PNAS RESEARCH ARTICLE | NEUROSCIENCE

A semantic-based community model for high-fidelity tuning of olfactory mixture distances

Vahid Satarifard[a,1], Laura Sisson[b,2], Yikun Han[c,2], Pedro Ilídio[d,e,2], Matej Hladiš[f,g,2], Maxence Lalis[f,2], Xuebo Song[h], Tiffany Yang[h], Wenjie Yin[i], Aharon Ravia[j], CiCi Xingyu Zheng[k], Gaia Andreoletti[l], Jake Albrecht[l], Robert Pellegrino[h,m], Zehua Wang[c], Stephen Yang[c], Robbe D'hondt[d,e], Achilleas Ghinis[d,e], Jasper de Boer[d,e], Felipe Kenji Nakano[d,e,n], Alireza Gharahighehi[d,e], Jose M. G. Vilar[o,p], Leonor Saiz[q], DREAM Olfactory Mixtures Prediction Consortium[3], Benjamin Sanchez-Lengeling[r], Andreas Keller[s], Leslie B. Vosshall[s,t], Sébastien Fiorucci[f], Ambuj Tewari[c], Jérémie Topin[f], Celine Vens[d], Mårten Björkman[i], Danica Kragic[i], Noam Sobel[u], Nicholas A. Christakis[a], Joel D. Mainland[h,v], and Pablo Meyer[w,1]

Affiliations are included on p. 9.

Edited by Martin S. Banks, University of California, Berkeley, CA; received April 10, 2026; accepted July 6, 2026

A central goal in sensory science is to establish quantitative mappings between physical stimuli and perceptual experience. Although such mappings are well defined in vision and audition, they remain elusive in olfaction, particularly for complex odor mixtures. Here, we show that perceptual distances between odor mixtures can be predicted with high fidelity and are unexpectedly well captured by a compact semantic space derived from single-molecule representations. In the Dialogue for Reverse Engineering Assessment and Methods Olfactory Mixtures Prediction Challenge, we assembled a unified dataset of odor-mixture pairs, benchmarked predictions on a hidden test set of 46 pairs, and integrated the top-performing models into a postchallenge ensemble. This model outperformed existing state-of-the-art approaches on the hidden test set, reducing RMSE by about 33% to 0.08 and increasing Pearson correlation by 53% to 0.57, and maintained strong performance on an independent validation set of 50 newly designed mixture pairs. An ensemble, retaining only olfactory semantic features for each model included, further improved predictions, raising the Pearson correlation by 7% to 0.61 on the test set and by 15% to 0.54 on the validation set. Given that semantic features were extracted from pure molecules, it suggests that mixture perception may not require fundamentally different representational principles from single-molecule olfaction. Together, these results establish a reproducible quantitative framework for olfactory mixture perception and advance efforts to measure, model, and engineer smell.

olfaction | odor mixtures | machine learning | psychophysics | digital olfaction

A central goal in sensory science is to discover how physical stimuli are represented in perceptual experience and to establish quantitative mappings between them. In vision, the trichromatic theory helps explain how combinations of three primary colors create our rich visual world. In audition, the superposition of sound waves at different frequencies produces complex tones and harmonies. These quantitative frameworks have enabled technologies from color television to digital audio, transforming how we capture, transmit, and reproduce sensory information. In olfaction, however, no comparable framework exists. We lack fundamental rules for predicting how molecular combinations produce odor perceptions, hindering both scientific understanding and technological applications in flavor and fragrance design.

However, recent advances in machine learning have made significant progress toward predicting the olfactory perception of single molecules from chemical structure. The first DREAM (Dialogue for Reverse Engineering Assessment and Methods) olfaction challenge (1) demonstrated that computational models could predict perceptual descriptors like "garlic," "sweet," and "floral," achievements later expanded to broader descriptor sets (2). The development of graph neural networks and larger datasets enabled construction of a Principal Odor Map (POM) (3) that maps molecular structures into perceptual space and predicts perceptual descriptors better than the median human panelist.

Most odors we encounter, however, are complex mixtures of dozens or hundreds of molecules. Although some models predict mixture perception from chemical structure (4–7), open competitions set clear benchmarks and often lead to improved predictions. With this motivation, we launched the second DREAM olfactory mixtures prediction challenge (8), inviting international teams to develop predictive models of perceptual

Significance

Olfaction lacks a general quantitative framework comparable to those long established for vision and hearing, especially for complex mixtures that dominate real-world smells. Here, a community effort shows that perceptual distances between odor mixtures can be predicted with high fidelity by combining machine learning, shared benchmarks, and psychophysical data. Strikingly, the best-performing models rely heavily on compact semantic features learned from individual molecules, rather than on molecular structure alone, meaning that linguistic representations of single molecules can capture the space of mixture perception. Beyond a conceptual advance, this work provides a validated metric and benchmark for comparing smells, laying a foundation for digital olfaction, machine-readable smell metrics, next-generation electronic nose technologies, and odor teleportation.

[1]To whom correspondence may be addressed. Email: vahid.satarifard@yale.edu or pmeyerr@us.ibm.com.

[2]L. Sisson, Y.H., P.I., M.H., and M.L. contributed equally to this work.

[3]The list of consortium authors and affiliations are listed in *SI Appendix*.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2611057123/-/DCSupplemental.

Published August 4, 2026.

similarity between pairs of molecular mixtures. For training, we curated a unified dataset from multiple psychophysical studies (4, 5, 9). Over three months, 26 teams competed to minimize prediction error on a hidden test-set of 46 mixture pairs, employing diverse machine-learning architectures and molecular feature representations.

The competition ended in a four-way tie among top performers. We combined predictions from the six highest-scoring teams into an ensemble model that outperformed all individual submissions and surpassed existing state-of-the-art methods, including the Snitz angle metric (4), POM-based approaches (3), semantic models (6), and aroma-chemical pair models (7). The ensemble achieved a median RMSE of 0.08 and Pearson correlation of 0.57 on the test-set. To validate these results, we generated an independent dataset of 50 mixture pairs and confirmed the model's superior performance while establishing test–retest reliability benchmarks for human olfactory perception. Parsimonious models reveal the importance of perceptual features in accurate mixture predictions. By publishing all models in open-source format, we provide a quantitative, machine-readable framework for odor similarity that represents a crucial step toward understanding mixture perception and achieving the goal of capturing, transmitting, and reproducing olfactory information.

Results

The DREAM Olfactory Mixtures Prediction Challenge and Dataset. For this competition, we standardized and compiled six datasets of odor-similarity measurements from three different studies (4, 5, 9) (*Materials and Methods* and *SI Appendix*, Table S2) into a unified dataset, comprising 168 unique monomolecules, 731 unique mixtures, and 507 mixture pairs measurements. Mixture pair distances were mapped onto a continuous perceptual scale from 0 (indistinguishable) to 1 (maximally distinct), see Fig. 1*A* and *SI Appendix*, Fig. S6.

Teams competed for roughly three months, competition rules are presented in *Materials and Methods*, to develop machine-learning models to predict how close two molecular mixtures are in odor perceptual space, with the objective of minimizing the RMSE and maximizing Pearson correlation, respectively. In the competition phase of the challenge, teams were allowed to submit predictions for a Leaderboard holdout set of 46 pairs of mixtures, up to three times per week and receive feedback on the performance of their model. The ultimate goal of the challenge was to predict olfactory mixture distance for a hidden test-set of 46 mixture pairs (Fig. 1 *A*, *Right*). The test-set consisted of three subtypes: random, manual, and angle. The random subtype contained mixture pairs generated without design constraints. The manual subtype included mixtures designed to be perceptually similar, and the angle subtype consisted of mixtures designed to be similar according to the Snitz angle distance metric (4). Further methodological details are provided in *Materials and Methods*. The identities of all subtypes were concealed from participants throughout the competition.

