Supplementary Materials:

Materials and Methods

Figures S1-S7

Data File S1

Supplementary Materials:

Materials and Methods:

Perceptual Data: DREAM Challenge

The psychophysical data for this project were collected between February 2013 and July 2014 as part of the Rockefeller University Smell Study. Data from 49 individuals (28 women, median age 36) were used for the DREAM challenge. The dataset represents a subset of that presented in the original study (*14*), which was unpublished until the DREAM challenge was completed in early 2016. Six individuals declined permission to have their data used in the DREAM challenge. We excluded data on familiarity and edibility ratings for all stimuli, as well as data about whether the individual recognized the smell and how they described it in their own words, as well as data from 4 molecules [compound identification number (CID) 6202: thiamine hydrochloride; CID 24203: sodium phosphate dibasic; CID 2537: camphor; CID 106441: 2-methoxy-3(5 or 6)-isopropylpyrazine]. Of those subjects who agreed to provide information on race and ethnicity, 24 self-identified as Black, 14 as White, 5 as Asian, and 2 as Native American. Nine individuals self-identified as Hispanic. Individuals provided perceptual ratings of 992 stimuli, 476 different monomolecular chemicals at two different concentrations with 20 molecules tested twice.

Each molecule was presented to individuals at two different concentrations, diluted in paraffin oil so that the "high" and “low” concentrations for each molecule were empirically set to about equal intensity. While molecules were obtained at high purity (> 97%), we cannot exclude the possibility that trace contaminants or degradation products account for or add to the odor of the molecule. In the DREAM challenge, teams were asked for predictions of pleasantness and the 19 descriptors only for the “high” concentrations, and intensity predictions were made only for the subset of odors that were tested at a dilution of 1/1,000. Individuals were asked to rate each stimulus using 21 perceptual attributes (intensity, pleasantness, and 19 semantic descriptors), by moving an unlabeled slider. The default location of the slider was 50 for intensity and pleasantness, and 0 for the 19 descriptors. For each task, the final position of the slider was translated into a scale from 0 to 100, where 100 signified highest intensity or pleasantness, and the best match of a descriptor for a given stimulus. Further details on the psychophysical procedures and all raw data are available in the original study (*14*).

Perceptual Data: Out of Sample Analysis on New Subjects

Analysis in Fig. S7B used an unpublished study (Rockefeller University IRB Protocol LVO-0869) that tested intensity and pleasantness ratings of 403 subjects who sampled 47 molecules, comprising 32 that overlap with the DREAM Challenge study and 15 new molecules (Data File S1). Data were collected between June 2015 and October 2016 using the same methods as the DREAM Challenge dataset, with the exception that the stimuli were not intensity-matched for pleasantness prediction, and intensity predictions were performed on the available dilutions in the study, and not solely the 1/1000 dilution used in the DREAM Challenge. All subjects gave their written informed consent to participate in this study. 403 healthy subjects participated (246 women, median age 31). A subset of subjects provided self-identified information on their race and ethnicity: 152 Black, 131 White, 48 Asian. 81 self-identified as Hispanic.. Information on molecules and dilutions, and all raw psychophysical data are in Data

File S1.

Molecular features

We provided challenge participants with the CID for each molecule, useful for PubChem (https://pubchem.ncbi.nlm.nih.gov/) or other database searches. We used the Dragon software package (version 6; http://www.talete.mi.it) to generate a large number of chemical features for each molecule and made these available to participants.

Baseline model for splitting data for the challenge

We developed a linear model with a second layer cubic correction based on a PCA-reduced version of Dragon features to predict the perception of the population. The underlying methodology was used to solve the population prediction and is a multi-linear regression for each of the attributes based on the responses of all individuals and the molecular features of each molecule. The only pre-processing of the data we did was dimensionality reduction of the number of Dragon features, and a log transformation of the values. Based on the above, we chose a random partition that yields good predictive accuracy. We chose the partitions for the leaderboard set and hidden test set based on the distribution of median correlation over test molecules obtained with the model, for different random partitions. The median correlation across molecules selected for the selected partition is above 0.

