Cite as: A. Keller *et al.*, *Science* 10.1126/science.aal2014 (2017).

Predicting human olfactory perception from chemical features of odor molecules

Andreas Keller,[1]* Richard C. Gerkin,[2]* Yuanfang Guan,[3]* Amit Dhurandhar,[4] Gabor Turu,[5,6] Bence Szalai,[5,6] Joel D. Mainland,[7,8] Yusuke Ihara,[7,9] Chung Wen Yu,[7] Russ Wolfinger,[10] Celine Vens,[11] Leander Schietgat,[12] Kurt De Grave,[12,13] Raquel Norel,[4] DREAM Olfaction Prediction Consortium,† Gustavo Stolovitzky,[4,15] Guillermo A. Cecchi,[4] Leslie B. Vosshall,[1,14] Pablo Meyer[4,15]‡

[1]Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, NY 10065, USA. [2]School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA. [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [4]Thomas J. Watson Computational Biology Center, IBM, Yorktown Heights, NY 10598, USA. [5]Department of Physiology, Faculty of Medicine, Semmelweis University, 1085 Budapest, Hungary. [6]Laboratory of Molecular Physiology, Hungarian Academy of Science, Semmelweis University (MTA-SE), 1085 Budapest, Hungary. [7]Monell Chemical Senses Center, Philadelphia, PA 19104, USA. [8]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104, USA. [9]Institution for Innovation, Ajinomoto Co., Inc., Kawasaki, Kanagawa 210-8681, Japan. [10]SAS Institute, Inc., Cary, NC 27513, USA. [11]Department of Public Health and Primary Care, KU Leuven, Kulak, 8500 Kortrijk, Belgium. [12]Department of Computer Science, KU Leuven, 3001 Leuven, Belgium. [13]Flanders Make, 3920 Lommel, Belgium. [14]Howard Hughes Medical Institute, New York, NY 10065, USA. [15]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

*These authors contributed equally to this work.

†DREAM Olfaction Prediction Consortium authors and affiliations are listed in the supplementary materials.

‡Corresponding author. E-mail: pmeyerr@us.ibm.com

It is still not possible to predict whether a given molecule will have a perceived odor or what olfactory percept it will produce. We therefore organized the crowd-sourced DREAM Olfaction Prediction Challenge. Using a large olfactory psychophysical data set, teams developed machine-learning algorithms to predict sensory attributes of molecules based on their chemoinformatic features. The resulting models accurately predicted odor intensity and pleasantness and also successfully predicted 8 among 19 rated semantic descriptors ("garlic," "fish," "sweet," "fruit," "burnt," "spices," "flower," and "sour"). Regularized linear models performed nearly as well as random forest–based ones, with a predictive accuracy that closely approaches a key theoretical limit. These models help to predict the perceptual qualities of virtually any molecule with high accuracy and also reverse-engineer the smell of a molecule.

In vision and hearing, the wavelength of light and frequency of sound are highly predictive of color and tone. In contrast, it is not currently possible to predict the smell of a molecule from its chemical structure (*1*, *2*). This stimulus-percept problem has been difficult to solve in olfaction because odors do not vary continuously in stimulus space, and the size and dimensionality of olfactory perceptual space is unknown (*1*, *3*, *4*). Some molecules with very similar chemical structures can be discriminated by humans (*5*, *6*), and molecules with very different structures sometimes produce nearly identical percepts (*2*). Computational efforts developed models to relate chemical structure to odor percept (*2*, *7*–*11*), but many relied on psychophysical data from a single 30-year-old study that used odorants with limited structural and perceptual diversity (*12*, *13*).

Twenty-two teams were given a large, unpublished psychophysical data set collected by Keller and Vosshall from 49 individuals who profiled 476 structurally and perceptually diverse molecules (*14*) (Fig. 1A). We supplied 4884 physicochemical features of each of the molecules smelled by the subjects, including atom types, functional groups, and topological and geometrical properties that were computed using Dragon chemoinformatic software (version 6, Talete S.r.l., see supplementary materials) (Fig. 1B).

Using a baseline linear model developed for the challenge and inspired by previous efforts to model perceptual responses of humans (*8*, *11*), we divided the perceptual data into three sets. Challenge participants were provided with a training set of perceptual data from 338 molecules that they used to build models (Fig. 1C). The organizers used perceptual data from an additional 69 molecules to build a leaderboard to rank performance of participants during the competition. Toward the end of the challenge, the organizers released perceptual data from the 69 leaderboard molecules so that participants could get feedback on their model and to enable refinement with a larger training + leaderboard data set. The remaining 69 molecules were kept as a hidden test set available only to challenge organizers to evaluate the performance of the final models (Fig. 1C). Participants developed models to predict the perceived intensity,

pleasantness, and usage of 19 semantic descriptors for each of the 49 individuals and for the mean and standard deviation across the population of these individuals (Fig. 1, D and E).

We first examined the structure of the psychophysical data using the inverse of the covariance matrix (*15*) calculated across all molecules as a proxy for connection strength between each of the 21 perceptual attributes (Fig. 1F and fig. S1). This yielded a number of strong positive interactions, including those between "garlic" and "fish"; "musky" and "sweaty"; and "sweet" and "bakery"; and among "fruit," "acid," and "urinous"; and a negative interaction between pleasantness and "decayed" (Fig. 1F and fig. S1A). The perception of intensity had the lowest connectivity to the other 20 attributes. To understand whether a given individual used the full rating scale or a restricted range, we examined subject-level variance across the ratings for all molecules (Fig. 1G). Applying hierarchical clustering on Euclidean distances for the variance of attribute ratings across all the molecules in the data set, we distinguished three clusters: subjects that responded with high-variance for all 21 attributes (left cluster in green), subjects with high-variance for four attributes (intensity, pleasantness, "chemical," and "sweet") and either low variance (middle cluster in blue) or intermediate variance (right cluster in red) for the remaining 17 attributes (Fig. 1G).