DREAM organizers evaluated all submitted models using 10,000 bootstrap iterations, with RMSE and Pearson correlation as performance metrics, see *Materials and Methods*. The competition resulted in a four-way tie among the top-performing teams: *D2Smell*, *UMich CASI*, *ChemSenSim Lab*, and *belfaction*. We then built an ensemble model by averaging the predictions from these four teams with two additional high-performing models (*PL21*, and *Tamarin*, see *SI Appendix*, Table S1 for inclusion and exclusion criteria). A summary of these six models is presented in *SI Appendix*, Table S3 with detailed descriptions of each model available in *SI Appendix*. All other models were excluded either because of underperformance compared to random shuffled baseline, or they lacked open-source code for reproducibility, see *SI Appendix*, Fig. S4. Distributions of RMSE and Pearson correlation for individual models and the ensemble are shown in Fig. 1 *B* and *C*. To estimate the performance ceiling imposed by intrinsic measurement noise, we calculated within-subject split-half reliability, where the first half of the trial measurements is used as test, and the second half is used as retest (red dashed lines in Fig. 1 *B* and *C*; see *SI Appendix*, Fig. S5). The ensemble model (pink distributions in Fig. 1 *B* and *C*) outperforms all individual models with a median RMSE of 0.08 and median Pearson correlation of 0.57 as indicated by gray dashed lines in Fig. 1 *B* and *C*. The ensemble model performed best on manual and angle subtypes, with Pearson correlation of 0.38 for both, and RMSE of 0.07 and 0.06, respectively. For the random subtype, it achieved an RMSE of 0.10 and slightly higher Pearson correlation of 0.45 (see *SI Appendix*, Fig. S7 for subtype distributions and *SI Appendix*, Table S4 for full performance summary).

Analysis of the residuals (the difference between model predictions and ground truth) for each test-set measurement shows that models generally tend to underestimate perceptual similarity (*SI Appendix*, Fig. S7). All teams except *PL21* exhibit a net negative residual sum. The random subtype also showed larger residuals than the manual and angle subtypes. Among all models, *PL21* displayed a distinct performance profile, performing relatively well on the random subtype but poorly on the manual and angle subtypes (*SI Appendix*, Fig. S7 and Table S4).

Model performance also correlated with the density of training data (Fig. 1*D*). Prediction error is lowest for mixtures that were rated in the midrange (approximately 0.5 and 0.8), where data are most abundant. In contrast, errors increase for distances below 0.5 or above 0.8, likely reflecting the scarcity of training examples in these regions.

The Ensemble Model Outperforms the State of the Art. To evaluate the performance of our ensemble model, we compared it with four state-of-the-art (SOTA) models: the Snitz angle model (4, 5), the open-POM model [an open source implementation of POM (3, 10)], a semantic model (6), and an aroma chemical pair model (7). Computational details are provided in *Materials and Methods* and *SI Appendix*, Fig. S8.

To estimate the distributions of RMSE and Pearson correlation we performed bootstrap sampling ($n = 10{,}000$) (*Materials and Methods*). To establish a chance-level baseline, we shuffled the experimental distance values. Fig. 2 *A* and *B* shows the resulting distributions for the shuffled ensemble, individual SOTA models (linear fits) and the ensemble model. All SOTA models outperformed the random shuffle baseline (Fig. 2*A*) and all achieved Pearson correlations above chance level, with the ensemble model performing best (Fig. 2*B*). Across test-set subtypes, the largest improvements occurred for the manual and random subtypes: RMSE decreased by 0.039 ± 0.004 and 0.047 ± 0.005 while Pearson correlation increased by 0.29 ± 0.07 and 0.26 ± 0.10, respectively (*SI Appendix*, Fig. S9). When averaged across all subtypes, the ensemble model reduced RMSE by 0.035 ± 0.004 and increased Pearson correlation by 0.22 ± 0.05. Overall, the ensemble model consistently outperformed all SOTA models and combining SOTA predictions with the ensemble models decreased the performance (*SI Appendix*, Fig. S9 and Table S4). Furthermore, we quantified the magnitude

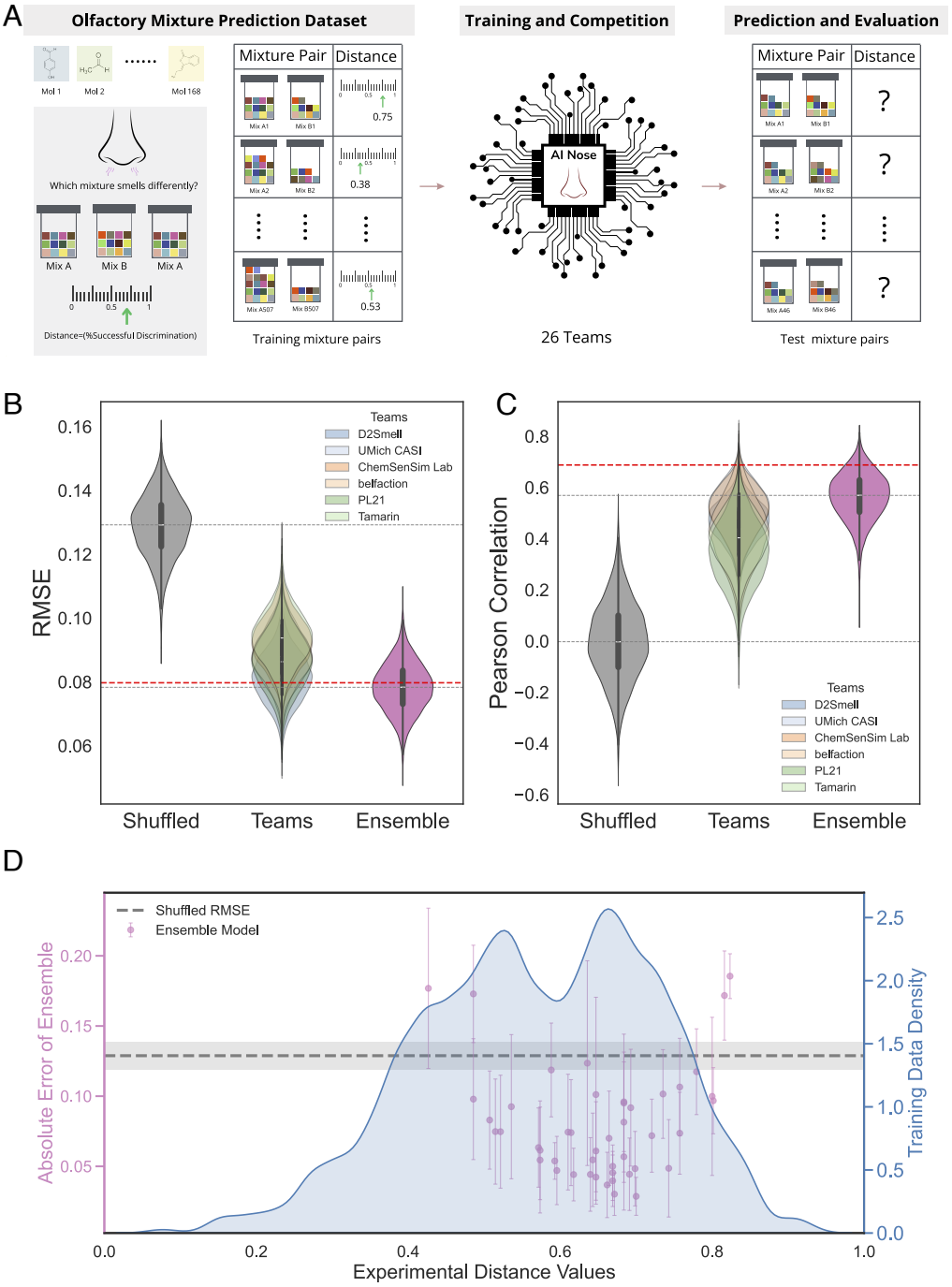

Fig. 1. Overview of the challenge. (*A*) Schematic of the challenge workflow. *Left*: Participating teams received a curated training dataset of 507 olfactory mixture pairs with experimentally measured distances ranging from 0 (indistinguishable) to 1 (maximally distinct). *Middle*: Teams developed machine learning (ML) models to predict the perceptual distance of unseen mixture pairs, optimizing for minimal RMSE and maximal Pearson correlation. The hidden test-set comprised 46 mixture pairs. *Right*: 19 teams submitted final predictions for all 46 hidden test-set comparisons. (*B* and *C*) Distributions of RMSE (*B*) and Pearson correlation (*C*) for the six top-performing teams, the ensemble model, and a shuffled ground truth baseline, obtained via bootstrap sampling ($n = 10{,}000$). Gray dashed lines indicate the median of the ensemble and shuffled distributions; red dashed lines denote within-subject split-half reliability (first half of the trials is used as test, and the second half as retest). (*D*) Absolute prediction error of the ensemble model (*Left* y axis) as a function of experimental perceptual distance (x axis) overlaid with the training data density distribution (*Right* y axis). The gray dashed line marks the shuffled baseline, and the gray band represents its SD. Error bars show SD across teams.

of performance differences using Cohen's d, where we observe a large effect size for the ensemble models relative to individual SOTA models (*SI Appendix*, Table S5).