Models

A graphical illustration of one of many decision trees generated by the random-forest algorithm as it evaluates how different structural and physical components determine "garlic" smell is shown in Fig. 2a. In each tree, the training data are sequentially partitioned such that each branch point helps increase the accuracy of a prediction. These trees are then aggregated, with their predictions averaged, through a process called bagging. Because the dimensionality of the structural data is high with 4884 Dragon features per molecule and the perception data matrix is sparse, random-forest models are well suited as they help reduce the dimension of the structural data by ignoring unimportant features, and help determine the decision boundary between perceptual ratings of zero and the more informative values. Because most perceptual attributes appeared to depend non-linearly on molecular features, and interactions between features may explain some of the perceptual experience, random-forest models—which can account for these complexities–performed best in this study. However, regularized linear models fared a close second for individual predictions (Data File S1). Linear models (Fig. 2b), which have previously been used to predict perceptual attributes (*8, 19*), served as a baseline model for the challenge. Their simplicity and good interpretability makes them appealing. Because the number of Dragon features far exceeds the number of molecules, simple linear models such as ordinary least squares regression will produce over-fitting and fail to generalize to untested molecules. Such models will also be sensitive to the highly non-normal distribution of the data and obviously fail to capture non-linear relationships between structural features and perceptual attributes. To overcome these problems, the best linear models used not only the original features, but also their squares (scaled between 0 and 1), and thus were quadratic in the original feature values. To reduce over-fitting, these models used randomized Lasso feature selection, so the summed magnitude of all the regression coefficients is minimized along with the mean-squared error; this automatically selects for models in which many coefficients are zero. Such models

were fit on resampled datasets to find the best-fitting and most informative features (Data File S1).

Scoring

The training set contained perceptual attribute data from 338 of the 476 molecules. The leaderboard set used for model validation and a hidden test set used for final predictions contained perceptual attribute data from 69 molecules each (Data File S1). Participants had access to the Dragon features for all 476 molecules. However, none of the challenge participants had access to the perceptual attribute data for the 69 molecules in the final hidden test set at any point during the challenge or the community phase. Scoring was handled by the organizers, including authors P.M. and R.N. Models were scored as follows: for individual prediction, the Pearson correlation between model and data, across test-set molecules, was computed for each individual and attribute. The mean correlations over individuals resulted in 21 attribute-level correlations. These were reduced to (1) the correlation for intensity, (2) the correlation for pleasantness, and (3) the mean of the correlations for the 19 semantic descriptors. These three items were normalized into Z-scores by using the mean and standard deviation for the same dataset with molecule identities shuffled. The final score is the mean of the three Z-scores. Population prediction was scored similarly except that the data were aggregated into means and standard deviations across individuals for each molecule and attribute. Models were asked to predict these means and standard deviations. Here six Z-scores were used, with three corresponding to the means and three to the standard deviations. In both cases we re-scored the models in 1000 bootstraps of the hidden test set.

For individual prediction, the best-performing model remained first in 80 per cent of the bootstrap runs, whereas the second model ranked first in 8 per cent of the runs. For population prediction, the best-performing model remained first in 38 per cent of the bootstrap runs, whereas the second model ranked first in 26 per cent of the runs.

Significance of the correlation between subjects' usage of the scale and correlation of predictions

We computed significance by shuffling subjects' identity 1000 times, such that variance and prediction accuracy were uncorrelated across subjects in expectation, and then computed the 97.5% percentile to obtain the threshold for a two-tailed $p < 0.05$, which is shown in Fig. 2f. Technically each descriptor should have a different $p < 0.05$ cutoff, but these were sufficiently similar across descriptors that we simply reported the one obtained by pooling the shuffles across descriptors.

Significance for the correlation between connectivity structure and model performance for "garlic"/ "fish", "sweet" / "fruit" and "musky" / "sweaty".

The p-value was calculated by randomizing 10,000 times the attributes' identities across both the connectivity and correlation axis and counting the proportion of cases were the connectivity strength had a value above or equal the value for "garlic/fish" and the correlation distance was below. The same procedure was used for the 2 other pairs of attributes.

Aggregation of models

Participant models were aggregated by first ranking by descending Z-score, then averaging one-by-one following these ranks (the 2 highest ranked models, the 3 highest ranked models, etc.) until all models were aggregated to obtain the same number of aggregations as models.

Post-challenge community phase

Five teams (Teams IKW Allstars, GuanLab, KU Leuven, Russ Wolfinger, and Joel Mainland) participated in this phase of the challenge where we discussed ways to enhance the predictions. Each team submitted one new model for both individual and population predictions based on these discussions, which was scored against the same test-set as during the open phase of the challenge. An aggregate model built from these five models was also scored (Fig. 1h).

Assessing the reverse-engineering of perceptual profiles using the aggregate model

One way to assess the sensitivity of the model's sensory profile predictions is to calculate the probability of having exactly k correct sensory profile predictions from a list of *n* molecules, that is: $p_k^n = \binom{n}{k} q^{n-k}(1-q)^k$ where $q = \frac{(n-1)}{n}$ is the probability of matching incorrectly one profile to the list of n molecules.