We assessed the performance of models submitted to the DREAM Challenge by computing for each attribute the correlation between the predictions of the 69 hidden test molecules and the actual data. We then calculated a *Z*-score by subtracting the average correlations and scaling by the standard deviation of a distribution based on a randomization of the test-set molecule identities. Of the 18 teams who submitted models to predict individual perception, Team GuanLab (author Y.G.) was the best performer with a *Z*-score of 34.18 (Fig. 1H and table S1). Team IKW Allstars (author R.C.G.) was the best performer of 19 teams to submit models to predict population perception, with a *Z*-score of 8.87 (Fig. 1H and table S1). The aggregation of all participant models gave *Z*-scores of 34.02 (individual) and 9.17 (population) (Fig. 1H), and a postchallenge community phase where initial models and additional molecular features were shared across teams gave even better models with *Z*-scores of 36.45 (individual) and 9.92 (population) (Fig. 1H).

Predictions of the models for intensity were highly correlated with the observed data for both individuals ($r = 0.56$; *t* test, $P < 10^{-228}$) and the population ($r = 0.78$; $P < 10^{-9}$) (Fig. 1, I and J). Pleasantness was also well predicted for individuals ($r = 0.41$; $P < 10^{-123}$) and the population ($r = 0.71$; $P < 10^{-8}$) (Fig. 1, I and J). The 19 semantic descriptors were more difficult to predict, but the best models performed respectably (individual: $r = 0.21$; $P < 10^{-33}$; population: $r = 0.55$; $P < 10^{-5}$) (Fig. 1, I and J). Previously described models to predict pleasantness (*8*, *10*) performed less well on this data set than our best model (Fig. 1J). To our knowledge, there are no existing models to predict the 19 semantic descriptors.

Random-forest (Fig. 2A and table S1) and regularized linear models (Fig. 2B and table S1) outperformed other common predictive model types for the prediction of individual and population perception (Fig. 2, fig. S2, and table S1). Although the quality of the best-performing model varied greatly across attributes, it was exceptionally high in some cases (Fig. 2C), and always considerably higher than chance (dotted line in Fig. 1I), while tracking the observed perceptual values (fig. S2 for population prediction). In contrast to most previous studies that attempted to predict olfactory perception, these results all reflect predictions of a hidden test set and avoid the pitfall of inflated correlations due to overfitting of the experimental data.

The accuracy of predictions of individual perception for the best-performing model was highly variable (Fig. 2C), but the correlation of six of the attributes was above 0.3 (white circles in Fig. 2D). The best-predicted individual showed a correlation above 0.5 for 16 of 21 attributes (Fig. 2D). We asked whether the usage of the rating scale (Fig. 1G) could be related to the predictability of each individual. Overall, we observed that individuals using a narrow range of attribute ratings—measured across all molecules for a given attribute—were more difficult to predict (Fig. 2, E and F, derived from the variance in Fig. 1G). The relations between range and prediction accuracy did not hold for intensity and pleasantness (Fig. 2, E and F).

We next compared the results of predicting individual and population perception. The seven best-predicted attributes overall (intensity, "garlic," pleasantness, "sweet," "fruit," "spices," and "burnt") were the same for both individuals and the population (Figs. 2D and 3A except "fish"). Similarly, the seven attributes that were the most difficult to predict ("acid," "cold," "warm," "wood," "urinous," "chemical," and "musky") were the same for both individual and population predictions (Fig. 2D and Fig. 3A), and except for a low correlation for "warm," these attributes are anticorrelated or uncorrelated to the "familiarity" attribute (*14*). This suggests some bias in the predictability of more familiar attributes, perhaps due to a better match to a well-defined reference molecule (*14*), and that, in this categorization, individual perceptions are similar across the population. For the population predictions, the first 10 attributes have a correlation above 0.5 (Fig. 3A). The connectivity structure in Fig. 1F follows the model's performance for the population (Fig. 3A). "Garlic"-"fish" ($P < 10^{-4}$), "sweet"-"fruit" ($P < 10^{-3}$), and "musky"-"sweaty" ($P < 10^{-3}$) are pairs with strong connectivity that were also similarly difficult to predict.

We analyzed the quality of model predictions for specific molecules in the population (Fig. 3B). The correlation between predicted and observed attributes exceeded 0.9 (*t* test, $P < 10^{-4}$) for 44 of 69 hidden test-set molecules when we used aggregated model predictions, and 28 of 69 when we averaged all model correlations (table S1). The quality of predictions varied across molecules, but for every molecule, the aggregated models exhibited higher correlations (Fig. 3B). The two best-predicted molecules were 3-methyl cyclohexanone followed by ethyl heptanoate. Conversely, the five molecules that were most difficult to predict were L-lysine and L-cysteine, followed by ethyl formate, benzyl ether, and glycerol (Fig. 3B and fig. S3).

To better understand how the models successfully predicted the different perceptual attributes, we first asked how many molecular features were needed to predict a given population attribute. Although some attributes required hundreds of features to be optimally predicted (Fig. 3, C to E), both the random-forest and linear models achieved prediction quality of at least 80% of that optimum with far fewer features. By that measure, the algorithm to predict intensity was the most complex, requiring 15 molecular features to reach the 80% threshold (Fig. 3C). "Fish" was the simplest, requiring only one (Fig. 3D). Although Dragon features are highly correlated, these results are remarkable because even those attributes needing the most molecular features to be predicted required only a small fraction of the thousands of chemoinformatic features.