We next analyzed relative error, defined as $(\frac{\text{prediction}}{\text{ground truth}} - 1)$ for each SOTA model and the ensemble model (*SI Appendix*, Fig. S10). Across all models and test-set subtypes, relative error showed a negative correlation with experimental distance values. This pattern indicates that all models tend to bias predictions toward intermediate values, overestimating lower perceptual distances and underestimating higher ones. The increase in relative error at smaller distances is expected, since dividing by small ground truth values amplifies relative differences. Error growth at both extremes also reflects limited training data in

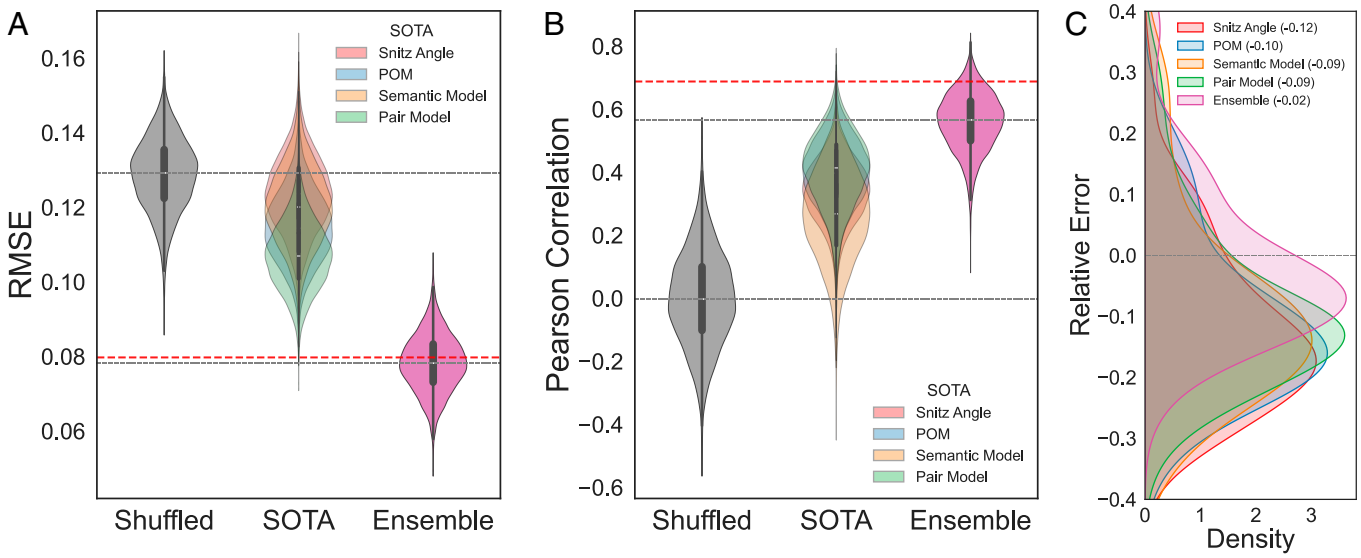

Fig. 2. The ensemble model outperforms existing state-of-the-art approaches. (*A* and *B*) Distributions of RMSE (*A*) and Pearson correlation (y axis) (*B*) for the ensemble model (pink), individual state-of-the-art (SOTA) models and a shuffled ground-truth baseline (gray) obtained via bootstrap sampling ($n = 10{,}000$). Gray dashed lines indicate median values for the ensemble model and shuffled baseline, and the red dashed line marks the across-subject split-half reliability (first half as test, second half as retest). (*C*) Kernel density estimates of relative-error distributions for each SOTA model and the ensemble model on the test-set. The dashed line at 0 indicates no bias; legend values report the mean values of relative error for each model.

these regions (Fig. 1*D*). As shown in Fig. 2*C*, the ensemble model's relative error distribution is shifted toward lower errors compared with the SOTA models, with a mean value of -0.02 ± 0.13.

Optimal Features Emerge from Feature Importance Analysis. To quantify feature importance across six different models we conducted a SHAP (SHapley Additive exPlanations) value analysis (11, 12) (Fig. 3). Each model included over 100 individual features, so we grouped them into feature sets for interpretability (x-axis labels in Fig. 3*A*). For each set, feature importance was defined as the mean absolute SHAP value across test-set predictions; error bars represent the SD of these means (x-axis in Fig. 3*A*). Detailed definitions of the feature sets used in each model are provided in *SI Appendix*.

For the D2Smell team, features derived from pair model labels (7) contributed most strongly to predictions across the test-set, with a mean absolute SHAP value of 0.11 ± 0.019. The UMich CASI team used 138 semantic labels from the POM model, grouped into 12 olfactory categories. The top three contributors were "Fruity," "Green & Herbal," and "Nutty, Fatty& Roasted," with mean absolute SHAP values of 0.04 ± 0.007, 0.03 ± 0.006, and 0.03 ± 0.005, respectively. The ChemSenSim Lab team utilized the full POM embedding, rather than probabilities, along with features such as predicted olfactory receptor responses, experiment type, and relative mixture size. SHAP analysis showed that the POM embedding contributed most to predictive performance (0.06 ± 0.014). The belfaction team used 30 principal components of the POM embedding and summarized mixture-level properties using statistical descriptors (mean, product, absolute difference, minimum, maximum, and U statistic). Among these, the absolute difference features provided the greatest contribution (0.10 ± 0.004). The PL21 team used a combination of chemical and semantic descriptors; among these, Dragon descriptors contributed most to predictions (0.081 ± 0.040). Finally, team Tamarin's predictions were driven primarily by a combination of DeepNose and Mordred features (0.06 ± 0.042).

Among all feature types, POM-derived representations were the most widely used, appearing as probabilities (UMich CASI), full embeddings (ChemSenSim Lab), and dimension-reduced embeddings (belfaction). To further investigate their influence, we analyzed the relationship between SHAP values and prediction errors for the probability-based semantic labels used by UMich CASI (Fig. 3*B*). Semantic labels showing statistically significant correlations between SHAP values and prediction errors are shown along the x-axis of Fig. 3*B*. Positive Spearman correlations (red) indicate that higher SHAP values were associated with larger prediction errors, whereas negative correlations (blue) indicate that greater SHAP values were associated with smaller prediction errors. The labels "violet," "aromatic," and "natural" most strongly increased prediction errors, whereas "green," "clean," and "cognac" were most strongly associated with decreased errors. Although these relationships are not straightforward to interpret, owing to the lack of experimental data directly linking semantic labels to perceptual mixture outcomes, they nonetheless provide an empirical basis for selecting informative POM features in future model development. In addition, this pattern cannot be explained by label frequency of the Good Scents and Leffingwell data used for training POM (3, 10) (*SI Appendix*, Fig. S12).

