Drawing the Borders of Olfactory Space

Emily J. Mayhew[1,*], Charles J. Arayata[1], Richard C. Gerkin[2], Brian K. Lee[3], Jonathan M. Magill[1], Lindsey L. Snyder[1], Kelsie A. Little[1], Chung Wen Yu[1], and Joel D. Mainland[1,4]

[1]Monell Chemical Senses Center, Philadelphia, PA 10104, USA.
[2]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA.
[3]Google Research, Brain Team, 355 Main St, Cambridge, MA 02142, USA.
[4]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104, USA.
[*]To whom correspondence should be addressed; E-mail: *emayhew@monell.org*.

December 4, 2020

Abstract: In studies of vision and audition, stimuli can be chosen to span the visible or audible spectrum; in olfaction, the axes and boundaries defining the analogous odorous space are unknown. Features governing the physical transport of molecules to olfactory receptors are sufficient to reliably classify novel molecules as odorous or odorless (AUROC = 0.97). Applying this model to a database of all possible small organic molecules, we estimate that over 30 billion possible compounds are odorous, 6 orders of magnitude larger than current estimates of 10,000. Remarkably, nearly all transport-capable molecules are odorous, suggesting broad collective tuning of olfactory receptors. Defining the boundaries of odor perception will enable design of experiments that representatively sample olfactory space and efficient search for novel odor compounds.

The number of molecules humans can smell is disputed, with published estimates ranging from 10,000 (*1*) to infinitely many (*2*). Chemical space is vast, and we cannot resolve this dispute until we define the subset of chemical space that has an odor. Critically, this knowledge gap means we cannot currently assess how well olfaction research represents the totality of odor space, nor where novel odorants may be found.

The journey of an odorous molecule to an olfactory receptor (OR) can be viewed as a mass transport problem (*3*). For a molecule to reach olfactory receptors and produce an odor percept it must be 1) volatile enough to enter the air phase and then the nasal cavity, 2) non-volatile and hydrophilic enough

to leave the air phase and sorb into the mucous layer coating the olfactory epithelium, 3) hydrophobic enough to enter an OR binding pocket, and 4) activate at least one OR (Fig. 1A). These requirements point to the types of molecular criteria we must consider (e.g. vapor pressure, hydrophobicity), but are insufficient to define precise limits or interactions between these constraints that determine which molecules are odorous.

Knowledge of the constraints also does not tell us how extensively each constraint reduces the pool of olfactory stimuli: does failure to complete the transport process (*4*) or failure of transport-capable molecules to activate an OR (*5*) eliminate more potential odorants? Here, we developed a quantitative model that can predict if a molecule is odorous or odorless and show that the transport process is the dominant limiting factor driving whether molecules are odorous.

A model using only three features that drive transport (boiling point, vapor pressure, and hydrophobicity) reliably classifies molecules as odorous or odorless (Fig. 1B). If odor classification can be explained by so few molecular properties, why were the classification rules not previously known? The likely reason is that available data are both noisy and poorly curated; we needed to gather odor classification data from multiple sources and correct errors in the data – both in transport features and odorous/odorless labels – before the three-parameter transport model matched the performance of more complicated models. Ultimately, we generated a large and chemically diverse dataset of over 1,900 molecules, classified as odorous (84%) or odorless (16%) through a combination of literature- and web-scraping, human discrimination tasks, and chemical analysis. A portion of the dataset, 60 of the molecules classified by human subjects and confirmed through chemical analysis, provided high-confidence odor classifications critical for model tuning (30 molecules) and measurement of final performance (30 molecules). The remaining molecules formed our training set, and we applied machine learning (ML) algorithms to train a variety of odor classification models. Our models represent molecular structure as a vector of physicochemical features (Dragon v6, Talete; EPI Suite, U.S. EPA), and calculate a probability that the molecule is odorous; all code used to generate models and figures is publicly available https://github.com/emayhew/OlfactorySpace.

We optimized ML models and measured performance using the area under the receiver operating characteristic curve (AUROC); AUROC penalizes false positives and false negatives equally, an important metric feature in classification problems where one class outnumbers the other. An AUROC value of 1.0 represents perfect classification, while 0.5 represents chance-level classification accuracy. We achieved near-perfect AUROC values in cross-validation with several algorithms (extreme gradient boosting – XGB, random forest – RF, support vector machine – SVM) when paired with a synthetic minority oversampling technique (SMOTE) to address the imbalance in odorous:odorless training examples. When applied to a test set of 30 molecules with high-confidence classifications held-out during model training, our best-performing XGB model separates odorous from odorless molecules with few errors and a high AUROC of 0.97 (Fig. 1C). Strong predictive performance on this held-out set suggests that the model will generalize well to new molecules.