Here *n*=69 and the aggregated model was able to reverse-engineer k=14 sensory profiles perfectly (20%), so $p_{14}^{69} = 1.2492 * 10^{-12}$.

Another way to measure the performance is to measure the area under the model prediction rank curve (AUC) of Fig. 4f. For a perfect model, the prediction rank for every molecule is 1 and so the AUC is the entire plot area: 69*69 (normalized to 1); for a random model all ranks are equally likely and 5f would show a diagonal line (in expectation), with area $\sum_1^{69} i$.=69*68/2 normalized to an AUC of 0.5. For our model presented here has an AUC equal to the perfect model area minus the sum of the ranks of the aggregate model i.e 69*69-830 (normalized to 0.826).

Out of sample analysis

We employed two methods to build and test our predictions on different subjects. First, we trained a random forest model on 25 subjects split arbitrarily from the DREAM Challenge dataset and tested it on the remaining 24 subjects. This process was repeated 50 times to yield the predictions in Fig. S7A. Second, we used a new unpublished dataset from 403 subjects to test intensity and pleasantness predictions for 47 molecules (15 not previously tested in the DREAM Challenge) using a random forest model trained on all the molecules of the 49 DREAM Challenge subjects (Fig. S7B).

Additional Author notes:

DREAM Olfaction challenge consortium:
Agnieszka Kitlas Golińska[15], Aleksandar Dimitriev[16], Amol P. Bhondekar[17], Andrej Dolenc[16], Andrew Matteson[18,19], Aneta Polewko-Klim[15], Barbara F. Huang[20], Bharat Panwar[3], Blaz Zupan[16], Bor-Wei Cherng[21], Chien-Yu Chen[21,22], Delia Yao[20], G.P.S Raghava[23], Jose M.G. Vilar[24,25], László Hunyady[5,6], Leonor Saiz[26],Marat D. Kazanov[27], Marinka Zitnik[16],Marko Toplak[16], Michael Xie[20], Ming-Yi Hong[22], Nancy Yu[20], Paul C. Boutros[20,28,29], Peter Us[16], Péter Várnai[5], Ping-Han Hsieh[30], Radosław Piliszek[15], Ren X. Sun[20,28], Rishemjit Kaur[17], Ritesh Kumar[17], Witold R. Rudnicki[15,31], Wojciech Lesiński[15], Xihui Lin[20], Yen-Jen Oyang[30], Yi-An Tung[22], Yu-Chuan Chang[30].

[15] Institute of Informatics, University of Bialystok, 15-245 Białystok, Poland.

[16] Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia.
[17]CSIR-Central Scientific Instruments Organisation, Chandigarh, India Academy of Scientific and Innovative Research, New Delhi, India.
[18] The Mathworks, Natick, MA 01760 USA.
[19] Applied BioMath, LLC, Lincoln, MA 01773 USA.
[20]Informatics & Biocomputing Program, Ontario Institute for Cancer Research, Toronto, M5G 0A3 Canada.
[21]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei 106, Taiwan.
[22]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 106, Taiwan.
[23]CSIR-Institute of Microbial Technology Chandigarh, India, Academy of Scientific and Innovative Research, New Delhi, India.
[24]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain.
[25] Biofisika Institute (CSIC,UPV/EHU), Department of Biochemistry and Molecular Biology, University of the Basque Country,48080 Bilbao, Spain.
[26]Department of Biomedical Engineering, University of California, Davis, California 95616 USA.
[27]Research and Training Center on Bioinformatics, Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.
[28]Department of Pharmacology & Toxicology, University of Toronto, Toronto, Canada.
[29]Department of Medical Biophysics, University of Toronto, Toronto, Canada.
[30]Graduate Institute of Biomedical Electronic and Bioinformatics, National Taiwan University, Taipei 106, Taiwan.
[31]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland.