To assess model compactness, we analyzed cumulative feature importance derived from SHAP values as a function of the number of features. Fig. 3*C* compares how efficiently each model concentrates its explanatory power within its highest-ranked features (*Materials and Methods*). Using the 50% cumulative importance threshold (horizontal gray dashed line in Fig. 3*C*), the Tamarin, D2Smell, and UMich CASI models were the most compact, achieving this threshold with fewer than 100 features. In contrast, the ChemSenSim Lab and belfaction models were less compact, requiring approximately 100 and 1,000 features, respectively, to reach the same threshold. All models eventually plateaued to 100%, though the least compact models (i.e., ChemSenSim Lab and belfaction) exhibit long tails of low-importance features that add little explanatory power. Finally, to assess the stability of feature-importance

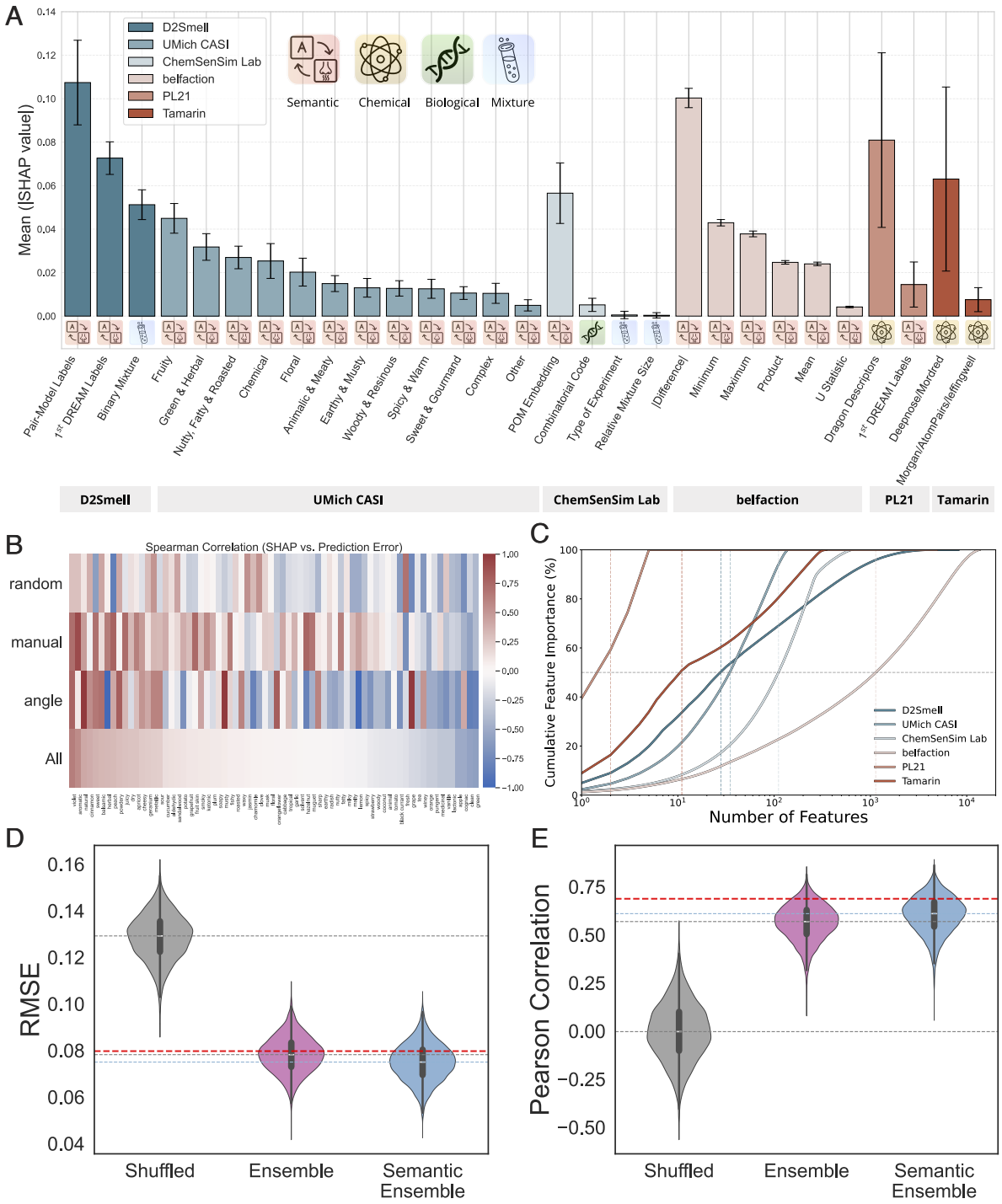

Fig. 3. Feature importance analysis of top-performing models. (*A*) Mean absolute SHAP (SHapley Additive exPlanations) values for grouped feature classes across six teams. Larger values indicate greater contribution to model predictions. Error bars represent the SD. Different classes of features (semantic, chemical, biological, and mixture) are differentiated with different symbols in the figure *Inset*. (*B*) Heatmap of Spearman's rank correlations between SHAP values and prediction errors of POM semantic labels across test-set subtypes (y-axis). (*C*) Cumulative feature importance of different models (y-axis) as a function of the number of features; the horizontal dashed line marks the 50% importance threshold; vertical dashed lines mark the number of features required to cross 50% importance threshold. (*D* and *E*) Distributions of RMSE (*D*) and Pearson correlation (y axis) (*E*) for the ensemble model (pink), the semantic ensemble model (blue) and a shuffled ground-truth baseline (gray) obtained via bootstrap sampling ($n = 10{,}000$). Gray dashed lines indicate median values for the ensemble model and shuffled baseline, and blue dashed line for the median of semantic ensemble model and the red dashed line marks the across-subject split-half reliability.

ranking, we calculated Kendall's coefficient of concordance (W) (*Materials and Methods*). The resulting stability distributions are shown as ridgeline plots in *SI Appendix*, Fig. S12 for each test-set subtype and for all subtypes combined. Across models, W values peaked around 0.5, indicating moderate, but not strong, agreement in feature rankings. For the angle subtype, stability varied more widely across models, likely due to its smaller sample size (10 mixture pairs) relative to random (16 mixture pairs) and manual (20 mixture pairs) subtypes.

To test whether semantic features are necessary and sufficient for accurate prediction, we conducted a systematic ablation study, by removing all chemical, biological, and mixture

feature sets and retraining models using only semantic features (*Materials and Methods*). We then constructed a semantic ensemble by averaging the predictions of these semantic-only models. This ensemble represents a prediction based exclusively on predicted semantic features, without access to any chemical, biological, or mixture information. The semantic ensemble achieved a median Pearson correlation of 0.61 and a median RMSE of 0.075, exceeding both the shuffled baseline and the full ensemble model predictions. (Fig. 3 *D* and *E* and *SI Appendix*, Table S4). In contrast, the nonsemantic ensemble model performed drastically worse, with higher RMSE 0.086 and a sharp drop in Pearson correlation 0.41 (*SI Appendix*, Fig. S15 and Tables S4 and S5).

In the postchallenge phase, insights from the SHAP analysis, guided the development of a model that integrated a unified feature set combining open-POM, pair model, Dragon, Mordred, and DeepNose features. Detailed specifications of this model are provided in *SI Appendix*, Table S3.

Ensemble Model Maintains High Predictive Accuracy in Novel Validation Set. In the postchallenge phase, we developed a validation set to evaluate both the test–retest reliability of experimental measurements and the performance of the ensemble model on an independent dataset. We first generated 10 base mixtures using molecular co-occurrence probabilities designed to mimic naturally occurring mixtures. Then we used POM predictions to generate nonoverlapping target mixtures spanning 5 different distance quintiles. In total, 50 nonoverlapping mixture pairs were experimentally tested using direct odor similarity ratings (Fig. 4*A* and *Materials and Methods*). For calibration, we also measured similarity ratings for two single-molecule reference pairs: diallyl disulfide (garlic odor) versus allyl caproate (pineapple odor), and strawberry aldehyde (strawberry odor) compared with itself.

Fig. 4*B* shows the mean experimental distance values for the first (test, x-axis) and second (retest, y-axis) sessions. Blue points represent 50 mixture pairs, while the two reference molecular pairs are shown in red. The distributions of test and retest similarities are shown in light blue and light peach, respectively. A linear regression fitted to the regular mixture pairs (blue points) is shown as a gray solid line with a 95% confidence band. Mean experimental distances were strongly correlated between test and retest sessions (Pearson's $r = 0.74$, P-value = 7.27e-10). The diallyl disulfide (garlic odor) and allyl caproate (pineapple odor) are judged far apart, and Strawberry aldehyde (strawberry odor) compared to itself was judged very close, for both test and retest sessions. For regular mixture pairs, the distance measurements ranged in [0.29,0.76] interval for test and [0.31,0.78] interval for retest sessions, respectively.

