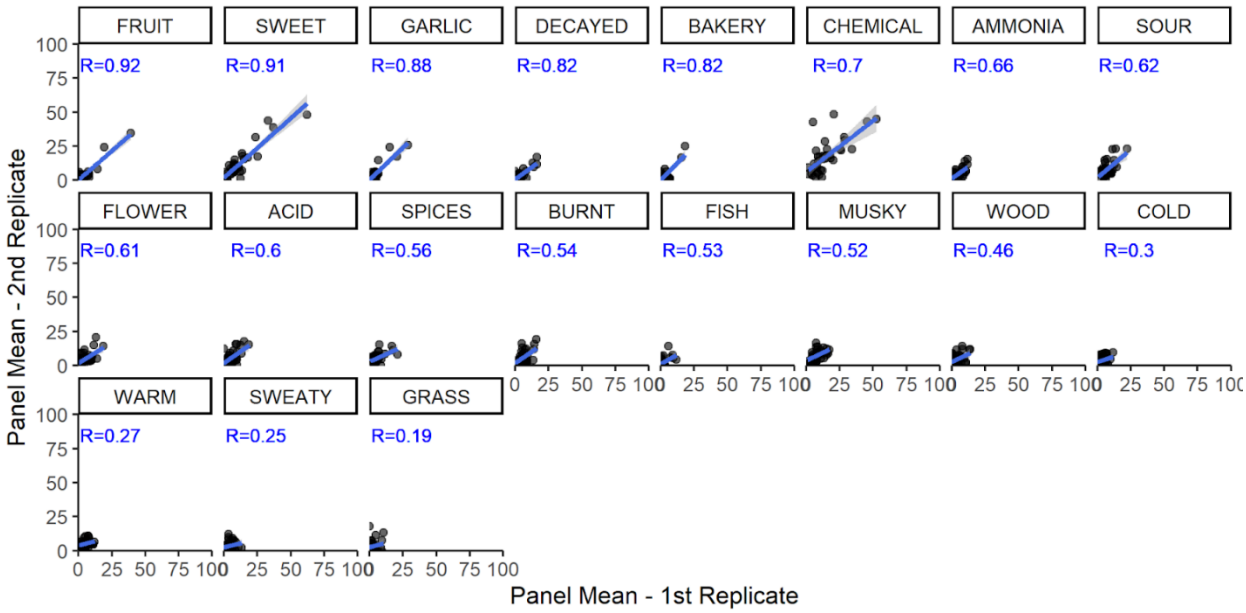


Fig. S10. Panel mean (n=49) test-retest correlation (R) for the 19 descriptors in the DREAM olfaction challenge dataset (6). Each dot represents 1 molecule. Descriptors are ordered by descending correlation.

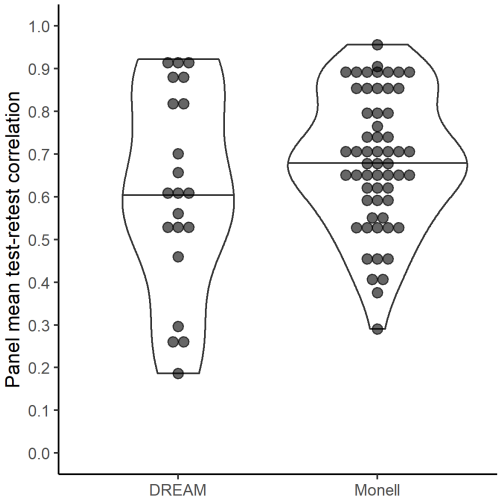


Fig. S11. Panel mean test-retest correlation for the 19 descriptors in the DREAM olfaction challenge (left) (*6*) and for the 55 descriptors in the present study (right). Each dot represents test-retest correlation one odor descriptor.

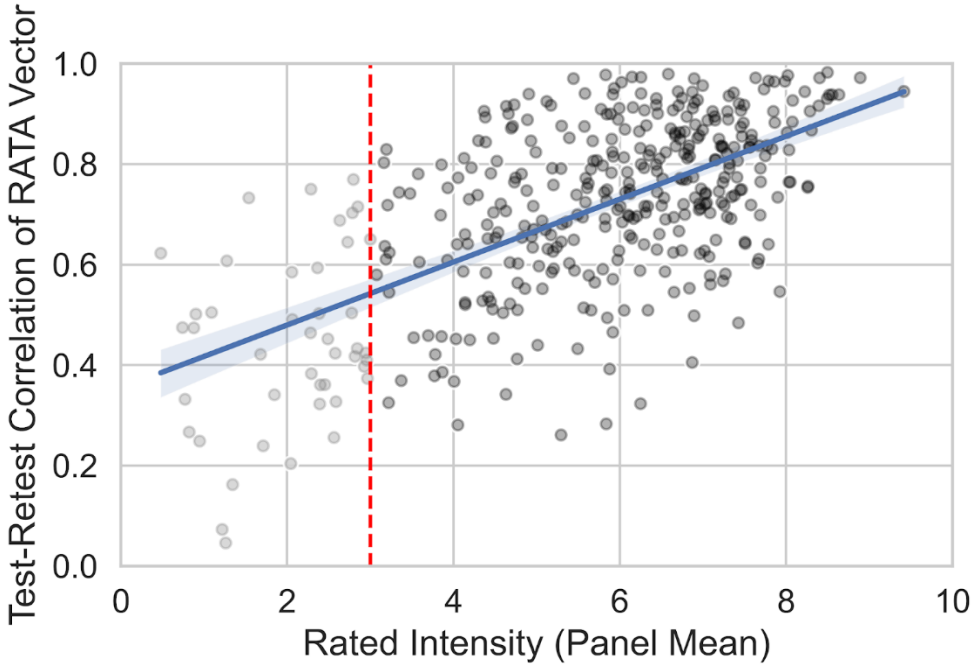


Fig S12. Test-retest correlation for 400 novel odorants as a function of panel mean intensity rating for that odorant. Molecules with lower rated intensity have weaker test-retest correlation of the panel mean, and a threshold of 3 was used to exclude molecules with low intensity from the study findings.