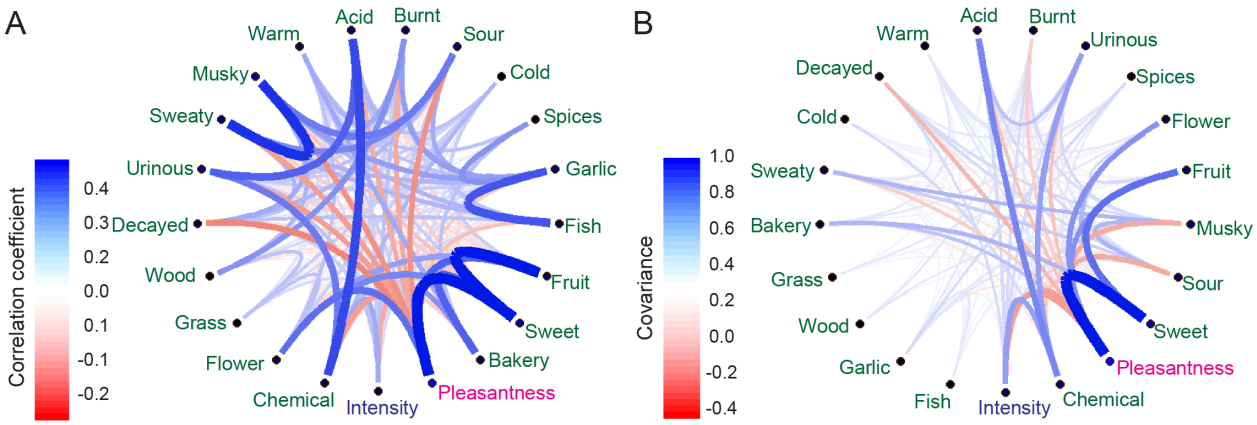

Fig. S1. Correlation and covariance of perceptual attributes. (**A-B**) Line width and color represent the strength of the pairwise correlation (**A**) and normalized covariance (**B**) between 21 attributes for all molecules and individuals.

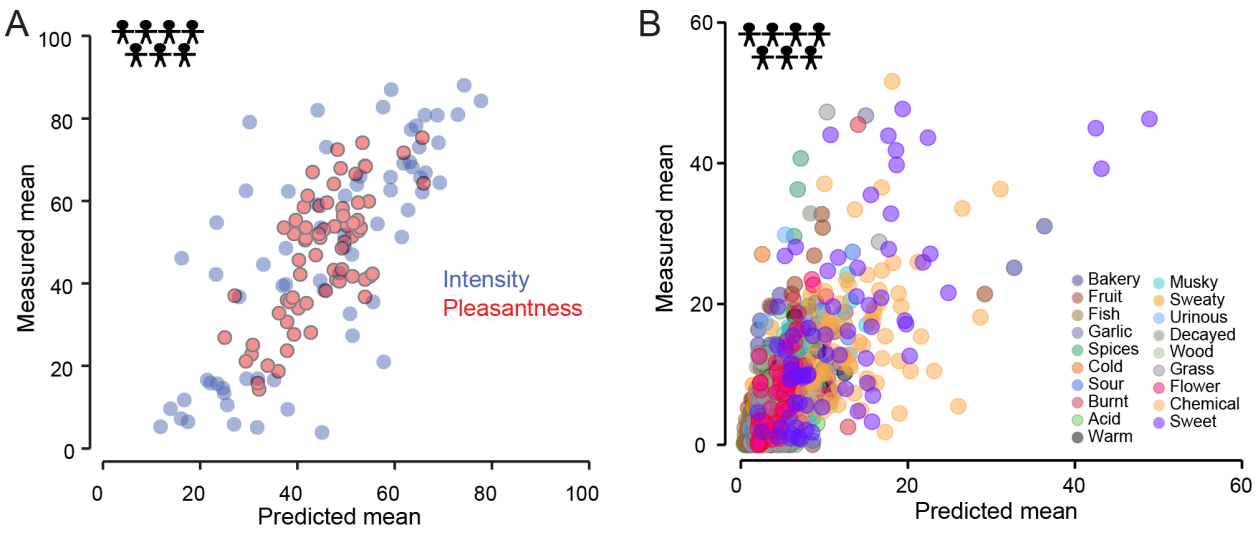

Fig. S2. Best performer outcomes for mean and standard deviation for population prediction. (**A-B**) Intensity and pleasantness (**A**) and 19 descriptor (**B**) predictions of the mean of the best-performing team plotted against the observed values for the 69 hidden test set molecules used for model validation.

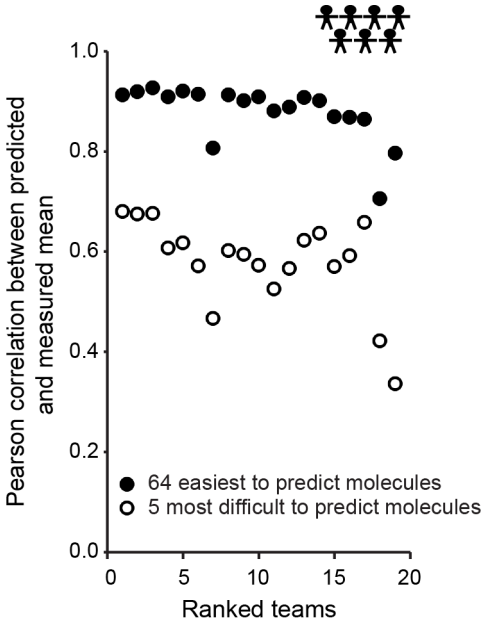

Fig. S3. Prediction performance. Pearson correlation between predicted and measured mean perception of the 64 molecules that were the easiest (black dots) and the five molecules that were the most difficult (white dots) to predict. Teams are ordered by their final score for population prediction, with the best performer ranked 1.