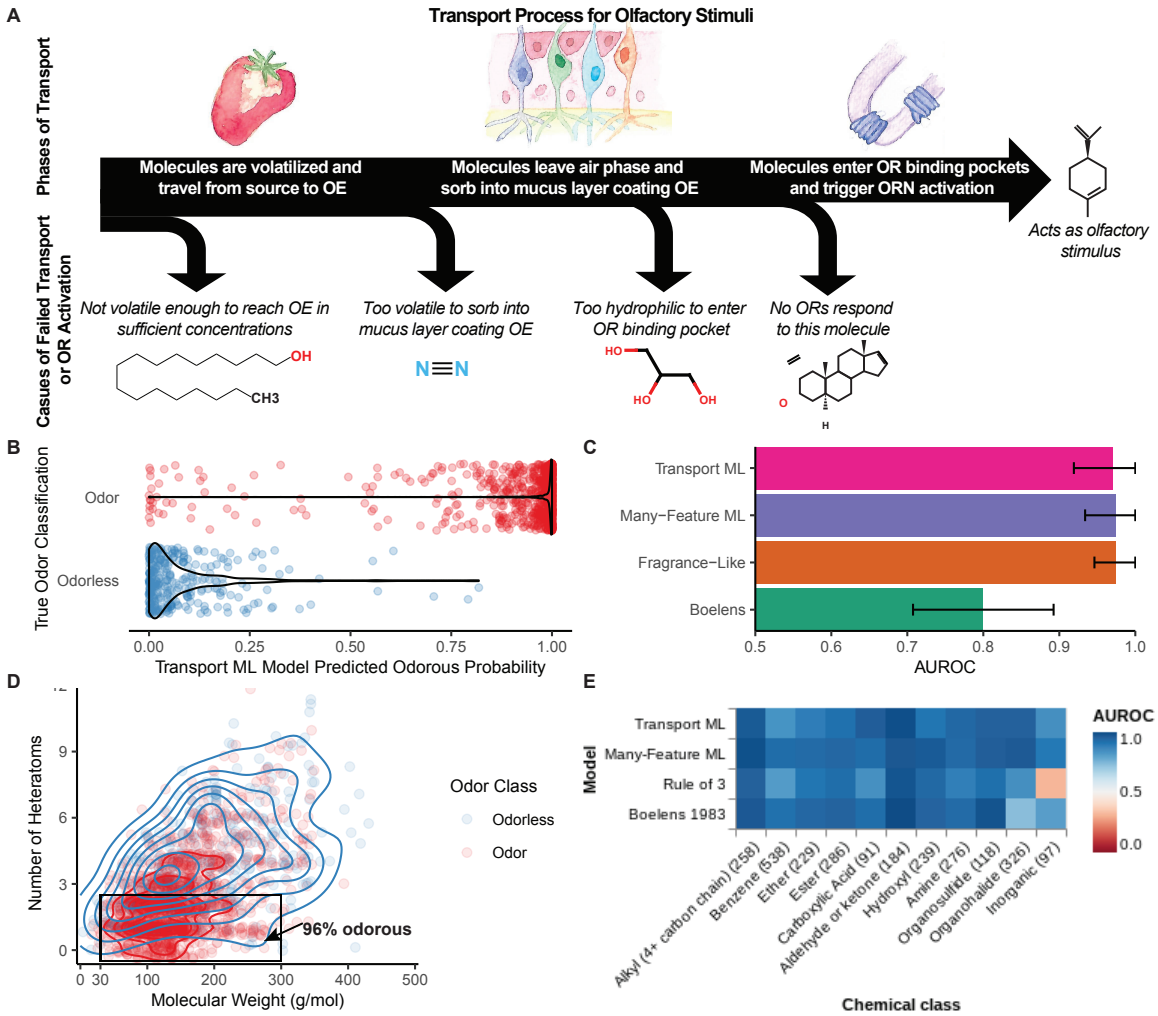

Figure 1: A machine learning model can accurately classify molecules as odorous or odorless based only on transport features. (**A**) Schematic of the transport process molecules must complete to act as olfactory stimuli. To elicit an odor, molecules must reach the olfactory epithelium (OE), enter olfactory receptor (OR) binding pockets, and trigger OR neuron (ORN) activation. (**B**) Transport-feature ML model-generated odorous probabilities for all molecules in the dataset. Each dot represents one molecule colored by the ground truth, and the width of the violin plot is the density of molecules at a given prediction value. (**C**) Model performance on a held-out test set of 30 molecules, quantified by the Area Under the Receiver Operating Characteristic (AUROC) curve, for transport-feature and many-feature machine learning models and published fragrance-like (*6*) and Boelens' (*4*) models. (**D**) Density of odorous and odorless molecules in transport space defined by molecular weight and number of heteroatoms. Each successive contour line indicates a step increase in density (red = 0.05%, blue = 0.015%). Molecules have a discrete number of heteroatoms, but are jittered along the y-axis to better show density. Plotted within the black box, molecules that obey the "rule of three" with fewer than 3 heteroatoms and molecular weight between 30 and 300 g/mol are generally odorous. Relative performance of models depends on functional group/chemical class. (**E**) Heatmap of mean AUROC generated by four models for molecules of common chemical classes (number of matching molecules in parentheses).

We next compared the performance of our transport-feature ML model to a "many-feature" ML model trained with over 3700 physicochemical features. Even when regularization is applied to reward use of few, highly relevant features to generate predictions, adding features does not improve classification accuracy (AUROC = 0.98, Fig. 1C). This finding suggests that transport features are sufficient to classify a molecule as odorous or odorless.

There are few published attempts to predict the odor status of molecules based on structure. The simplest, a transport-based model proposed by Boelens (*4*), classifies molecules based solely on boiling point and hydrophobicity, represented by the log of the octanol-water partition coefficient (log P). Boelens' transport model consists of a parabola in log P vs. boiling point space where molecules above the parabola are predicted to be odorous and molecules below it odorless; notably, the publication does not