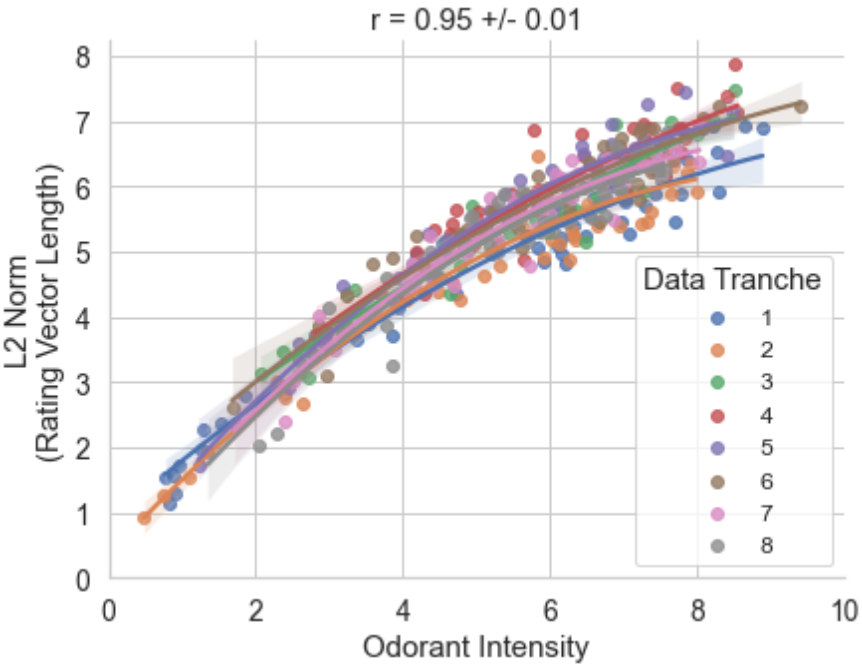
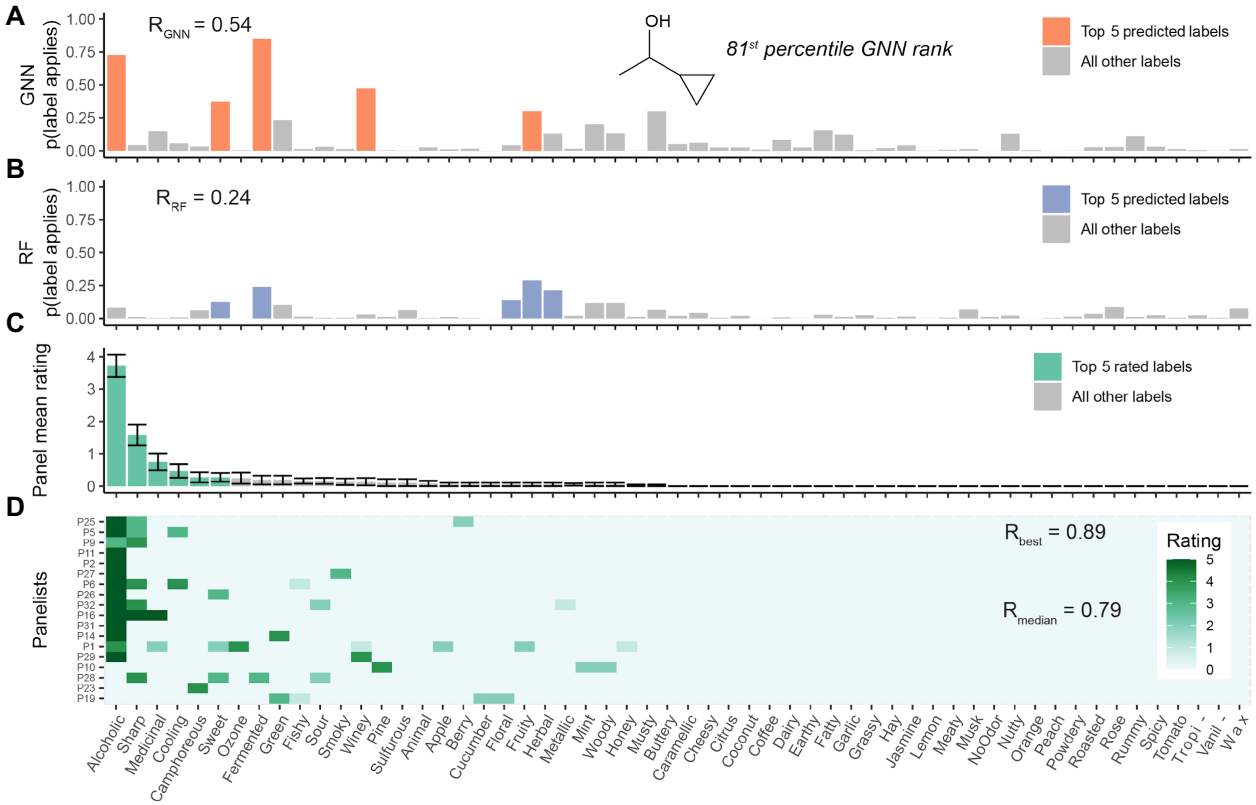
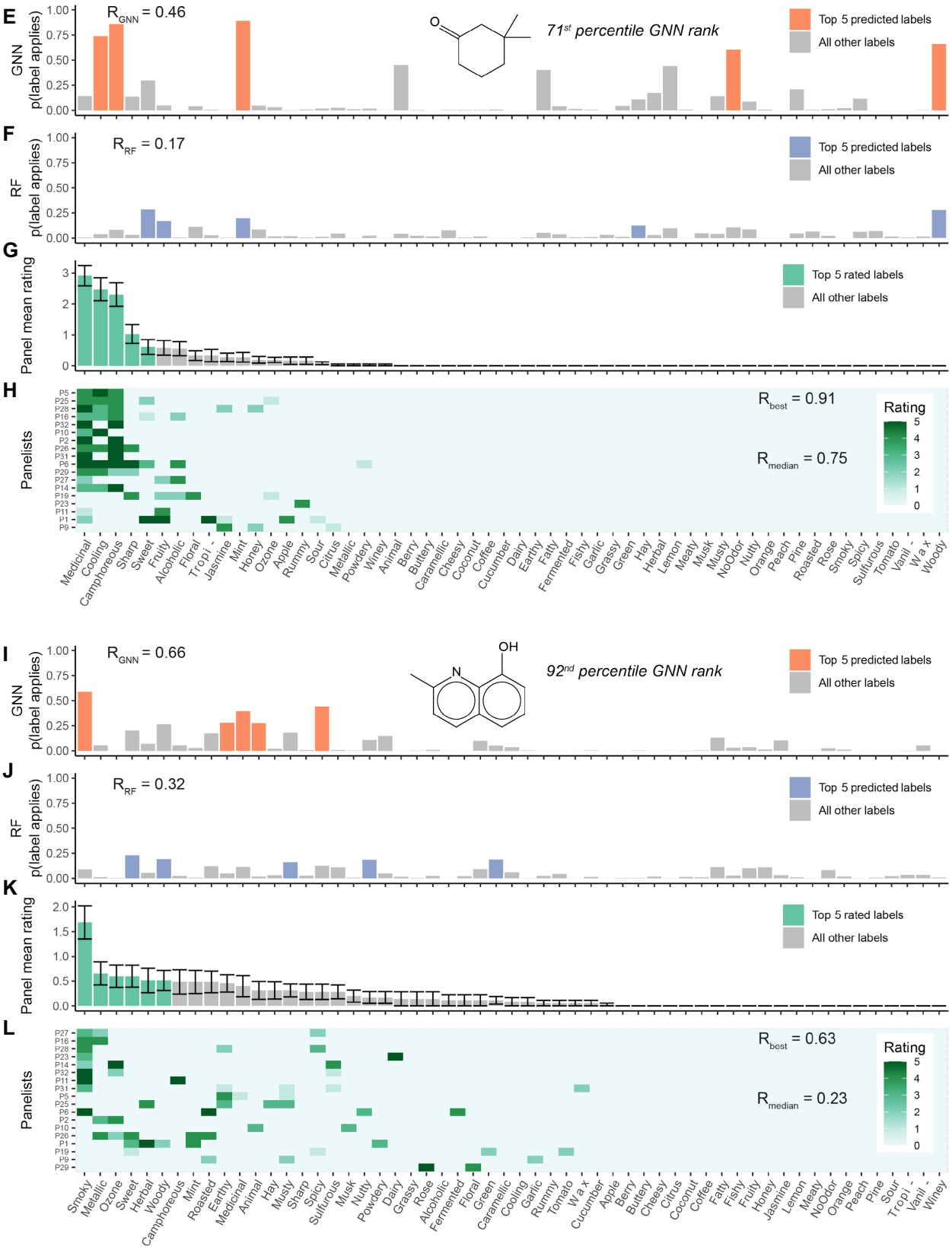