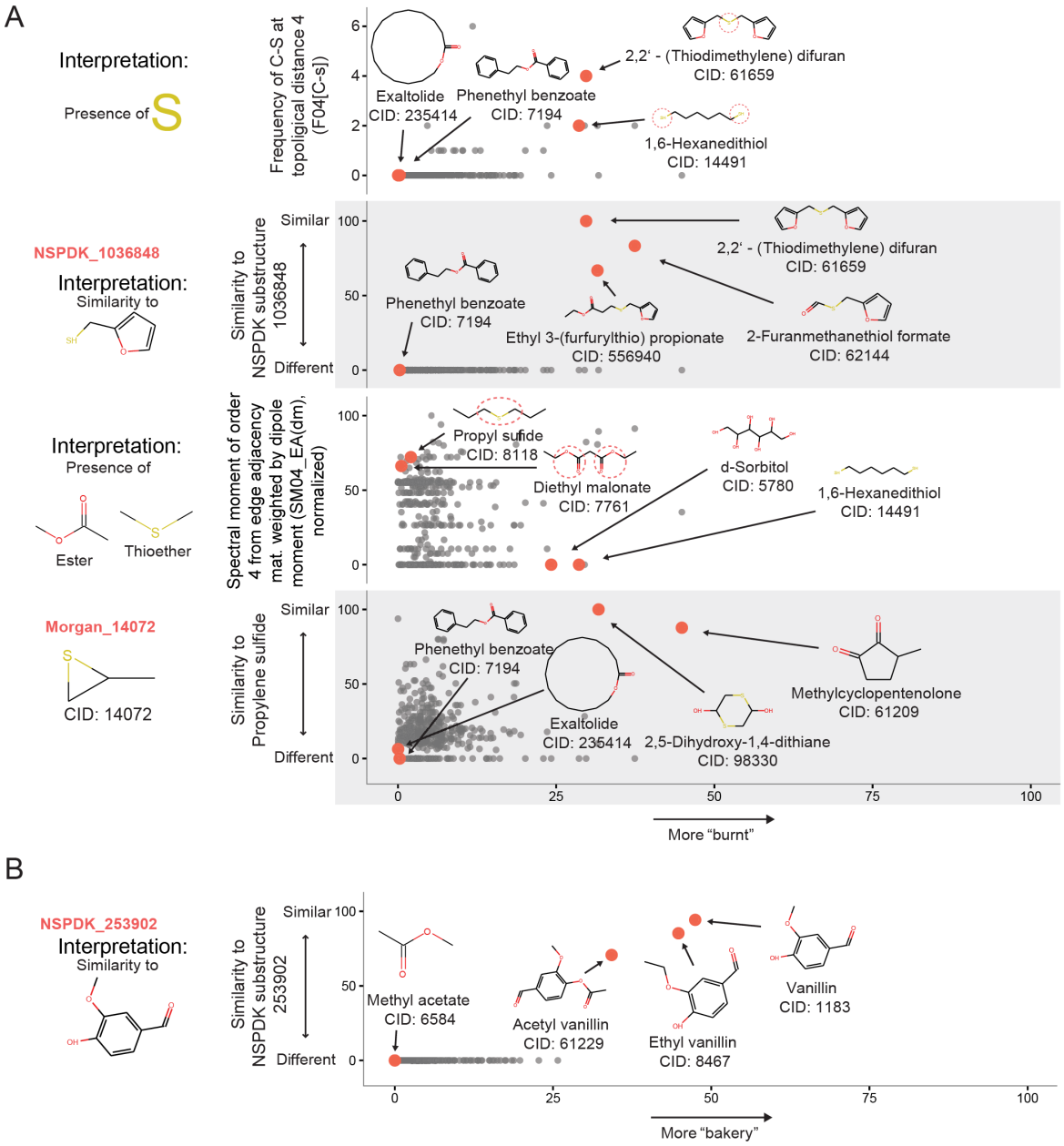

Fig. S4. Top molecular features used by the random-forest model from the post-challenge phase as predictors for "burnt" and "bakery." (**A-B**) Each grey dot represents predictions of each of the 407 molecules in the training+leaderboard set for "burnt" (**A**) and "bakery" (**B**), with example molecules indicated by red dots. In (**A**) only four of the five top features are shown. The fifth feature (R3p+; R maximal autocorrelation of lag 3 / weighted by polarizability) is very similar to the feature depicted in the top panel. In (**B**) only the top feature is shown. The four other features in the top 5 (NSPDK_1022278, NSPDK_722140, NSPDK_250366, NSPDK_555472) are very similar to the one shown.