We asked what features are most important for predicting a given attribute (figs. S4 to S6 and table S1). The Dragon software calculates a large number of molecular features but is not exhaustive. In a postchallenge phase (triangles in Fig. 1H), four of the original teams attempted to improve their model predictions by using additional features. These included Morgan (*16*) and neighborhood subgraph pairwise distance kernel (NSPDK) (*17*), which encode features through the presence or absence of particular substructures in the molecule; experimentally derived partition coefficients from EPI Suite (*18*); and the common names of the molecules. We used cross-validation on the whole data set to compare the performance of the same models using different subsets of Dragon and these additional molecular features. Only Dragon features combined with Morgan features yielded decisively better results than Dragon features alone, both for random-forest (Fig. 4A) and linear (Fig. 4B) models. We then examined how the random-forest model weighted each feature (table S1 for a similar analysis using the linear model). As observed previously, intensity was negatively correlated with molecular size but was positively correlated with the presence of polar groups, such as phenol, enol, and carboxyl features (fig. S6A) (*1*, *7*). Predictions of intensity relied primarily on Dragon features.

There is already anecdotal evidence that some chemical features are associated with a sensory attribute. For example, sulfurous molecules are known to smell "garlic" or "burnt," but no quantitative model exists to confirm this. Our model confirms that the presence of sulfur in the Dragon descriptors used by the model correlated positively with both "burnt" ($r = 0.661$; $P < 10^{-62}$) (fig. S4A) and "garlic" ($r = 0.413$; $P < 10^{-22}$; table S1). Pleasantness was predicted most accurately using a mix of both Dragon and Morgan-NSPDK features. For example, pleasantness correlated with both molecular size ($r = 0.160$; $P < 10^{-3}$) (*9*) and similarity to paclitaxel ($r = 0.184$; $P < 10^{-4}$) and citronellyl phenylacetate ($r = 0.178$; $P < 10^{-4}$) (fig. S6B). "Bakery" predictions were driven by similarity to the molecule vanillin ($r = 0.45$; $P < 10^{-24}$) (fig. S4B). Morgan features improved prediction in part by enabling a model to template-match target molecules against reference molecules for which the training set contains perceptual data. Thus, structural similarity to vanillin or ethyl vanillin predicts "bakery" without recourse to structural features.

Twenty of the molecules in the training set were rated twice ("test" and "retest") by each individual, providing an estimate of within-individual variability for the same stimulus. This within-individual variability places an upper limit on the expected accuracy of the optimal predictive model. We calculated the test-retest correlation across individuals and molecules for each perceptual attribute. This value of the observed correlation provides an upper limit to any model, because no model prediction should produce a better correlation than data from an independent trial with an identical stimulus and individual. To examine the performance of our model compared with the theoretically best model, we calculated a correlation coefficient between the prediction of a top-performing random-forest model and the test data. All attributes except "burnt" were statistically indistinguishable from the test-retest correlation coefficients evaluated at the individual level (Fig. 4C). The slope for the best linear fit of the test-retest and model-test correlation coefficients was 0.80 ± 0.02, with a slope of 1 expected for optimal performance (Fig. 4C). Similar results were obtained using a model-retest correlation. Thus, given this data set, performance of the model is close to that of the theoretically optimal model.

We evaluated the specificity of the predictions of the aggregated model by calculating how frequently the predicted sensory profile had a better correlation with the actual sensory profile of the target molecule than it did with the sensory profiles of any of the other 68 molecules in the hidden test-set (Fig. 4, D and E). For 14 of 69 molecules, the highest correlation coincided with the actual sensory profile ($P < 10^{-11}$). For an additional 20%, it was second highest, and 65% of the molecules ranked in the top-ten predictions [Fig. 4F

and table S1; area under the curve (AUC) = 0.83]. The specificity of the aggregated model shows that its predictions could be used to reverse-engineer a desired sensory profile by using a combination of molecular features to synthesize a designed molecule.

Finally, to ensure that the performance of our model would extend to new subjects, we trained it on random subsets of 25 subjects from the DREAM data set and consistently predicted the attribute ratings of the mean across the population of the 24 left-out subjects (fig. S7A). To test our model across new subjects and new molecules, we took advantage of a large unpublished data set of 403 volunteers who rated the intensity and pleasantness of 47 molecules, of which only 32 overlapped with the stimuli used in the original study (table S1). Using a random-forest model trained on the original 49 DREAM Challenge subjects and all the molecules, we are able to show that the model robustly predicts the average perception of all of these molecules across the population (fig. S7B).

The DREAM Olfaction Prediction Challenge has yielded models that generated high-quality personalized perceptual predictions. This work substantially expands on previous modeling efforts (*2*, *3*, *7*–*11*) because it predicts not only pleasantness and intensity, but also 8 out of 19 semantic descriptors of odor quality. The predictive models enable the reverse-engineering of a desired perceptual profile to identify suitable molecules from vast databases of chemical structures and closely approach the theoretical limits of accuracy when accounting for within-individual variability. Although highly significant, there is still much room for improving in particular the individual predictions. Although the current models can only be used to predict the 21 attributes, the same approach could be applied to a psychophysical data set that measured any desired sensory attribute (e.g., "rose," "sandalwood," or "citrus"). How can the highly predictive models presented here be further improved? Recognizing the inherent limits of using semantic descriptors for odors (*12*–*14*), we think that alternative perceptual data, such as ratings of stimulus similarity, will be important (*11*).