Fig. S13. L2 norm length of the mean RATA rating vector, plotted against mean intensity. Human psychophysical ratings were gathered in 8 data collection waves; differently colored dots and fits come from different tranches of data collection. Panelists use more descriptors and give higher RATA ratings for higher-intensity stimuli.

Evaluating Model Performance on Prospective Validation Set

The main text presents raw panelist ratings, mean panel ratings, and GNN and RF model predictions for one molecule (Fig. 2A-D); four additional examples are presented here (Fig. S14) to show a broader range of panel rating and model prediction patterns. The following 4 molecules were all subjected to the quality control procedure described in the next section and determined to be free from odorous contaminants. This analysis and visualization can be generated for any molecule using code provided at https://github.com/osmoai/publications.




E
GNN p(label applies)
R_GNN = 0.46
71st percentile GNN rank
Top 5 predicted labels
All other labels
F
RF p(label applies)
R_RF = 0.17
Top 5 predicted labels
All other labels
G
Panel mean rating
Top 5 rated labels
All other labels
H
Panelists
R_best = 0.91
R_median = 0.75
Rating
Medicinal Cooling Camphoreous Sharp Sweet Fruity Alcoholic Floral Tropi_ Jasmine Mint Honey Ozone Apple Rummy Sour Citrus Metallic Powdery Winey Animal Berry Buttery Caramellic Cheesy Coconut Coffee Cucumber Dairy Earthy Fatty Fermented Fishy Garlic Grassy Green Hay Herbal Lemon Meaty Musk Musty NoOdor Nutty Orange Peach Pine Roasted Rose Smoky Spicy Sulfurous Tomato Vanil_ Wax Woody
I
GNN p(label applies)
R_GNN = 0.66
92nd percentile GNN rank
Top 5 predicted labels
All other labels
J
RF p(label applies)
R_RF = 0.32
Top 5 predicted labels
All other labels
K
Panel mean rating
Top 5 rated labels
All other labels
L
Panelists
R_best = 0.63
R_median = 0.23
Rating
Smoky Metallic Ozone Sweet Herbal Woody Camphoreous Mint Roasted Earthy Medicinal Animal Hay Musty Sharp Spicy Sulfurous Musk Nutty Powdery Dairy Grassy Rose Alcoholic Fermented Floral Green Caramellic Cooling Garlic Rummy Tomato Wax Cucumber Apple Berry Buttery Cheesy Citrus Coconut Coffee Fatty Fishy Fruity Honey Jasmine Lemon Meaty NoOdor Orange Peach Pine Sour Tropi_ Vanil_ Winey

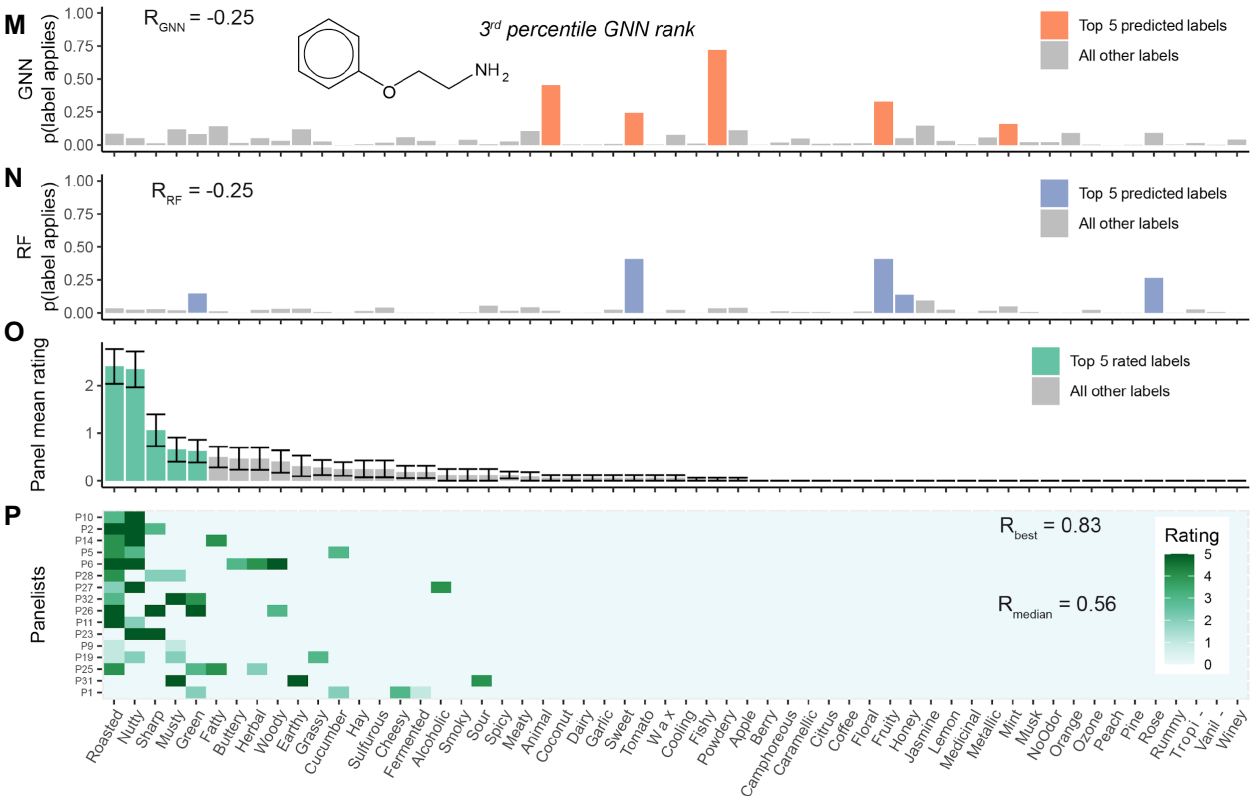


Fig. S14. For the molecules (A-D) 1-cyclopropylethanol, (E-H) 3,3-dimethylcyclohexan-1-one, (I-L) 2-Methylquinolin-8-ol, (M-P) 2-Phenoxyethylamine; (A, E, I, M) GNN model label predictions, (B, F, J, N) random forest (RF) model label predictions, (C, G, K, O) panel mean ratings with standard error bars, and (D, H, L, P) individual panelist ratings (where 0 means the label does not apply and 5 means the label is very applicable), averaged over 2 replicates. The top 5 ranked descriptors for each molecule are in orange (GNN), purple (RF), or green (panel). Descriptors in panels are ordered by panel mean ratings for the given molecule. Figure panels are annotated with the Pearson correlation coefficient of our model predictions to the panel mean rating. Panels D, H, L, and P include panelist-panel correlation coefficients for the panelist that best matches the panel mean and for the panelist with the median match.