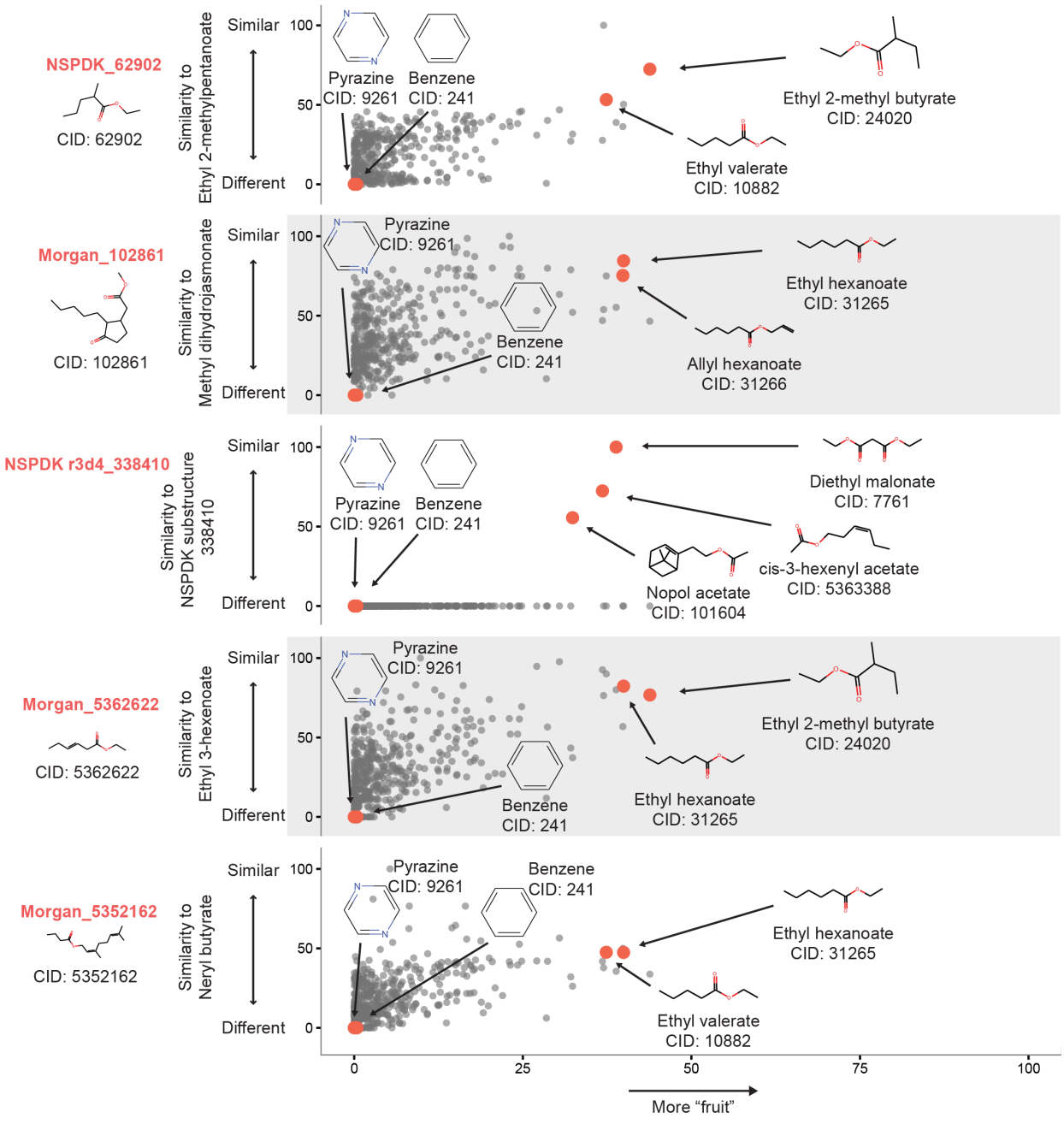

Fig. S5. Top 5 molecular features used by the random-forest model from the post-challenge phase as predictors for “fruit”. Each grey dot represents predictions of each of the 407 molecules in the training+leaderboard set for “fruit”, with example molecules indicated by red dots.