What do our results imply about how the brain encodes an olfactory percept? We speculate that, for each molecular feature, there must be some quantitative mapping, possibly one to many, between the magnitude of that feature and the spatiotemporal pattern and activation magnitude of the associated olfactory receptors. If features rarely or never interact to produce perception, as suggested by the strong relative performance of linear models in this challenge, then these feature-specific patterns must sum linearly at the perceptual stage (*19*). Peripheral events in the olfactory sensory epithelium, including receptor binding and sensory neuron firing rates might have nonlinearities, but the numerical representation of perceptual magnitude must be linear in these patterns. It is possible that stronger nonlinearity will be discovered when odor mixtures or the temporal dynamics of odor perception are investigated. Many questions regarding human olfaction remain that may be successfully addressed by applying this method to future data sets that include more specific descriptors; more molecules that represent different olfactory percepts than those studied here; and subjects of different genetic, cultural, and geographic backgrounds.

Results of the DREAM Olfaction Prediction Challenge may accelerate efforts to understand basic mechanisms of ligand-receptor interactions, and to test predictive models of olfactory coding in both humans and animal models. Finally, these models have the potential to streamline the production and evaluation of new molecules by the flavor and fragrance industry.

References and Notes

1. H. Boelens, Structure-activity relationships in chemoreception by human olfaction. *Trends Pharmacol. Sci.* **4**, 421–426 (1983). doi:10.1016/0165-6147(83)90475-3
2. C. Sell, Structure-odor relationships: On the unpredictability of odor. *Angew. Chem. Int. Ed.* **45**, 6254–6261 (2006). doi:10.1002/anie.200600782
3. A. A. Koulakov, B. E. Kolterman, A. G. Enikolopov, D. Rinberg, In search of the structure of human olfactory space. *Front. Syst. Neurosci.* **5**, 65 (2011). doi:10.3389/fnsys.2011.00065 Medline
4. J. B. Castro, A. Ramanathan, C. S. Chennubhotla, Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLOS ONE* **8**, e73289 (2013). doi:10.1371/journal.pone.0073289 Medline
5. M. Laska, P. Teubner, Olfactory discrimination ability for homologous series of aliphatic alcohols and aldehydes. *Chem. Senses* **24**, 263–270 (1999). doi:10.1093/chemse/24.3.263 Medline
6. S. Boesveldt, M. J. Olsson, J. N. Lundström, Carbon chain length and the stimulus problem in olfaction. *Behav. Brain Res.* **215**, 110–113 (2010). doi:10.1016/j.bbr.2010.07.007 Medline
7. P. A. Edwards, P. C. Jurs, Correlation of odor intensities with structural properties of odorants. *Chem. Senses* **14**, 281–291 (1989). doi:10.1093/chemse/14.2.281
8. R. M. Khan, C. H. Luk, A. Flinker, A. Aggarwal, H. Lapid, R. Haddad, N. Sobel, Predicting odor pleasantness from odorant structure: Pleasantness as a reflection of the physical world. *J. Neurosci.* **27**, 10015–10023 (2007). doi:10.1523/JNEUROSCI.1158-07.2007 Medline
9. F. Kermen, A. Chakirian, C. Sezille, P. Joussain, G. Le Goff, A. Ziessel, M. Chastrette, N. Mandairon, A. Didier, C. Rouby, M. Bensafi, Molecular complexity determines the number of olfactory notes and the pleasantness of smells. *Sci. Rep.* **1**, 206 (2011). Medline
10. M. Zarzo, Hedonic judgments of chemical compounds are correlated with molecular size. *Sensors (Basel)* **11**, 3667–3686 (2011). doi:10.3390/s110403667 Medline
11. 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
12. A. Dravnieks, Odor quality: Semantically generated multidimensional profiles are stable. *Science* **218**, 799–801 (1982). doi:10.1126/science.7134974 Medline
13. A. Dravnieks, *Atlas of Odor Character Profiles* (ASTM, Philadelphia, 1985).
14. A. Keller, L. B. Vosshall, Olfactory perception of chemically diverse molecules. *BMC Neurosci.* **17**, 55 (2016). doi:10.1186/s12868-016-0287-2 Medline
15. R. J. Prill, R. Vogel, G. A. Cecchi, G. Altan-Bonnet, G. Stolovitzky, Noise-driven causal inference in biomolecular networks. *PLOS ONE* **10**, e0125777 (2015). doi:10.1371/journal.pone.0125777 Medline
16. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010). doi:10.1021/ci100050t Medline
17. F. Costa, K. De Grave, in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Quebec, Canada, 14 to 18 June 2009 (Association

for Computing Machinery, New York, 2010), pp. 255–262.
18. U.S. Environmental Protection Agency, Estimation Programs Interface Suite for Microsoft Windows (EPI Suite), v 4.11 (EPA, Washington, DC, 2014).
19. P. Gupta, D. F. Albeanu, U. S. Bhalla, Olfactory bulb coding of odors, mixtures and sniffs is a linear sum of odor time profiles. *Nat. Neurosci.* **18**, 272–281 (2015). doi:10.1038/nn.3913 Medline
20. K. R. Varshney, L. R. Varshney, Olfactory signal processing. *Digit. Signal Process.* **48**, 84–92 (2016). doi:10.1016/j.dsp.2015.09.012