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 our model and panelists were not directly comparable. Our 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, our model's predictions had a positive nonzero score, indicating a systematic scoring bias in our 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. S15). Increasing rater test-retest reliability and agreement is a necessary precursor to increasing measured model accuracy.

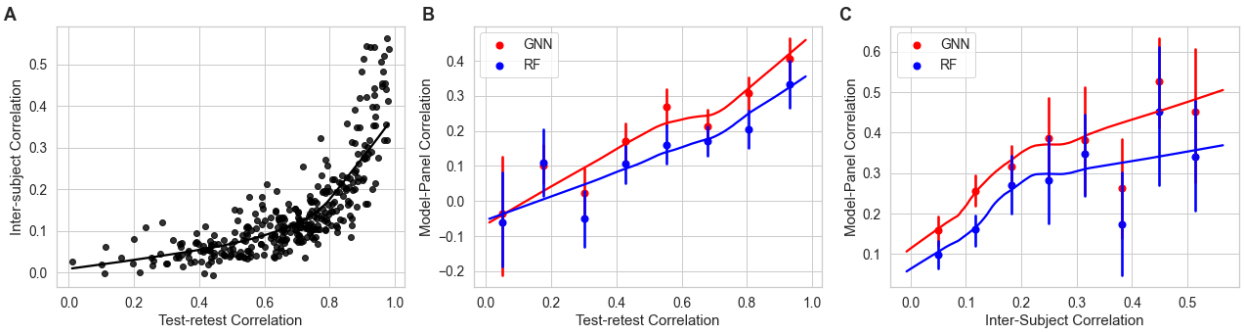


Fig. S15. Relationship between model predictive performance and inter- and intra-subject agreement. (A) Inter-subject correlation is plotted against intra-subject correlation (test-retest correlation). Each dot represents 1 molecule. (B) RF and GNN model-panel correlation plotted against binned test-retest correlation. (C) RF and GNN model-panel correlation plotted against binned inter-subject correlation. Model performance is capped by panelists' rating consistency.

Model successes and failures in predicting labels for each molecule are visualized in Fig. S16. Model performance is further benchmarked against the panelists per odorant (Fig. S17 left), and the fraction of odorants for which our model comes within 1 standard deviation of the panel mean is quantified per label (Fig. S17 right).

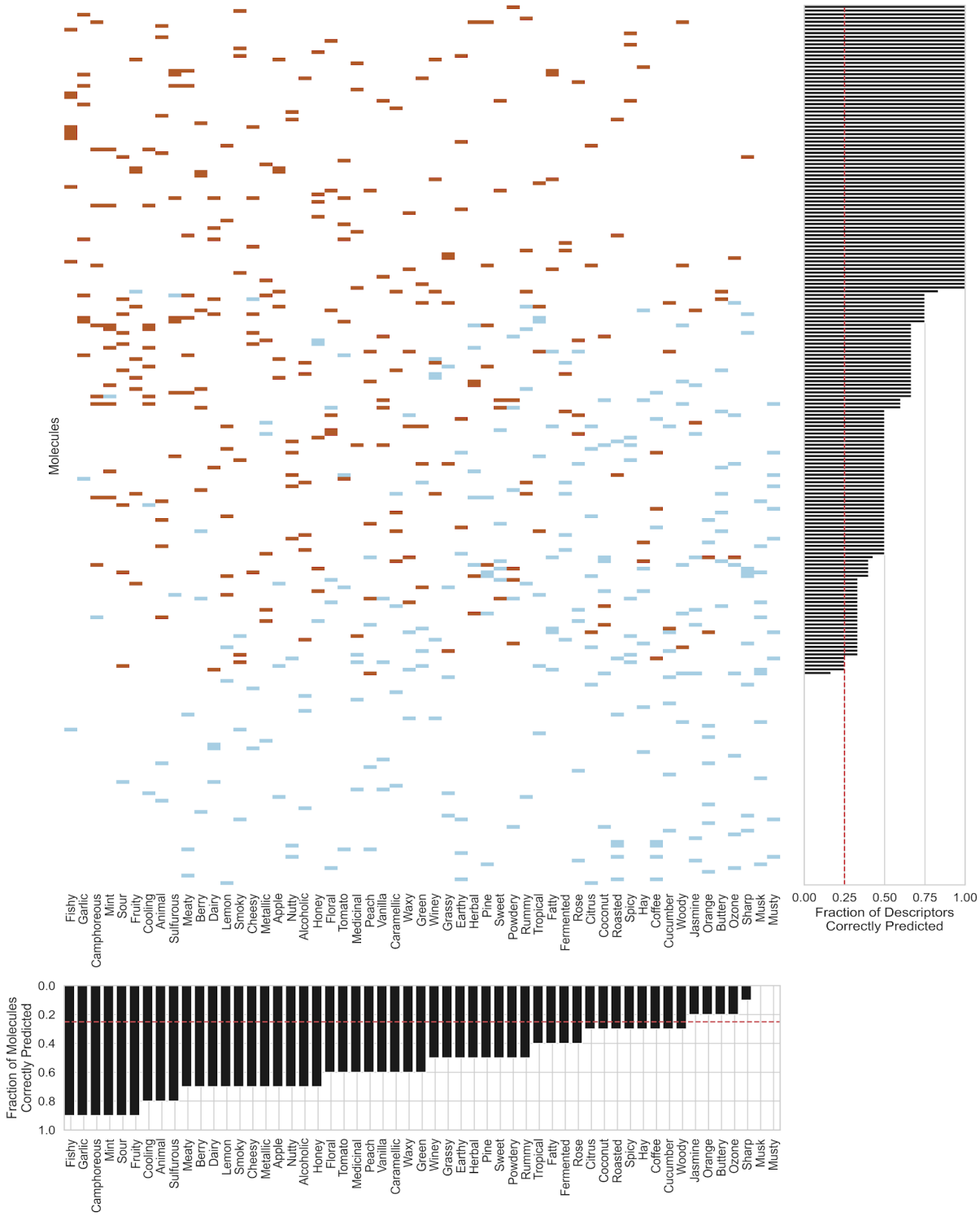


Fig. S16. Successes and failures in model prediction per molecule and descriptor. Central heatmap shows each individual case colored red (success) or blue (failure). Successes are cases where our model predicted that the molecule would have one of the top 10 ratings (across molecules) for that descriptor, and the panel rating was in the top 25% for that descriptor (brown). Failures are the same but when the panel ratings were not in the top 25% (blue). Cases where our model did not predict that the molecule would be in the top 10 for that descriptor are ignored (white). Each row corresponds to all descriptors for one molecule. Each column corresponds to all molecules for one descriptor. Marginal plots show the proportion of successes in each row and column. The red dashed lines indicate chance value.