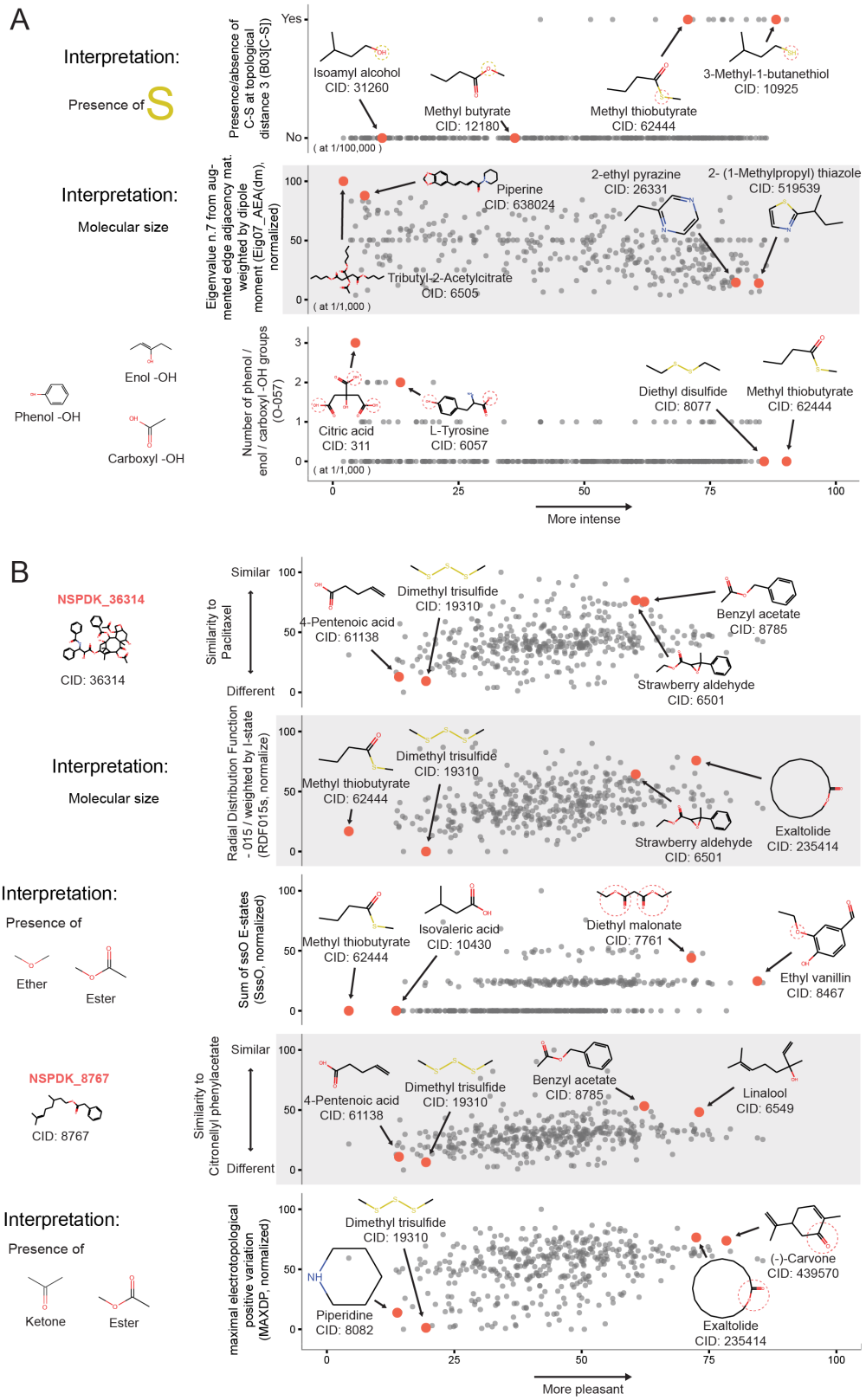
A
Interpretation:
Presence of S
Presence/absence of C-S at topological distance 3 (B03[C-S])
Yes
No
(at 1/100,000)
Isoamyl alcohol
CID: 31260
Methyl butyrate
CID: 12180
Methyl thiobutyrate
CID: 62444
3-Methyl-1-butanethiol
CID: 10925
Interpretation:
Molecular size
Eigenvalue n.7 from augmented edge adjacency mat. weighted by dipole moment (Eig07_AEA(dm), normalized)
(at 1/1,000)
Piperine
CID: 638024
2-ethyl pyrazine
CID: 26331
2- (1-Methylpropyl) thiazole
CID: 519539
Tributyl-2-Acetylcitrate
CID: 6505
Phenol -OH
Enol -OH
Carboxyl -OH
Number of phenol / enol / carboxyl -OH groups (O-057)
(at 1/1,000)
Citric acid
CID: 311
L-Tyrosine
CID: 6057
Diethyl disulfide
CID: 8077
Methyl thiobutyrate
CID: 62444
More intense
B
NSPDK_36314
CID: 36314
Similarity to Paclitaxel
Similar
Different
4-Pentenoic acid
CID: 61138
Dimethyl trisulfide
CID: 19310
Benzyl acetate
CID: 8785
Strawberry aldehyde
CID: 6501
Interpretation:
Molecular size
Radial Distribution Function - 015 / weighted by I-state (RDF015s, normalize)
Methyl thiobutyrate
CID: 62444
Dimethyl trisulfide
CID: 19310
Strawberry aldehyde
CID: 6501
Exaltolide
CID: 235414
Interpretation:
Presence of
Ether
Ester
Sum of ssO E-states (SssO, normalized)
Methyl thiobutyrate
CID: 62444
Isovaleric acid
CID: 10430
Diethyl malonate
CID: 7761
Ethyl vanillin
CID: 8467
NSPDK_8767
CID: 8767
Similarity to Citronellyl phenylacetate
Similar
Different
4-Pentenoic acid
CID: 61138
Dimethyl trisulfide
CID: 19310
Benzyl acetate
CID: 8785
Linalool
CID: 6549
Interpretation:
Presence of
Ketone
Ester
maximal electrotopological positive variation (MAXDP, normalized)
Dimethyl trisulfide
CID: 19310
Piperidine
CID: 8082
(-)-Carvone
CID: 439570
Exaltolide
CID: 235414
More pleasant