Acknowledgments

This research was supported in part by grants from the NIH (R01DC013339 to J.D.M., R01MH106674 and R01EB021711 to R.C.G., UL1RR024143 to Rockefeller University); the Russian Science Foundation (#14-24-00155 to M.D.K. of the DREAM Olfaction Prediction Consortium [see supplementary materials (SM)]; the Slovenian Research Agency (P2-0209 to B.Z. of the DREAM Olfaction Prediction Consortium); the Research Fund KU Leuven (C.V.); and Flemish Agency for Innovation by Science and Technology-Flanders–Strategic Basic Research Project (IWT-SBO) NEMOA (L.S.). A.K. was supported by a Branco Weiss Science in Society Fellowship. G.S. is an employee of IBM Research. G.T. is partly funded by the Hungarian Academy of Sciences. L.B.V. is an investigator of the Howard Hughes Medical Institute. P.C.B. of the DREAM Olfaction Prediction Consortium has the support of the Ontario Institute for Cancer Research through funding provided by the Government of Ontario and a Terry Fox Research Institute New Investigator Award and a Canadian Institutes of Health Research New Investigator Award. R.K. is supported by a grant from the Council of Scientific and Industrial Research–Central Scientific Instruments Organisation, Chandigarh, India, K.D.G. by the Flemish Council of Scientific and Industrial Research (IWT) "InSPECtor" and European Research Council (ERC) Proof of Concept "SNIPER." L.B.V. is a member of the scientific advisory board of International Flavors & Fragrances, Inc. (IFF), and receives compensation for these activities. IFF was one of the corporate sponsors of the DREAM Olfaction Prediction Challenge. Y.I. is employed by Ajinomoto Co., Inc. J.D.M. is a member of the scientific advisory board of Aromyx and receives compensation and stock for these activities. Web links for data and models are provided below. On the website pages, individual predictions are known as "Subchallenge 1," and population prediction as "Subchallenge 2." Model details and code from the best-performing team for individual prediction (Team GuanLab; authors Y.G. and B.P. of the DREAM Olfaction Prediction Consortium): https://www.synapse.org/#!Synapse:syn3354800/wiki/; model details and code for the best-performing team for population prediction (Team IKW Allstars and author R.C.G.): https://www.synapse.org/#!Synapse:syn3822692/wiki/231036. DREAM Olfaction challenge description, participants, leaderboards, and data sets: https://www.synapse.org/#!Synapse:syn2811262/wiki/78368; model descriptions and predictions: https://www.synapse.org/#!Synapse:syn2811262/wiki/78388; code and details to reproduce analysis for scoring and to reproduce all the analysis for the figures: http://dream-olfaction.github.io.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.aal2014/DC1
Materials and Methods
Figures S1 to S7
Table S1
Reference (*20*)