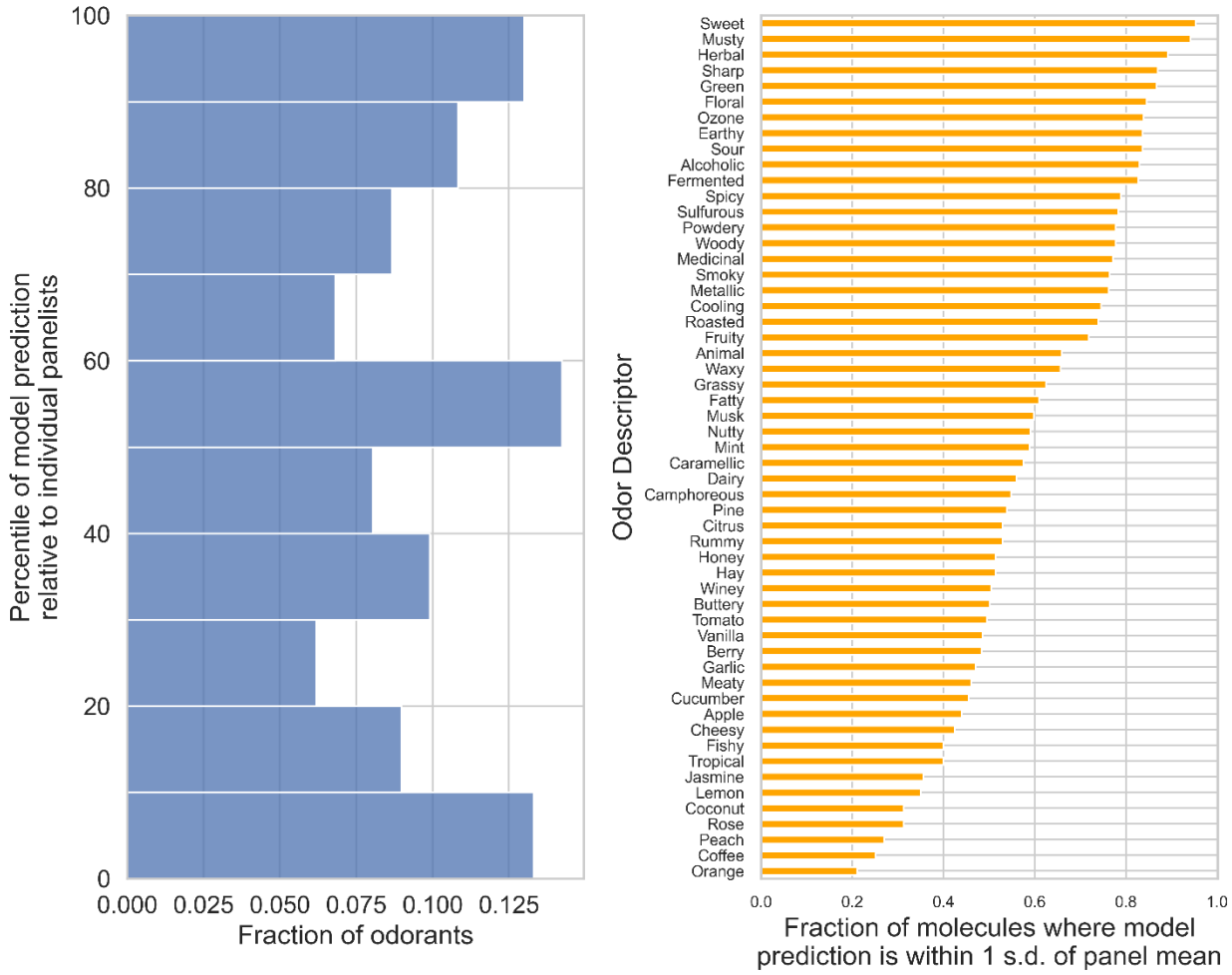


Fig. S17. Model performance relative to panelists and proximity to panel mean. Left: Model performance relative to individual panelists across odorants. The vertical axis indicates the percentile of our model (higher is better). 50 indicates that our model matched the panel mean odor profile as well as did the median panelist. Right: Model performance relative to individual panelists across descriptors. The x-axis shows the fraction of molecules where our model prediction was within 1 standard deviation (across panelists) of the panel mean. The y-axis is sorted from best to worst-predicted descriptors according to this fraction.

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 (GC-MS/O) 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, 19 molecules were dropped from the validation set due to confirmed contamination (Fig. S18) and an additional 13 were dropped due to 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.