We partitioned the measured perceptual distances into five quintiles, and computed model residuals for each mixture pair. Fig. 4*C* shows a combined residual heatmap and absolute-error bar plot across all teams. Consistent with the test-set results, the validation-set exhibited a regression-to-the-mean pattern: Models tend to overestimate perceptual similarity at lower distances and underestimate it at higher distances. The solid green, pink, and blue horizontal lines indicate the mean quintile errors for the POM, ensemble, and semantic ensemble model, respectively. In the most similar quintile (Q1 [0.31,0.44]), the POM model showed lower error, whereas across the remaining quintiles (Q2-Q5 [0.45,0.77]), the ensemble and semantic ensemble models consistently outperformed POM with smaller errors in each quintile.

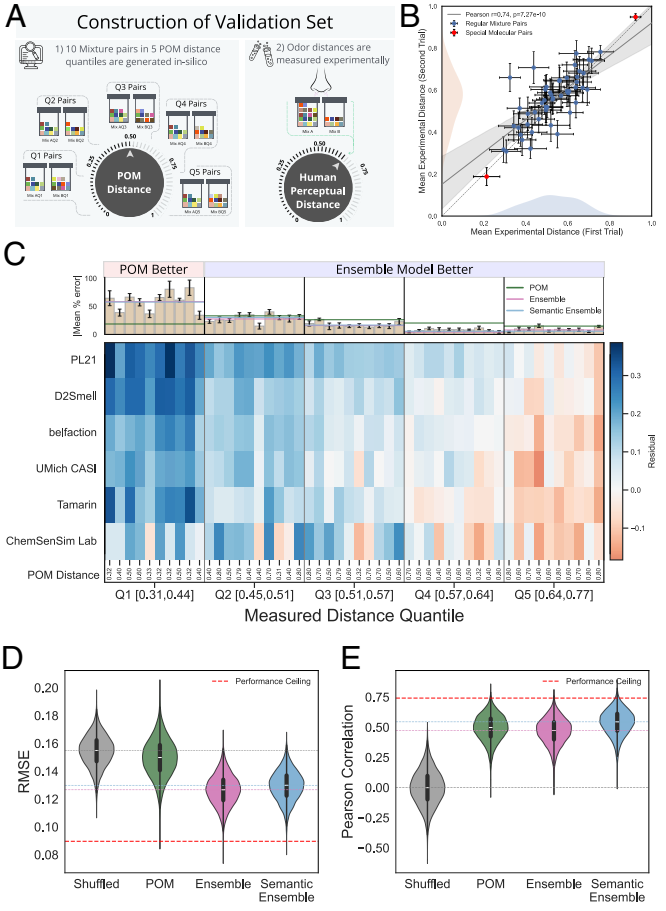

Fig. 4. Construction of the validation set. (*A*) The in silico mixture pair generations using POM in 5 distance quintiles (*Left*), and experimental measurement of perceptual distance by human subjects (*Right*), are shown. (*B*) The mean experimental distance values are shown for test (x-axis) and retest (y-axis) sessions using blue and red data-points. Blue data points correspond to 50 regular mixture pairs, and two red data points correspond to strawberry, garlic, and pineapple smelling molecule pairs; *Top Right* red point corresponds to garlic versus pineapple rating and *Bottom-Left* red point corresponds to strawberry rating against itself, respectively. The error bars are SEs of the mean. The gray solid line shows the linear regression of regular mixture pairs with 95% confidence band. The light blue and light peach distributions show the distance distributions of test and retest sessions, respectively. (*C*) Residual heatmap showing team-experiment residuals, defined as the difference of model predictions and ground truth. Columns are grouped by measured distance quintiles; labels under the heatmap give each pair's POM distance. *Top* bars plot mean absolute percent error along with SEs of the means across teams per pair; three reference lines show quantile means for POM (green), Ensemble (pink), and Semantic Ensemble (blue). (*D* and *E*) Distribution of RMSE (*D*) and Pearson correlation (y axis) (*E*) for POM, ensemble and semantic ensemble model (x axis), obtained via bootstrap sampling ($n = 10{,}000$). The gray distribution represents chance-level performance from shuffled experimental distance values. The pink and blue distribution represent ensemble and semantic ensemble model's predictions, respectively. Pink, blue, and gray dashed lines indicate median values for respectively the ensemble, semantic ensemble models, and shuffled baseline, and the red dashed lines are the performance ceilings obtained from test–retest reliability in two different trials.

To evaluate the ensemble and semantic ensemble model's performance on the validation set, we performed bootstrap sampling ($n = 10{,}000$) to estimate the distribution of RMSE and Pearson correlation for the POM, ensemble, and semantic ensemble models (*Materials and Methods*). We also established a chance-level baseline by randomly shuffling the experimental distance values. As shown in Fig. 4 *D* and *E*, all models outperformed the shuffled baseline. Both ensemble and semantic ensemble models exhibited comparable RMSE (Fig. 4*D*), but

the semantic ensemble model achieved the highest Pearson correlation (see Fig. 4*E* and *SI Appendix*, Tables S4 and S5 for performance summary and effect size). This was not the case for the nonsemantic ensemble model, which performed markedly worse on both metrics (*SI Appendix*, Fig. S15). No model reached the performance ceiling defined by the upper bound of the measurement error, as determined by the test–retest correlation.

Discussion

Prior to the training of advanced machine learning (ML) models made possible by the generation of the necessary data, predicting the odor perception of single molecules from their molecular structure had long remained a notoriously difficult problem (13). This challenge has been addressed in the last decade, as demonstrated through the first DREAM olfaction challenge (1) and the development of the Principal Odor Map (POM) (3). However, achieving high-fidelity prediction of olfactory mixture similarities remained an open problem. In this work, we leveraged a large-scale community effort, through the second DREAM Olfaction Challenge, a global collaborative effort to improve the prediction of olfactory mixture similarity. We unified six heterogeneous psychophysical datasets from three separate studies to create a comprehensive training dataset. Twenty-six international teams independently developed ML models to predict odor similarity, optimizing for low RMSE and high Pearson correlation on a hidden test-set of 46 mixture pairs. The competition resulted in a four-way tie among top-performing models. Following the challenge, an independent validation set of 50 olfactory mixtures was developed to estimate prospective model reliability.

From the 19 submitted models, we analyzed the six top-performing submissions, all of which achieved RMSE < 0.1 and provided fully reproducible and open-source code. The competition resulted in a four-way tie among top performers (D2Smell, UMich CASI, ChemSenSim Lab, and belfaction). This tie likely reflects both the relatively small size of the test-set (46 mixture pairs) and the fact that it was drawn from a different distribution than the training data, making fine discrimination among similar high-performing models difficult. The tie does not represent a performance ceiling, as split-half reliability measurements indicate further room for improvement. Averaging predictions from the six top models produced an ensemble model with a median RMSE of 0.08 and a Pearson correlation of 0.57 on the test-set. Additionally, using an ablation study, we constructed a semantic ensemble model that further improved prediction performance, achieving a median RMSE of 0.075 and a Pearson correlation of 0.61 on the test set. Both ensemble and semantic ensemble models outperformed each individual model as well as four SOTA baselines: the Snitz angle model, the POM model, the semantic model, and the aroma-chemical pair model. On the prospective validation set, both models maintained strong performance, with an RMSE of 0.13, while the semantic ensemble model outperformed POM with a Pearson correlation of 0.54. The ensemble's superior performance indicates that combining diverse weak learners can better approximate the underlying latent perceptual manifold while reducing noise, consistent with the advantages of ensemble learning observed in computer-vision, natural language processing, and many other domains (14–16).

Most individual models, and therefore the ensemble, performed best on the manual and angle subtypes of the test-set, particularly in reducing prediction error. This suggests that the top models captured higher-order perceptual structure of systematically designed olfactory mixture pairs compared to randomly generated mixtures. Several aspects of our evaluation are constrained by the size and composition of the test set. With 46 mixture pairs distributed across three subtypes, the bootstrap CIs for RMSE and Pearson correlation are correspondingly wide, particularly for subtype-specific estimates where the effective sample size is smallest (see *SI Appendix*, Table S4). This limited statistical power likely contributed to the four-way tie among top-performing models; although their median performances differ slightly, the overlapping CIs indicate that these differences cannot be reliably distinguished from sampling variability. The angle subtype ($n = 10$) is especially underpowered, and subtype-specific claims involving this subset should be interpreted with caution. Distinguishing among top-performing models, characterizing performance in the distributional extremes, and evaluating generalization to novel molecules will ultimately require substantially larger and more systematically balanced held-out datasets. We view this as a key priority for future iterations of olfactory mixture prediction benchmarks.