21 October 2016; accepted 27 January 2017
Published online 20 February 2017
10.1126/science.aal2014

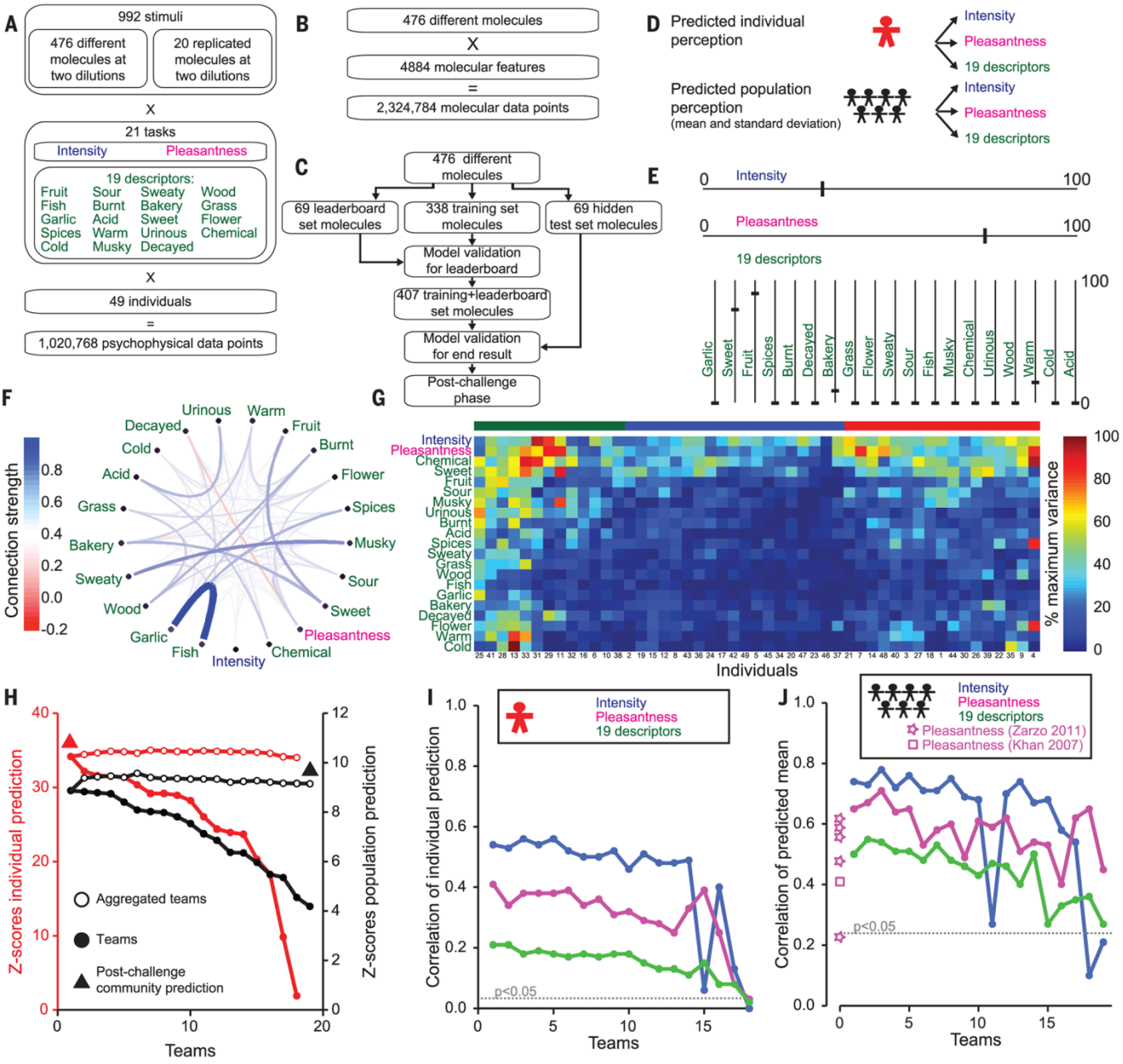

Fig. 1. DREAM Olfaction Prediction Challenge. (**A**) Psychophysical data. (**B**) Chemoinformatic data. (**C**) DREAM Challenge flowchart. (**D**) Individual and population challenges. (**E**) Hypothetical example of psychophysical profile of a stimulus. (**F**) Connection strength between 21 attributes for all 476 molecules. Width and color of the lines show the normalized strength of the edge. (**G**) Perceptual variance of 21 attributes across 49 individuals for all 476 molecules at both concentrations sorted by Euclidean distance. Three clusters are indicated by green, blue, and red bars above the matrix. (**H**) Model *Z*-scores, best performers at left. (**I** and **J**) Correlations of individual (I) or population (J) perception prediction sorted by team rank. The dotted line represents the $P < 0.05$ significance threshold with respect to random predictions. The performance of four equations for pleasantness prediction suggested by Zarzo (*10*) [from top to bottom: equations (10, 9, 11, 7, 12)] and of a linear model based on the first seven principal components inspired by Khan *et al*. (*8*) are shown.