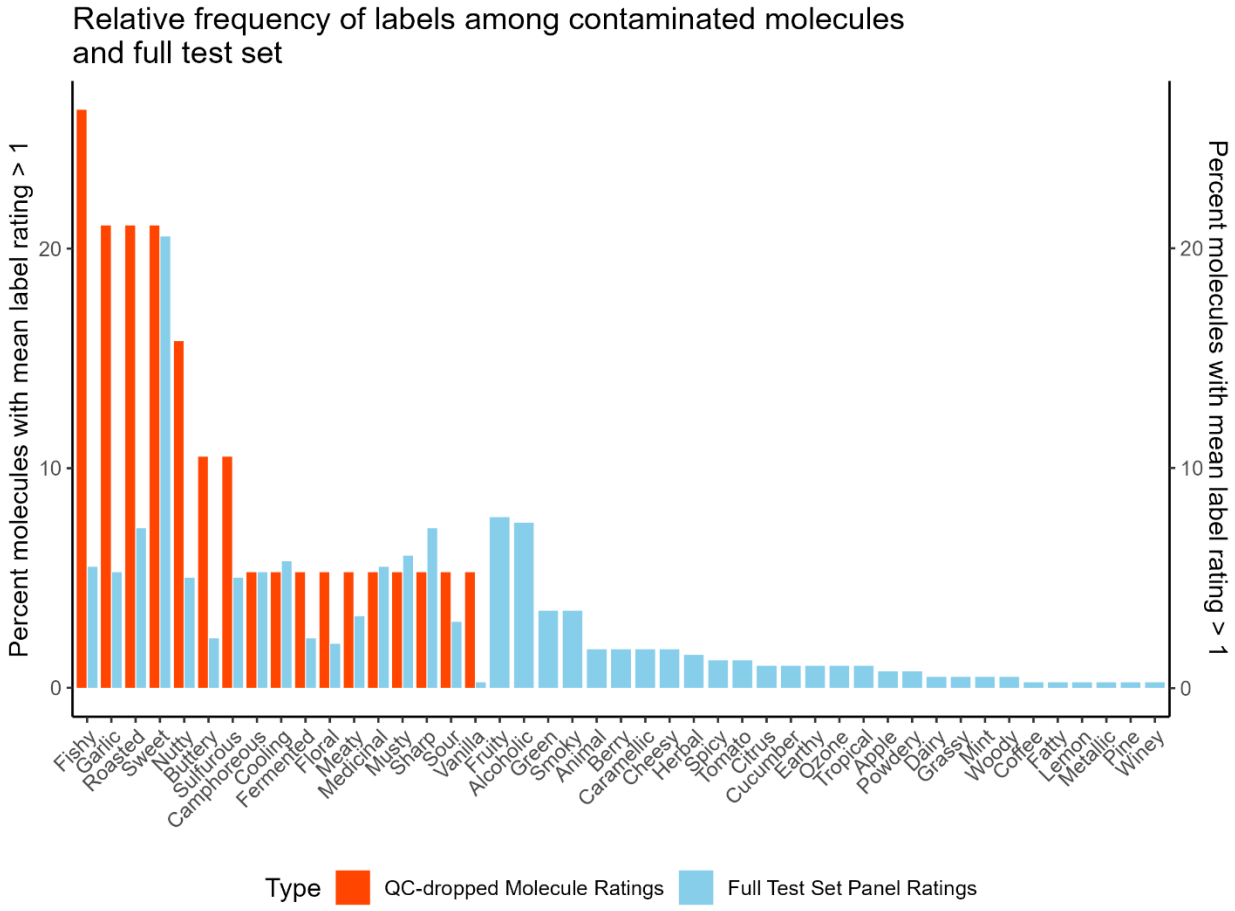


Fig. S18. Label frequency among contaminated molecules and full test set. Plot shows percent of molecules dropped due to odorous contamination ("QC-dropped", n = 19) with a given label (mean panel rating > 1) in orange and percent of molecules in the full test set (n = 399) with a given label (mean panel rating > 1) in blue. Odor labels such as "fishy" and "garlic" are much more common among contaminated molecules than in the full test set.

We estimate the final contamination rate in the full dataset at 31.5% with 95% CI of [27.4%, 35.6%] (Fig. S19). This estimate is accomplished by subgroup analysis, splitting the 50 molecules verified by GC-O into Fishy/Garlic/Sulfurous (FGS) and non-FGS groups (criteria: sum of the raw mean panelist Fishy/Garlic/Sulfurous ratings > 1 on a scale of 0-5). 10/12 of FGS molecules were found to be contaminated, whereas only 9/38 of non-FGS molecules were contaminated. Applying these statistics to

the 347 untested molecules (40 FGS, 307 non-FGS), we arrive at a 31.5% estimate for contamination rate across the full 397 molecules. PDFs for all four subgroups are obtained via the beta distribution and convolved to obtain the final CI.

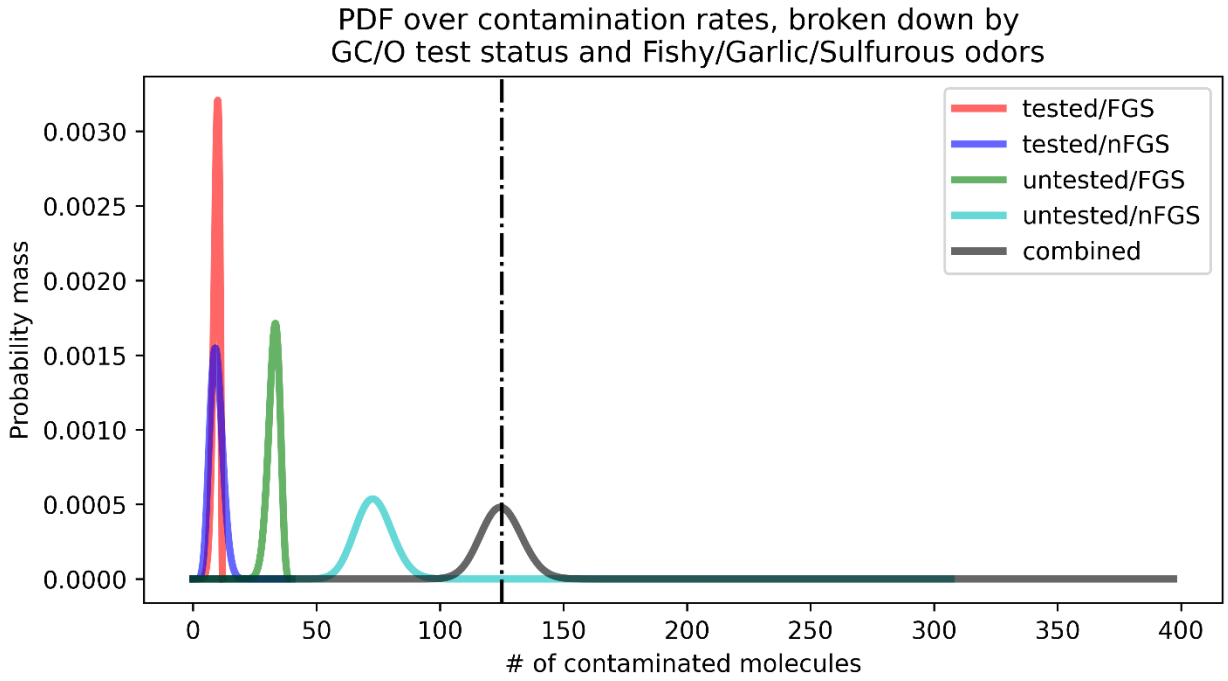


Fig. S19. Probability distribution functions for odorous contamination rates. Contamination rates of Fishy/Garlic/Sulfurous (FGS) contaminants and non-FGS (nFGS) contaminants from the GC/O QC set were applied to the full test set to yield untested FGS, untested nFGS, and combined contamination estimates.

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 Viscopearls (4 mm diameter, Rengo Co., Ltd., Osaka, Japan). For GC-O, approximately 10 Viscopearls (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 Data S1).

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 were 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 ionization 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 familiarized with the standard lexicon. One expert assessed 46, compounds whereas 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 analyzed 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 Viscopearls (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 fiber type as used for GC-O. For six less volatile samples (RedJade codes: 133, 136, 316, 728, 917), we used 10 Viscopearls, increased the incubation time to 20 min at 55 °C, and increased the extraction time to 20 min. We used a 7890A Gas Chromatograph coupled to a 5975C series GC/MSD from Agilent, 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 analyzed 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.

Retrospective Triplet Performance

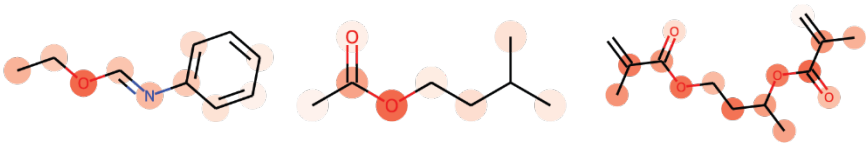
Sell's triplets (*3*) are sets of three molecules each for which there is an anchor molecule A, an "arm" with structural cliff molecule S, and an "arm" with perceptual cliff molecule P. S is chosen to have a very similar perceptual character to the anchor but dissimilar molecular structure. P is chosen to have a very similar molecular structure to the anchor but dissimilar perceptual character. For our analysis we searched for triplets using the training set to find anchor molecules, and molecules from the prospective validation set for structural and perceptual cliff molecules. We conceived this analysis after the prospective validation experiments were completed, so we selected the triplet using empirical labels (binary, for the training set, and binarized, for the prospective validation set).