Despite the wide range of modeling strategies and feature sets across teams, SHAP analysis revealed that semantic descriptors as well as semantic features derived from the POM and the pair model consistently provided the most predictive information. POM features were used in several forms, including semantic label probabilities, embeddings, and dimension-reduced representations, all of which contributed disproportionately to model accuracy. Semantic labels from the pair model also made substantial contributions, particularly in the D2Smell model. Together, these results suggest that although POM was trained exclusively on single molecules, the perceptual distance between olfactory mixtures can also be effectively represented in a low-dimensional semantic space. In other words, predicting mixture perception from component molecules may not be as challenging as previously theorized (17). This challenges the notion that mixture perception involves fundamentally different mechanisms than single-molecule perception and suggests that the primary challenge lies in the initial structure-to-percept mapping rather than the component-to-mixture transformation.

Our systematic ablation study revealed that semantic features carry superior information to predict olfactory mixture similarity. The semantic ensemble outperformed the full challenge ensemble both in test and validation sets, which combined semantic and nonsemantic features. Conversely, the nonsemantic ensemble was the poorest performer, with substantially higher RMSE and a Pearson correlation that fell to near-chance levels on the validation set, which indicates that nonsemantic features alone carry little information for predicting mixture similarity. These results provide direct evidence that semantic descriptors are the primary information-carrying features for predicting olfactory mixture similarity. The findings have two implications. First, it further validates the hypothesis that perceptual similarity between odor mixtures is largely determined by their overlap in perceptual descriptor space, a view that some oppose (18, 19). Second, it suggests that advances in single-molecule odor character prediction translate directly into improved mixture similarity prediction, which establishes a clear path for the development of future models.

Prior studies have shown that POM effectively captures olfactory relationships across varied datasets, outperforming generic molecular descriptors (20). Our results extend this finding to mixture perception and suggest that semantic representations provide the most efficient features for predicting olfactory

perceptual similarity. In terms of model compactness, three of the most accurate individual models reached 50% cumulative SHAP importance using fewer than 100 features, indicating that accurate predictions can be achieved with compact feature sets (Fig. 3*C*).

Despite the ensemble's strong performance, a gap remains between model predictions and human test–retest reliability ($r = 0.74$ on validation set, $r = 0.69$ split-half reliability on test-set). This gap may reflect the inherent difficulty of diagnosing mixture similarity errors. Error in predicting mixture similarity could arise at multiple levels: 1) the structure-to-percept mapping for individual components, 2) the component-to-mixture integration, or 3) the mixture-percept-to-similarity metric. Disentangling these error sources is challenging with current data, as we lack systematic measurements of component percepts, predicted mixture percepts, and similarity judgments within the same experimental framework.

Across all models and test-set subtypes, relative error showed a negative correlation with experimental distance values, indicating that models tend to bias predictions toward intermediate values—overestimating similarity at lower perceptual distances and underestimating it at higher ones. This regression-to-the-mean pattern reflects limited training data at the distributional extremes (Fig. 1*D*). Prediction error is lowest for mixtures rated in the midrange (0.5 to 0.8), where training data are most abundant, and increases for distances below 0.5 or above 0.8.

Several limitations constrain current model performance and generalization. First, although we unified multiple measurements paradigms, including triangle tests and similarity ratings, by mapping all measurements onto a 0 to 1 distance scale, this approach overlooks paradigm-specific biases (*SI Appendix*, Fig. S6) (21) and residual noise arising from different administration techniques [e.g, headspace via pads in jar (5) versus liquid in vials (4, 9)].

Second, the training data are unevenly distributed across the similarity scale, with higher density around moderate distances (0.5 to 0.8). This imbalance occurs because, in our experience, randomly generated mixtures tend to fall within this range. Without a preexisting model to guide mixture design, it is difficult to generate sufficient mixture pairs at the extremes of the distribution. At low perceptual distances, all models overestimated the perceptual distance, as shown in *SI Appendix*, Fig. S10. Improving performance at the distributional extremes will require actively designing and measuring new mixture pairs.

Third, most datasets used here include only intensity-matched components, with the exception of the Ravia dataset. Odor quality can shift with intensity (22–24), and, within mixtures, more intense components can mask weaker ones (25). Since real-world odors rarely consist of intensity-balanced components, future models must account for intensity to improve their generalization to natural odors.

Fourth, our postchallenge effort to develop a refined model using only the highest-contributing features (open-POM logits, pair model embeddings, Dragon and Mordred descriptors, and DeepNose features) achieved performance comparable to individual top performing teams rather than the ensemble (*SI Appendix*, Fig. S13). This suggests that while feature selection and model architecture are important for learning olfactory perceptual similarity, achieving ensemble-level performance requires the diversity of modeling approaches rather than simply identifying optimal features. The postchallenge model's performance may also reflect fundamental model limitations when training data are scarce.

Finally, both the test-set and validation set use molecules from a limited palette and whether models can generalize to completely novel molecules not represented in training remains an open question that will require future evaluation.

Closing the gap between model performance and human reliability will require several advances. Most critically, we need larger, more systematically designed datasets that span the full range of perceptual distances with more specific perceptual information and intensity variation. Ideally, future datasets would include measurements at multiple levels—component percepts, mixture percepts, and pairwise similarity judgments—enabling more precise diagnosis of error sources.

Our results provide a high-fidelity, machine-readable framework for compressing high-dimensional chemical and semantic information into a one-dimensional perceptual metric. This metric enables efficient in-silico experiments that were previously impractical or computationally expensive. For example, virtual olfactory metamers could now potentially be generated efficiently and tested experimentally. Accurate and robust mixture-similarity metrics are foundational for a wide range of technological applications, including content-aware compression for digital olfaction, noninvasive health monitoring (26, 27), generative fragrance design (28–30), next generation electronic nose sensors (31, 32), and development of olfactory foundation models (33). We encourage the community to expand publicly available psychophysical datasets, along with metadata on concentration and measurement techniques. Expanding these olfactory datasets is the most direct path to training more robust, generalizable models capable of learning the underlying structure of olfactory mixture perception and bringing us closer to understanding olfaction.

Materials and Methods

Training Dataset. The training dataset composed of six datasets of odor-similarity measurements from three different studies (4, 5, 9) (*SI Appendix*, Table S2). Each dataset is originally designed with different objectives. For instance, the Snitz dataset (4) focused on developing an odor-similarity metric between mixture pairs based on chemical descriptors, whereas the Ravia dataset (5) was designed to create olfactory metamers, i.e. nonoverlapping mixture pairs with similar olfactory perception. In contrast, the Bushdid dataset (9) was designed to measure human olfactory discrimination capacity by increasing the overlap of mixture pairs under comparison. Because these datasets were designed for different goals, they vary in mixture composition and measurement techniques (34), with some using triangle tests, while others leverage similarity ratings (*SI Appendix*, Fig. S5). For the DREAM Challenge, we standardized and merged available measurements into a unified dataset, comprising 168 unique monomolecules, 731 unique mixtures, and 507 mixture pairs measurements. Mixture pair distances were mapped onto a continuous perceptual scale from 0 (indistinguishable) to 1 (maximally distinct). A detailed discussion on unifying perceptual distance across psychophysical paradigms is provided in *SI Appendix*.

Development of the Test-Set. A total of 34 participants (20 females) aged between 18 and 50 y (median age: 40 y), were recruited at The Rockefeller University. All subjects provided written informed consent. The study procedures were approved by the Rockefeller University institutional review board (IRB: LVO-0857). A total of 92 unique olfactory mixtures that form 46 mixture pairs were used. Each mixture contained 10 nonoverlapping odorous components selected from a previously intensity-matched set of molecules (9). Participants performed triangle test, three-alternative forced-choice discrimination tasks in which they were asked to identify the odd smelling odor vial from three presented vials. For each mixture pair, a total of 136 or 140 measurements were collected (see *SI Appendix* for further details).