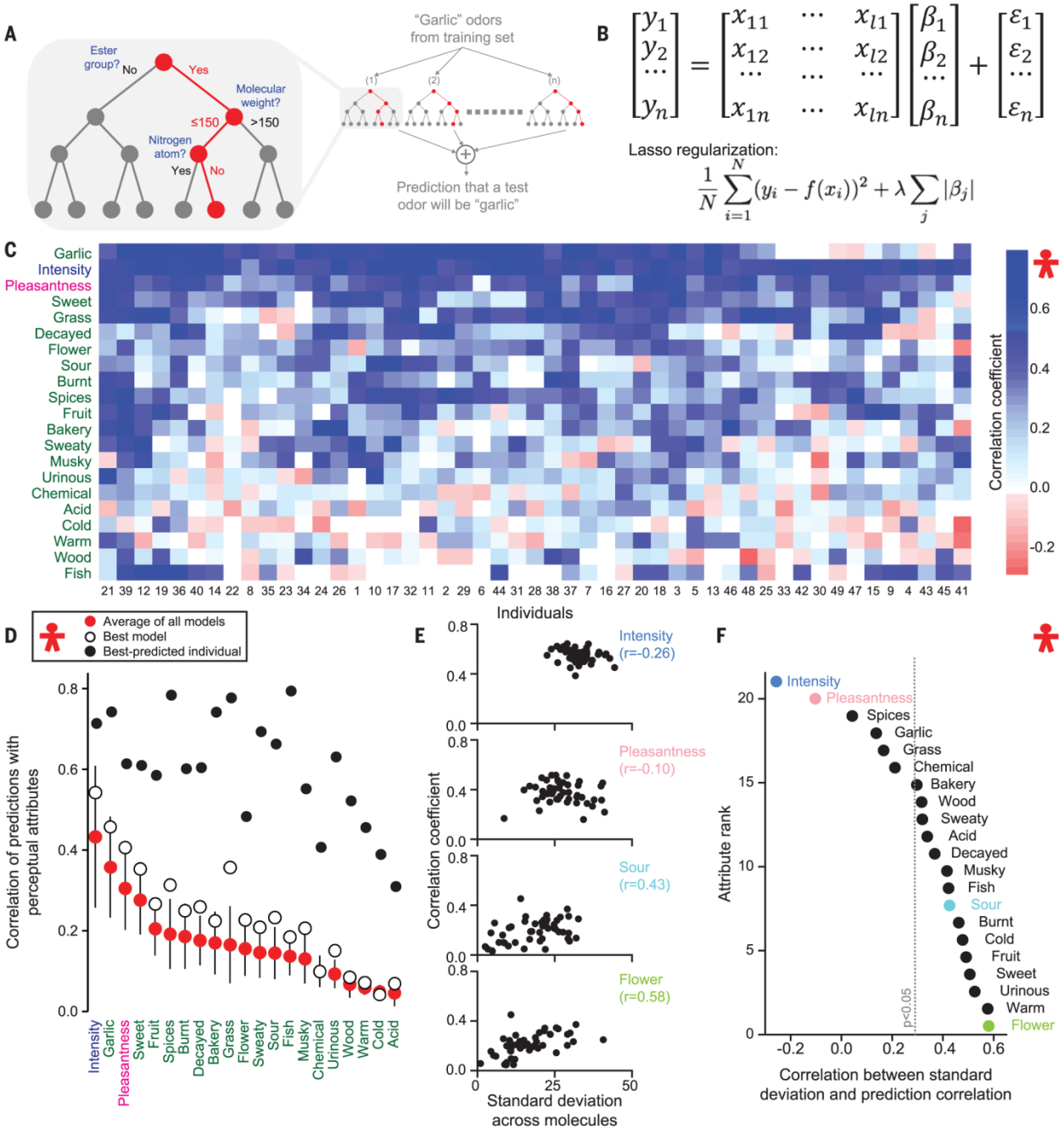

Fig. 2. Predictions of individual perception. (**A**) Example of a random-forest algorithm that utilizes a subset of molecules from the training set to match a semantic descriptor (e.g., "garlic") to a subset of molecular features. (**B**) Example of a regularized linear model. For each perceptual attribute y_i, a linear model utilizes molecular features $x_{i,j}$ weighted by β_i to predict the psychophysical data of 69 hidden test-set molecules, with sparsity enforced by the magnitude of λ. (**C**) Correlation values of best-performer model across 69 hidden test-set molecules, sorted by Euclidean distance across 21 perceptual attributes and 49 individuals. (**D**) Correlation values for the average of all models (red dots, mean ± SD), best-performing model (white dots), and best-predicted individual (black dots), sorted by the average of all models. (**E**) Prediction correlation of the best-performing random-forest model plotted against measured standard deviation of each subject's perception across 69 hidden test-set molecules for the four indicated attributes. Each dot represents one of 49 individuals. (**F**) Correlation values between prediction correlation and measured standard deviation for 21 perceptual attributes across 49 individuals, color coded as in (E). The dotted line represents the $P < 0.05$ significance threshold obtained from shuffling individuals.

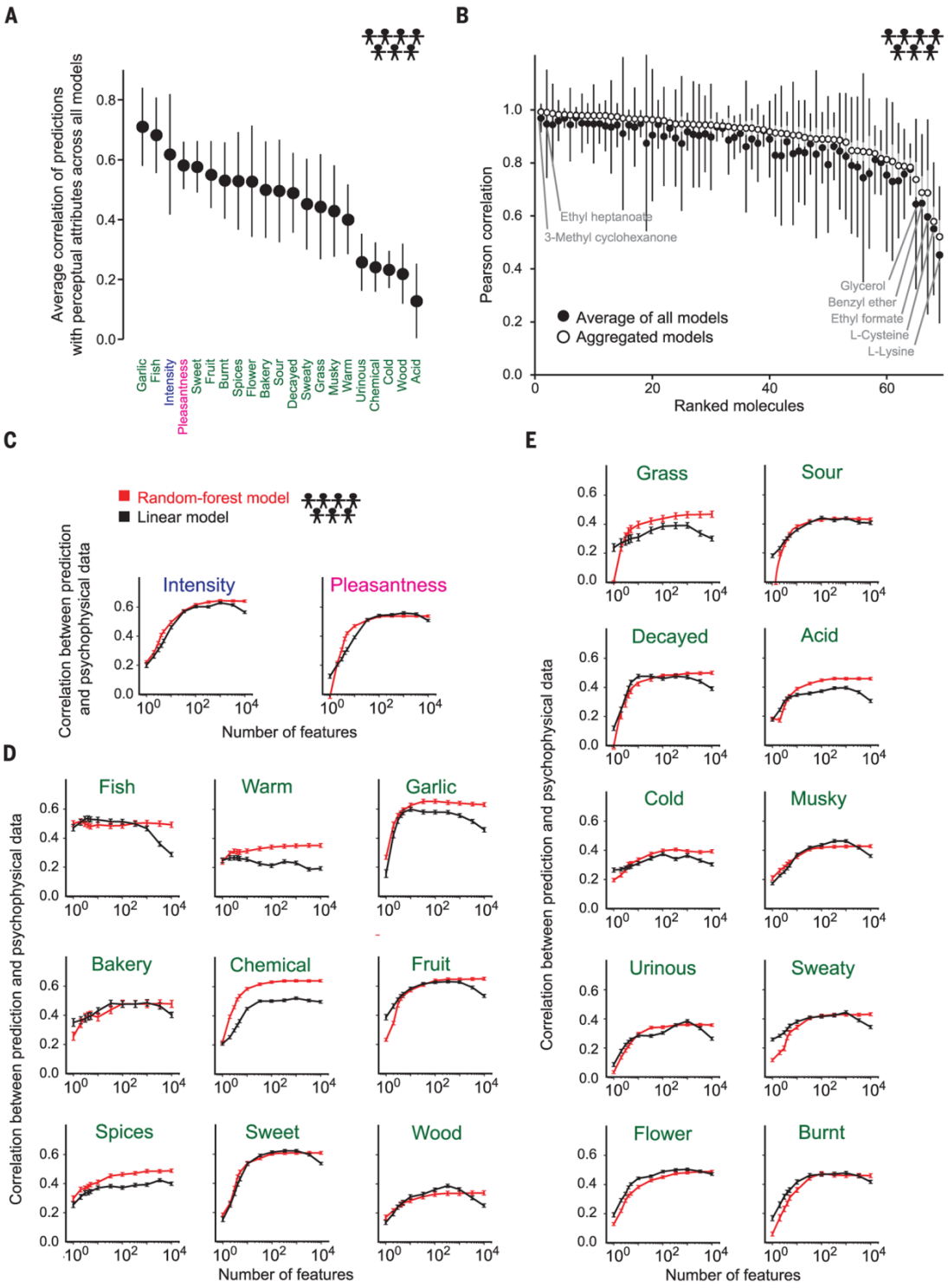

Fig. 3. Predictions of population perception. (**A**), Average of correlation of population predictions. Error bars, SDs calculated across models. (**B**) Ranked prediction correlation for 69 hidden test-set molecules produced by aggregated models (open black circles; gray bars, SD) and the average of all models (solid black dots; black bars, SD). (**C** to **E**) Prediction correlation with increasing number of molecular features using random-forest (red) or linear (black) models. Attributes are ordered from top to bottom and left to right by the number of features required to obtain 80% of the maximum prediction correlation using the random-forest model. Plotted are intensity and pleasantness (C), and attributes that required six or fewer (D) or more than six features (E). The combined training + leaderboard set of 407 molecules was randomly partitioned 250 times to obtain error bars for both types of models.