Fig. S6. Predicting the smell of specific molecules. (**A-B**), The five most important molecular features selected from Dragon, Morgan, and NSPDK (red text) for predicting (**A**) intensity and (**B**) pleasantness using the random-forest model from the post-challenge phase. Each grey dot represents one of the 407 molecules in the training+leaderboard set, with example molecules indicated by red dots. For (**A**), only three features are shown. The other two are very similar to the one shown in the top panel (B03[C-S]) and the one shown in the middle panel (Eig07_AEA(dm)), respectively.

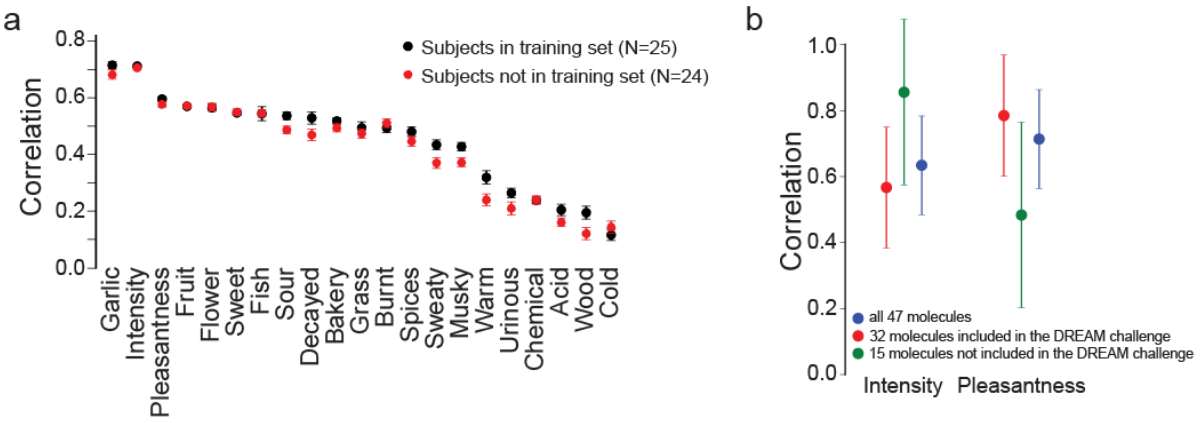

Fig. S7. Out of sample predictions. (**A**) Average of correlation of prediction using a random forest model trained with 25 subjects and tested on the remaining 24 subjects selected randomly 50 times with the odors of the training set. For each attribute, the correlation between the model predictions for the 69 hidden test set odors was calculated against the 25 subjects used for training (black dots) or the 24 subjects left out of sample (red dots). (**B**) Average of correlation of prediction for intensity and pleasantness trained on the DREAM challenge dataset but tested on a new population of 403 subjects. Predictions are correlated against perception of 32 molecules present both in the original DREAM challenge and the new study (red), 15 new molecules (green), or all 47 molecules used in the new study (blue). Error bars were generated from the Fisher transformation.

Data File S1. Raw data including prediction scores and methods, correlation values, molecule CIDs, and top molecular features.