Competition. In total, 26 teams submitted 159 Leaderboard predictions, and 19 teams submitted test-set predictions. Detailed information on the competition and its guidelines can be found on the DREAM Challenge webpage (https://www.synapse.org/Synapse:syn53470621/wiki/). The competition rules, results of the DREAM Challenge, including team performance and inclusion/exclusion criteria, are provided in the Appendix. In addition, the detailed implementation of all models presented in this study are provided in *SI Appendix*.

Performance Evaluation Metric. We calculated RMSE and Pearson correlation to evaluate olfactory distance prediction performance. By integrating these two metrics, we provided a comprehensive evaluation of each model's accuracy and predictive power, to ensure that both the magnitude of the prediction errors and the consistency of the predicted trends were taken into account. Throughout the paper, we define a chance-level baseline by randomly shuffling the experimental distance values in both test and validation sets.

State-of-the-Art (SOTA) Models. We implemented SOTA benchmarks using linear, logarithmic, or lasso regression methods. Specifically, we evaluated four established models: Snitz angle (4, 5), open-POM [an open-source implementation of POM (3, 10)], a semantic model (6), and an aroma chemical pair model (7). For the Snitz angle model, we computed the Snitz angle between all mixture pairs in the training dataset using 21 Dragon descriptors (4). We then performed linear and logarithmic regression to identify the best-fitting relationship between Snitz angle and experimental distance values of training set. The best fit is eventually used to predict the test-set olfactory perceptual distance. For POM, we first generated 138 dimensional semantic descriptor probabilities for each molecule individually. Each mixture was then represented by averaging the semantic descriptor probabilities of its constituent molecules, leading to a 138 dimensional vector per mixture. From these vector representations, we calculated distance metrics from a variety of quantities, including Pearson correlation, cosine similarity, Euclidean distance, and angle for all mixture pairs. As with the Snitz angle approach, we applied linear and logarithmic regressions to these metrics in training set and used the best fit as a predictive model to obtain olfactory perceptual distance for test-set. For the semantic model (6), predictions were directly generated from an open-source implementation based on lasso regression. Finally, for aroma chemical pair model (7), we adopted the same computational pipeline as the POM model, but used the mixture-level probabilities derived from pair model descriptors (see *SI Appendix*, Fig. S8 for more details).

Feature Importance Analysis Using SHAP Values. We used SHapley Additive exPlanations (SHAP) values (11, 12) to analyze the importance of different classes of features. To assess model compactness, features were first sorted by their mean absolute SHAP values, and cumulative sums were computed to determine each feature's contribution relative to total feature importance. The slope of this cumulative curve provides a measure of model compactness, indicating how many features are needed to reach a given explanatory threshold. To calculate Kendall's coefficient of concordance (W), test-set predictions were divided into two consecutive, equally sized blocks of mixture pairs. Within each block, features were ranked according to their SHAP values, and Kendall's W was then calculated across blocks to quantify rank stability for each feature.

Ablation Study and Semantic and Nonsemantic Ensemble Model. We performed a systematic ablation study using two complementary analyses, a semantic and a nonsemantic ablation. In the semantic ablation, the D2Smell, ChemSenSim Lab, and PL21 models were retrained using only their semantic feature subsets, and all chemical, biological, and mixture feature sets removed. The UMich CASI and belfaction models were not retrained, as they rely exclusively on semantic features. Predictions from the Tamarin model were excluded from the semantic ablation, as it uses only nonsemantic features. In the nonsemantic ablation, the D2Smell, ChemSenSim Lab, and PL21 models were instead retrained using only their nonsemantic feature subsets. The Tamarin model was used in full, as it relies exclusively on nonsemantic features, whereas the UMich CASI and belfaction models were excluded, as they use only semantic features. All ablated models were trained on the same training data and evaluated on the held-out test and validation sets using RMSE and Pearson correlation. The training-evaluation split was identical to that of the original challenge protocol.

Development of the Validation Set. A total of 16 participants (7 females) aged between 18 and 55 y (mean age: 32 y), were recruited at the Monell Chemical Senses Center. All subjects provided written informed consent. The study procedures were approved by the University of Pennsylvania institutional review board (IRB818208). A total of 59 unique odor mixtures which form 50 mixture pairs were used. Each mixture contained 9 to 11 nonoverlapping odorous components, selected from a set of 94 single molecules that had been intensity-matched. Participants rated the perceptual similarity of odor mixture pairs using a direct similarity rating task conducted in two sessions spaced at least one day apart. The first session was used for the test, and the second session was used for the retest. For each mixture pair, a total of 30 measurements were collected (see *SI Appendix* for further details).

Data, Materials, and Software Availability. All datasets, models, evaluation scripts, and codes to generate figures are available at the following repository https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge, https://github.com/Satarifard/CWYK-Olfboost, https://github.com/chemosim-lab/ChemoDREAM, https://github.com/YikunHan42/DREAM-Olfactory-Mixtures-Prediction-Challenge-CASI, https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge/tree/main/Models/PL21, https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge/tree/main/Models/Tamarin, and https://gitlab.kuleuven.be/u0131222/dream 2024 (35–41)

ACKNOWLEDGMENTS. This research was supported in part by grants from the NIH (R01 DC017757), the NOMIS Foundation; Pershing Square Philanthropies; the Stavros Niarchos Foundation; Rothberg Catalyzer; Paul Graham Foundation; Schmidt Futures; European Research Council SynGrant 101118977 D2Smell; William R. Miller Fellowship; French National Research Agency (ANR) [ANR-19-CE07-0044] (PhD fellowship to M.H.); Fondation Roudnitska under the aegis of Fondation de France (PhD fellowship to M.L.); the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01; ChemSenSim Lab team is grateful to the Université Côte d'Azur's Center for High-Performance Computing (Observatoire Pluridisciplinaire des Alpes-maritimes infrastructure) for providing resources and support; The belfaction team thanks the Flemish Government (AI Research Program) and Fonds Wetenschappelijk Onderzoek, 11A7U26N, 1235924N, and 1S38025N; The University of Michigan Computer Age Statistical Inference team acknowledges the support of a NINI (New Initiatives/New Instruction) grant from the College of Literature, Science, and the Arts at the University of Michigan. L.B.V. is supported by the HHMI.

Author affiliations: [a]Human Nature Lab, Yale University, New Haven, CT 06511; [b]Department of Computer Science, Boston University, Boston, MA 02215; [c]Department of Statistics, University of Michigan, Ann Arbor, MI 48109; [d]Department of Public Health and Primary Care, Katholieke Universiteit Leuven, Kortrijk 8500, Belgium; [e]Itec, imec Research Group at Katholieke Universiteit Leuven, Kortrijk 8500, Belgium; [f]Institut de Chimie de Nice, Université Côte d'Azur,CNRS, Nice 06108, France; [g]Department of Computer Science, University of Oxford, Oxford OX1 3QD, United Kingdom; [h]Monell Chemical Senses Center, Philadelphia, PA 19104; [i]School of Electrical Engineering and Computer Science, Kungliga Tekniska högskolan Royal Institute of Technology, Stockholm 10044, Sweden; [j]Runway Startup, Cornell Tech, Cornell University, New York, NY 10044; [k]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; [l]Sage Bionetworks, Seattle, WA 98109; [m]Department of Animal Science, Texas A&M University, College Station, TX 77843; [n]Computer Science Department, State University of Londrina, Paraná, 86057-970, Brazil; [o]Biofisika Institutua, Consejo Superior de Investigaciones Científicas, Euskal Herriko Unibertsitatea, University of the Basque Country, Bilbao 48080, Spain; [p]IKERBASQUE, Basque Foundation for Science, Bilbao 48009, Spain; [q]Department of Biomedical Engineering, University of California, Davis, CA 95616; [r]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S3E5, Canada; [s]Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, NY 10065; [t]HHMI, The Rockefeller University, New York, NY 10065; [u]Department of Neurobiology, Weizmann Institute of Science, Rehovot, 7610001, Israel; [v]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104; and [w]Health Care and Life Sciences, International Business Machines Research, Yorktown Heights, NY 10598

Author contributions: V.S., B.S.-L., L.B.V., S.F., A.T., J.T., C.V., M.B., D.K., N.S., N.A.C., J.D.M., and P.M. designed research; V.S., L. Sisson, Y.H., P.I., M.H., M.L., X.S., T.Y., W.Y., A.R., C.X.Z., R.P., Z.W., S.Y., R.D., A. Ghinis, J.d.B., F.K.N., A. Gharahighehi, J.M.G.V., L. Saiz, D.O.M.P.C., and A.K. performed research; V.S., L. Sisson, X.S., T.Y., G.A., J.A., A.K., L.B.V., J.D.M., and P.M. contributed new reagents/analytic tools; V.S., L. Sisson, Y.H., P.I., M.H., M.L., J.D.M., and P.M. analyzed data; G.A. and J.A. organized the Dialogue for Reverse Engineering Assessment

and Methods Challenge; B.S.-L., L.B.V., S.F., A.T., J.T., C.V., M.B., D.K., N.S., N.A.C., J.D.M., and P.M. supervised the research; and V.S., J.D.M., and P.M. wrote the paper.