We selected triplets by choosing molecule A at random from the training set, and then identifying a molecule S that is similar in perceptual character (jaccard distance < 0.6 between A and S using empirical labels) and dissimilar in molecular structure (tanimoto distance > 0.6 between A and S using fingerprints); and a molecule P that is the reverse (jaccard distance > 0.6 between A and P using empirical labels; tanimoto distance < 0.6 between A and P using fingerprints). We then computed for each triplet a structural contrast $(f_s(A, S) - f_s(A, P))^2$ and a perceptual contrast $(f_p(A, S) - f_p(A, P))^2$ where f_s is the tanimoto distance and f_p is the jaccard distance on labels. We ranked all such candidate triplets according to a "triplet score" which we set equal to the sum of the structural and perceptual contrasts. We then selected the top 42 (containing 61 unique "arms") for empirical evaluation. We then collected an independent dataset measuring explicit perceptual distances for all arms, i.e. between each A and S and each A and P. These distances were explicit perceptual judgements of similarity for each pair, rather than perceptual labels applied to each molecule independently, and thus represent a better "ground truth" about whether a set of three molecules meets the definition of a triplet. ~90% of the triplets selected using empirical labels satisfied the definition of a triplet using these explicit perceptual distance judgements.

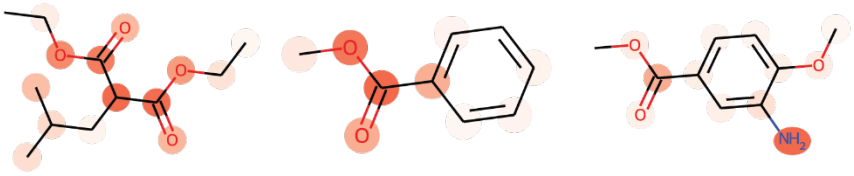
Thus, there are empirically observable triplets, and both label-distance and explicit pairwise similarity judgements are in close agreement.

A triplet represents an "adversarial attack" on our predictive model (*41*). In other words, it is an empirical example of an input designed specifically to flummox a perceptual model trained on molecular structure, leading it to give the wrong answer. A naive model should expect the pair of molecules that have the more similar molecular structure to have a more similar odor percept. However, some models are more robust to such attacks than others (*42*), i.e. they are more likely to give the correct answer rather than the naive model. Using the empirical triplets identified above, we asked whether the GNN model was robust to such attacks, and whether it was more robust than a baseline model trained on structural fingerprints. We found that the GNN model was robust to 50 % of such attacks, vs only 19% for the baseline model ($p<0.01$). Since the explicit perceptual judgments showed that only 90% of the triplets are consistent across measurement approaches, this means that the GNN model was "fooled" less than half of the time in this maximally challenging test scenario, vs 7/8th of the time for the baseline model.

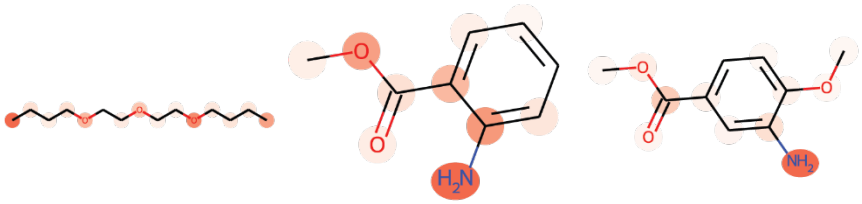
Model attributions are performed to understand how different atoms are contributed to the GNN predictions (Fig. 4A, S20). We used a perturbation-based model attribution method (*43*) where the importance of a specific atom is determined by how much noise on the related features would influence the label predictions. Specifically, we inject a uniform noise between -1 and +1 to the original atom features and connected bond features, and measure the cosine distance between the new model prediction and the original prediction. The atom importance is then measured by averaging the cosine distances across 100 different noise injection. Finally, the atom importance values are normalized across a molecule between 0 and 1. Attribution heatmaps for all 41 tested triplets are available at https://github.com/osmoai/publications.



	#1 struct_cliff	#1 anchor	#1 label_cliff
Triplet Role:			
Compound Name:	Ethyl N-phenylformimidate	isoamyl acetate	1,3-Butanediol Dimethacrylate
Model-predicted Labels:	green, sweet, fruity, floral, rose	sharp, green, sweet, apple, fruity, berry, rummy	herbal, sweet, fruity, floral, rose
Panel Mean Ratings:	Alcoholic 2 Sweet 1.3 Fruity 1.2	N/A	Tomato 1.1 Ozone 0.5 Earthy 0.3



	#2 struct_cliff	#2 anchor	#2 label_cliff
Triplet Role:			
Compound Name:	Diethyl isobutylmalonate	methyl benzoate	Methyl 3-amino-4-methoxybenzoate
Model-predicted Labels:	fatty, apple, fruity, winey	sweet, fruity, honey, floral, rose	sweet, fruity, berry, floral
Panel Mean Ratings:	Fruity 2.3 Berry 2.2 Sweet 1.7	N/A	Musty 2.1 Earthy 0.9 Ozone 0.9



	#3 struct_cliff	#3 anchor	#3 label_cliff
Triplet Role:			
Compound Name:	2,2-Dibutoxyethyl ether	methyl anthranilate	Methyl 3-amino-4-methoxybenzoate
Model-predicted Labels:	fermented, alcoholic, green	orange, sweet, fruity, floral	sweet, fruity, berry, floral
Panel Mean Ratings:	Fruity 0.8 Sweet 0.6 Floral 0.4	N/A	Musty 2.1 Earthy 0.9 Ozone 0.9

Fig. S20. Label frequency among contaminated molecules and full test set. Plot shows percent of molecules dropped due to odorous contamination ("QC-dropped", n = 19) with a given label (mean panel rating > 1) in orange and percent of molecules in the full test set (n = 399) with a given label (mean panel rating > 1) in blue. Odor labels such as "fishy" and "garlic" are much more common among contaminated molecules than in the full test set.

Supplemental Text

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" (*44*). 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. S21) (*44*, *45)*.