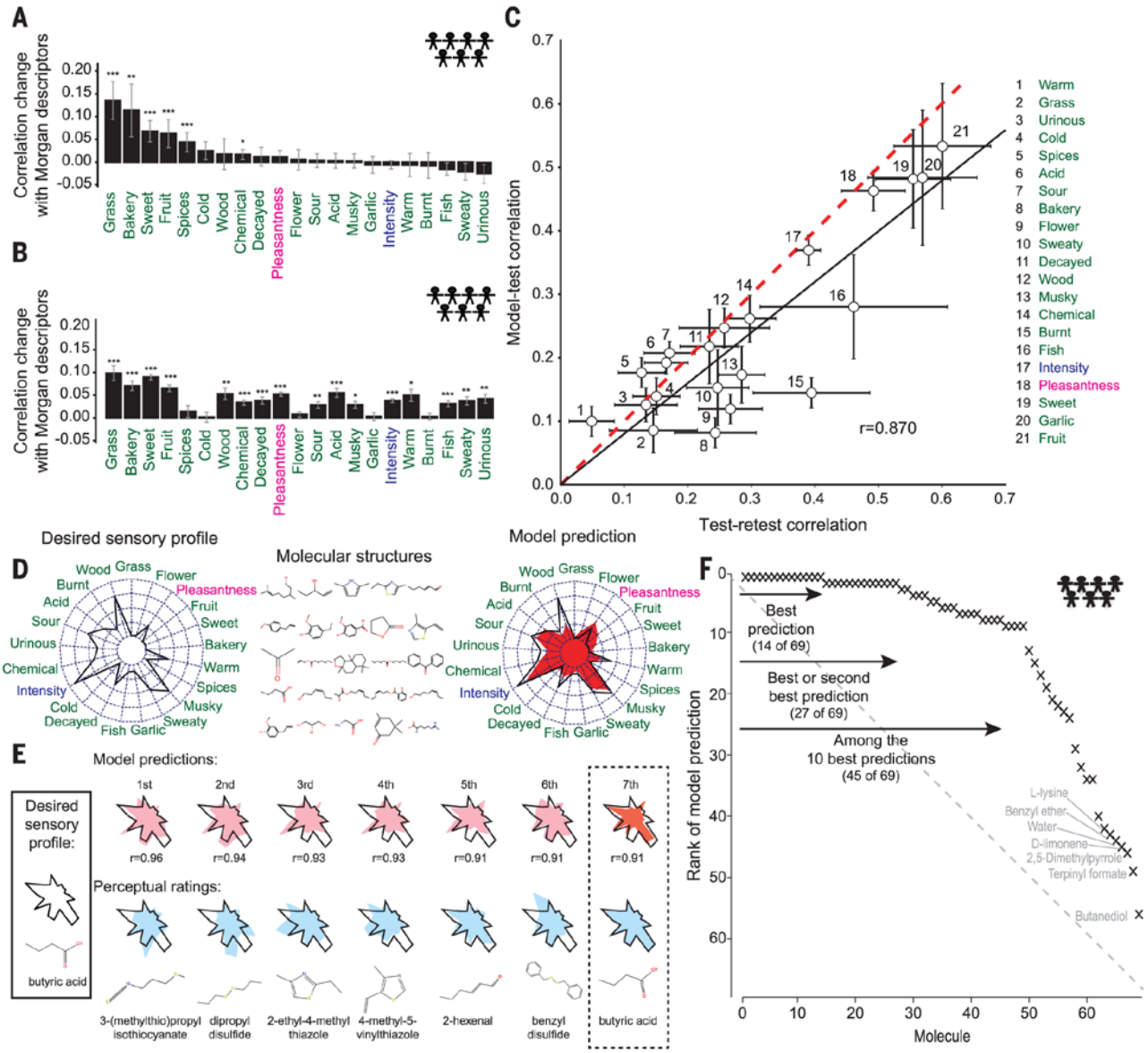

Fig. 4. Quality of predictions. (**A** and **B**) Community phase predictions for random-forest (A) and linear (B) models using both Morgan and Dragon features for population prediction. The training set was randomly partitioned 250 times to obtain error bars: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, corrected for multiple comparisons [false discovery rate (FDR)]. (**C**) Comparison between correlation coefficients for model predictions and for test-retest for individual perceptual attributes by using the aggregated predictions from linear and random-forest models. Error bars reflect standard error obtained from jackknife resampling of the retested molecules. Linear regression of the model-test correlation coefficients against the test-retest correlation coefficients yields a slope of 0.80 ± 0.02 and a correlation of $r = 0.870$ (black line) compared with a theoretically optimal model (perfect prediction given intraindividual variability, dashed red line). Only the model-test correlation coefficient for "burnt" (15) was statistically distinguishable from the corresponding test-retest coefficient ($P < 0.05$ with FDR correction). (**D**) Schematic for reverse-engineering a desired sensory profile from molecular features. The model was presented with the experimental sensory profile of a molecule (spider plot, left) and tasked with searching through 69 hidden test-set molecules (middle) to find the best match (right, model prediction in red). Spider plots represent perceptual data for all 21 attributes, with the lowest rating at the center and highest at the outside of the circle. (**E**) Example where the model selected a molecule with a sensory profile 7th closest to the target, butyric acid. (**F**) Population prediction quality for the 69 molecules in the hidden test-set when all 19 models are aggregated. The overall area under the curve (AUC) for the prediction is 0.83, compared with 0.5 for a random model (gray dashed line) and 1.0 for a perfect model.