Competing interest statement: J.D.M. serves on the scientific advisory board of Osmo Labs, Public Benefit Corporation and receives compensation for these activities. P.M. serves on the advisory board of Hynt. L. Sisson is currently the Chief Technology Officer of the startup Patina; she was not affiliated with Patina at the time the work reported in this manuscript was conducted.

This article is a PNAS Direct Submission.

1. A. Keller *et al*., Predicting human olfactory perception from chemical features of odor molecules. *Science* **355**, 820–826 (2017).
2. E. D. Gutiérrez, A. Dhurandhar, A. Keller, P. Meyer, G. A. Cecchi, Predicting natural language descriptions of mono-molecular odorants. *Nat. Commun.* **9**, 4979 (2018).
3. B. K. Lee *et al*., A principal odor map unifies diverse tasks in olfactory perception. *Science* **381**, 999–1006 (2023).
4. K. Snitz *et al*., Predicting odor perceptual similarity from odor structure. *PLoS Comput. Biol.* **9**, e1003184 (2013).
5. A. Ravia *et al*., A measure of smell enables the creation of olfactory metamers. *Nature* **588**, 118–123 (2020).
6. A. Dhurandhar, H. Li, G. A. Cecchi, P. Meyer, Expansive linguistic representations to predict interpretable odor mixture discriminability. *Chem. Senses* **48**, bjad018 (2023).
7. L. Sisson, A. A. Barsainyan, M. Sharma, R. Kumar, Deep learning for odor prediction on aroma-chemical blends. *ACS Omega* **10**, 8980–8992 (2025).
8. P. Meyer, J. Saez-Rodriguez, Advances in systems biology modeling: 10 years of crowdsourcing dream challenges. *Cell Syst.* **12**, 636–653 (2021).
9. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. *Science* **343**, 1370–1372 (2014).
10. A. A. Barsainyan, R. Kumar, P. Saha, M. Schmuker. Model and data from "Open Principal Odor Map" GitHub. https://github.com/BioMachineLearning/openpom. Deposited 14 August 2024.
11. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
12. S. M. Lundberg *et al*., From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
13. L. B. Vosshall, Laying a controversial smell theory to rest. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6525–6526 (2015).
14. C. Ju, A. Bibaut, M. van der Laan, The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* **45**, 2800–2818 (2018).
15. M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, P. N. Suganthan, Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).
16. J. Saez-Rodriguez *et al*., Crowdsourcing biomedical research: Leveraging communities as innovation engines. *Nat. Rev. Genet.* **17**, 470–486 (2016).
17. L. Xu, D. J. Zou, S. Firestein, Odor mixtures: A chord with silent notes. *Front. Ecol. Evol.* **11**, 1135486 (2023).
18. M. Kurfalı, P. Herman, S. Pierzchajlo, J. Olofsson, T. Hörberg, Representations of smells: The next frontier for language models? *Cognition* **264**, 106243 (2025).
19. T. Hörberg *et al*., A rose by another name? Odor misnaming is associated with linguistic properties. *Cogn. Sci.* **48**, e70003 (2024).
20. W. W. Qian *et al*., Metabolic activity organizes olfactory representations. *eLife* **12**, e82502 (2023).
21. E. Roessler, R. Pangborn, J. Sidel, H. Stone, Expanded statistical tables for estimating significance in paired–preference, paired-difference, duo-trio and triangle tests. *J. food Sci.* **43**, 940–943 (1978).
22. R. Gross-Isseroff, D. Lancet, Concentration-dependent changes of perceived odor quality. *Chem. Senses* **13**, 191–204 (1988).
23. D. G. Laing, P. Legha, A. L. Jinks, I. Hutchinson, Relationship between molecular structure, concentration and odor qualities of oxygenated aliphatic molecules. *Chem. Senses* **28**, 57–69 (2003).
24. M. Conway, M. Oncul, K. Allen, Z. Zhang, J. Johnston, Perceptual constancy for an odor is acquired through changes in primary sensory neurons. *Sci. Adv.* **10**, eado9205 (2024).
25. R. Pellegrino *et al*., A quantitative framework for predicting odor intensity across molecule and mixtures. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.08.08.668954 (Accessed 12 August 2025).
26. Y. Li, X. Wei, Y. Zhou, J. Wang, R. You, Research progress of electronic nose technology in exhaled breath disease analysis. *Microsyst. Nanoeng.* **9**, 129 (2023).
27. B. Va, P. Mathew, S. Thomas, L. Mathew, Detection of lung cancer and stages via breath analysis using a self-made electronic nose device. *Expert. Rev. Mol. Diagn.* **24**, 341–353 (2024).
28. M. Sharma, S. Balaji, P. Saha, R. Kumar, Navigating the fragrance space using graph generative models and predicting odors. *J. Chem. Inf. Model.* **65**, 4818–4832 (2025).
29. M. Aleixandre, D. Prasetyawan, T. Nakamoto, Generative diffusion network for creating scents. *IEEE Access* **13**, 57311–57321 (2025).
30. B. C. Rodrigues, V. V. Santana, S. Murins, I. B. Nogueira, Molecule generation and optimization for efficient fragrance creation. *Ind. Eng. Chem. Res.* **63**, 14480–14494 (2024).
31. N. Dennler *et al*., High-speed odor sensing using miniaturized electronic nose. *Sci. Adv.* **10**, eadp1764 (2024).
32. C. Wang *et al*., Biomimetic olfactory chips based on large-scale monolithically integrated nanotube sensor arrays. *Nat. Electron.* **7**, 157–167 (2024).
33. F. Taleb, M. Vasco, A. Ribeiro, M. Björkman, D. Kragic, Can transformers smell like humans? *Adv. Neural Inf. Process. Syst.* **37**, 72032–72060 (2024).
34. D. M. Ennis, The power of sensory discrimination methods. *J. Sens. Stud.* **8**, 353–370 (1993).
35. V. Satarifard *et al*., Data from "DREAM-olfactory-mixtures-prediction-challenge." GitHub. https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge. Deposited 7 October 2024.
36. V. Satarifard *et al*., Satarifard/CWYK-Olfboost. GitHub. https://github.com/Satarifard/CWYK-Olfboost. Deposited 22 August 2024.
37. M. Lalis *et al*., chemosim-lab/ChemoDREAM. GitHub. https://github.com/chemosim-lab/ChemoDREAM. Deposited 14 November 2025.
38. Y. Han *et al*., YikunHan42/DREAM-Olfactory-Mixtures-Prediction-Challenge-CASI. GitHub. https://github.com/YikunHan42/DREAM-Olfactory-Mixtures-Prediction-Challenge-CASI. Deposited 1 August 2024.
39. V. Satarifard *et al*., Satarifard/DREAM-olfactory-mixtures-prediction-challenge. GitHub. https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge/tree/main/Models/PL21. Deposited 10 December 2025.
40. V. Satarifard *et al*., Satarifard/DREAM-olfactory-mixtures-prediction-challenge. GitHub. https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge/tree/main/Models/Tamarin. Deposited 10 December 2025.
41. R. D'hondt *et al*., Robbe D'hondt/DREAM 2024 Olfactory Challenge. GitLab. https://gitlab.kuleuven.be/u0131222/dream2024. Deposited 17 May 2024.