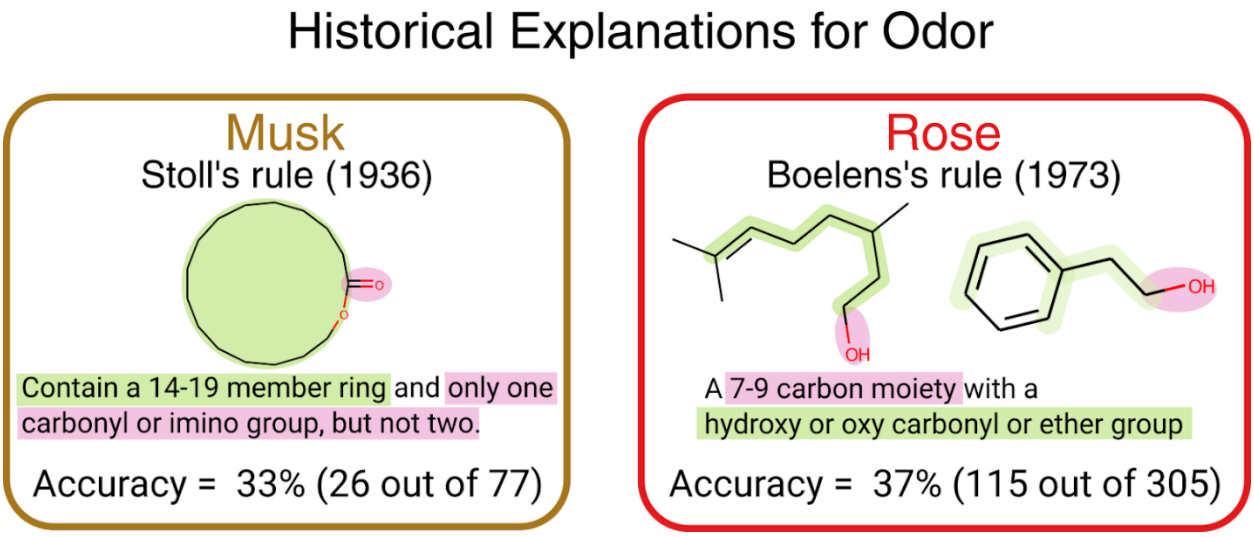


Fig. S21: Example of two historical odor rules with their recall as measured on the Goodscents, Leffingwell datasets.

Supplemental References

37. K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 6913 LNAI, 145–158 (2011).

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. M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey. Pattern Recognit Lett. 150, 228–234 (2021).

44. 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).

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

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 at https://github.com/osmoai/publications.

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

RedJade Code	Internal anonymizing tracking number for panelist responses [PRIMARY KEY]
Solvent	Diluting solvent, if needed for safety or for intensity balancing
Final []	Concentration of molecule (w/w) in final sample
GC-O no of beads used	Number of viscopearls with absorbed odorant used for headspace extraction prior to GC-O analysis
GC-O incubation and extraction temperature	GC-O methodological details
GC-O incubation time	
GC-O extraction time	
GC-O result	Verdict from GC-O analysis
GC-O contaminant, if identified	Canonical SMILES of the causal contaminant, if one was successfully identified
Impact on GNN performance	Whether the GC-O 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 397 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 397 molecules by a random forest model trained on GS-LF datasets.

Data S5. Odor attribute predictions on 397 molecules by a graph neural network model trained on GS-LF datasets. Final layer.

Data S6. Two dimensional coordinates (for visualization) of the high-dimensional graph neural network embeddings on 397 molecules. Penultimate layer.

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

References and Notes

1. T. Smith, J. Guild, The C.I.E. colorimetric standards and their use. *Trans. Opt. Soc.* **33**, 73–134 (1931). doi:10.1088/1475-4878/33/3/301

2. E. F. Evans, Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus. *Psychophys. Physiol. Hear.* **79**, 185–192 (1977).

3. C. S. Sell, On the unpredictability of odor. *Angew. Chem. Int. Ed.* **45**, 6254–6261 (2006). doi:10.1002/anie.200600782 Medline

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**, e1003184 (2013). doi:10.1371/journal.pcbi.1003184 Medline

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). doi:10.1038/s41586-020-2891-7 Medline

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; DREAM Olfaction Prediction Consortium, Predicting human olfactory perception from chemical features of odor molecules. *Science* **355**, 820–826 (2017). doi:10.1126/science.aal2014 Medline

7. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, "Message passing neural networks" in *Machine Learning Meets Quantum Physics*, K. T. Schütt, S. Chmiela, O. Anatole von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller, Eds., vol. 968 of Lecture Notes in Physics (Springer, 2020), pp. 199–214.

8. B. Sanchez-Lengeling, E. Reif, A. Pearce, A. Wiltschko, A gentle introduction to graph neural networks. *Distill* **2021**, 1 (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). doi:10.1021/c160017a018

10. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017). doi:10.1145/3065386

11. F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A unified embedding for face recognition and clustering" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), 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* (International Speech Communication Association, 2012).

13. The Good Scents Company, The Good Scents Company information system; http://www.thegoodscentscompany.com/.

14. Leffingwell & Associates, 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. arXiv:1412.6980 [cs.LG] (2014).

16. D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, D. Sculley, “Google Vizier: A service for black-box optimization” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2017), pp. 1487–1495.

17. B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, A. B. Wiltschko, Machine learning for scent: Learning generalizable perceptual representations of small molecules. arXiv:1910.10685 [stat.ML] (2019).

18. S. Kearnes, Pursuing a prospective perspective. *Trends Chem.* **3**, 77–79 (2021). doi:10.1016/j.trechm.2020.10.012

19. H. T. Lawless, H. Heymann, *Sensory Evaluation of Food: Principles and Practices*, Food Science Text Series (Springer, 2010).

20. 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.* **116**, 9475–9480 (2019). doi:10.1073/pnas.1804106115 Medline

21. 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). doi:10.1186/1471-2202-13-122 Medline

22. A. Dravnieks, Odor quality: Semantically generated multidimensional profiles are stable. *Science* **218**, 799–801 (1982). doi:10.1126/science.7134974 Medline

23. K. J. Rossiter, Structure–odor relationships. *Chem. Rev.* **96**, 3201–3240 (1996). doi:10.1021/cr950068a Medline

24. O. R. P. David, A chemical history of polycyclic musks. *Chemistry* **26**, 7537–7555 (2020). doi:10.1002/chem.202000577 Medline

25. 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). doi:10.1523/ENEURO.0070-17.2017 Medline

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

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

28. E. C. Crocker, L. F. Henderson, *Analysis and Classification of Odors: An Effort to Develop a Workable Method* (Robbins Perfumer Company, 1927).

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

30. J. E. Amoore, “A plan to identify most of the primary odors” in *Olfaction and Taste III: Proceedings*, C. Pfaffmann, Ed. (Rockefeller Univ. Press, 1969), pp. 158–171.

31. 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). doi:10.1016/j.neunet.2019.01.012 Medline

32. 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). doi:10.1073/pnas.2116576119 Medline

33. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. arXiv:2005.14165 [cs.CL] (2020).

34. G. Branwen, The scaling hypothesis (2020); https://gwern.net/scaling-hypothesis.

35. A. Dravnieks, *Atlas of Odor Character Profiles*, ASTM Data Series (ASTM, 1985).

36. 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). doi:10.1093/chemse/bjr094 Medline

37. K. Sechidis, G. Tsoumakas, I. Vlahavas, "On the stratification of multi-label data" in vol. 6913 of Lecture Notes in Computer Science, subseries Lecture Notes in Artificial Intelligence (Springer, 2011), pp. 145–158.

38. RDKit; 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). doi:10.1016/j.foodqual.2015.11.003

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* (Association for Computing Machinery, 2001), 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 and Data Mining* (Association for Computing Machinery, 2018), pp. 2847–2856.

42. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 [stat.ML] (2017).

43. M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognit. Lett.* **150**, 228–234 (2021). doi:10.1016/j.patrec.2021.06.030

44. 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).

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