include information on the data or fit method used to derive this boundary, and its accuracy is untested. More recently, Ruddigkeit et al. (*6*) proposed a set of criteria that define "fragrance-like" molecules, including heavy atom count (HAC ≤ 21), atom types (C,H,O,S only), and number of hydrogen bond donors (HBD ≤ 1). These criteria represent trends found in databases of fragrance molecules, but their classification accuracy is likewise untested.

We applied Boelens' transport model and Ruddigkeit et al.'s fragrance-like model to our test set to benchmark our ML transport model and found that although both models perform well (Boelens' model AUROC = 0.80; fragrance-like model AUROC = 0.98), our model substantially outperforms Boelens' transport model (Fig. 1C). Sharma and others recently published a deep neural network model that predicts odor classification with comparable accuracy (AUROC 0.98) using 1622 chemical features and molecules drawn from internet databases (*7*); however, our results indicate that this degree of model complexity is not necessary to generate reliable predictions.

Our transport ML model achieves high accuracy with few features, but experimental values of boiling point, vapor pressure, and log P are unavailable for many molecules. We also developed a simple rule of thumb that does not require measured experimental values, can be applied knowing only the molecular formula (e.g. $C_{10}H_{12}O_2$), and preserves most of the performance: the "rule of three" states that molecules with molecular weight between 30 and 300 Da and with fewer than 3 heteroatoms are generally odorous (Fig. 1D). In our dataset, 96% of the molecules that meet these criteria are odorous (positive predictive value); these criteria capture 68% of odorous molecules (sensitivity) and exclude 84% of odorless molecules (specificity).

Next, we asked if the relative accuracy of the transport ML, many-feature ML, "rule of three", and Boelens (*4*) models varied by chemical class. Fig. 1E shows the test set AUROC for common chemical classes, averaged over 80 models trained on randomized train/test splits. All models achieved strong predictive accuracy for alkanes, alcohols, and carbonyl-containing molecules, but only ML models accurately classified organohalides. The "rule of three" underperforms on inorganic compounds because molar mass is not a good proxy for volatility of inorganic compounds (e.g. NaCl: MW = 54 g/mol, BP = 1465). The strong performance of the transport ML model independent of chemical class supports the reliability of the model across common classes of molecules.

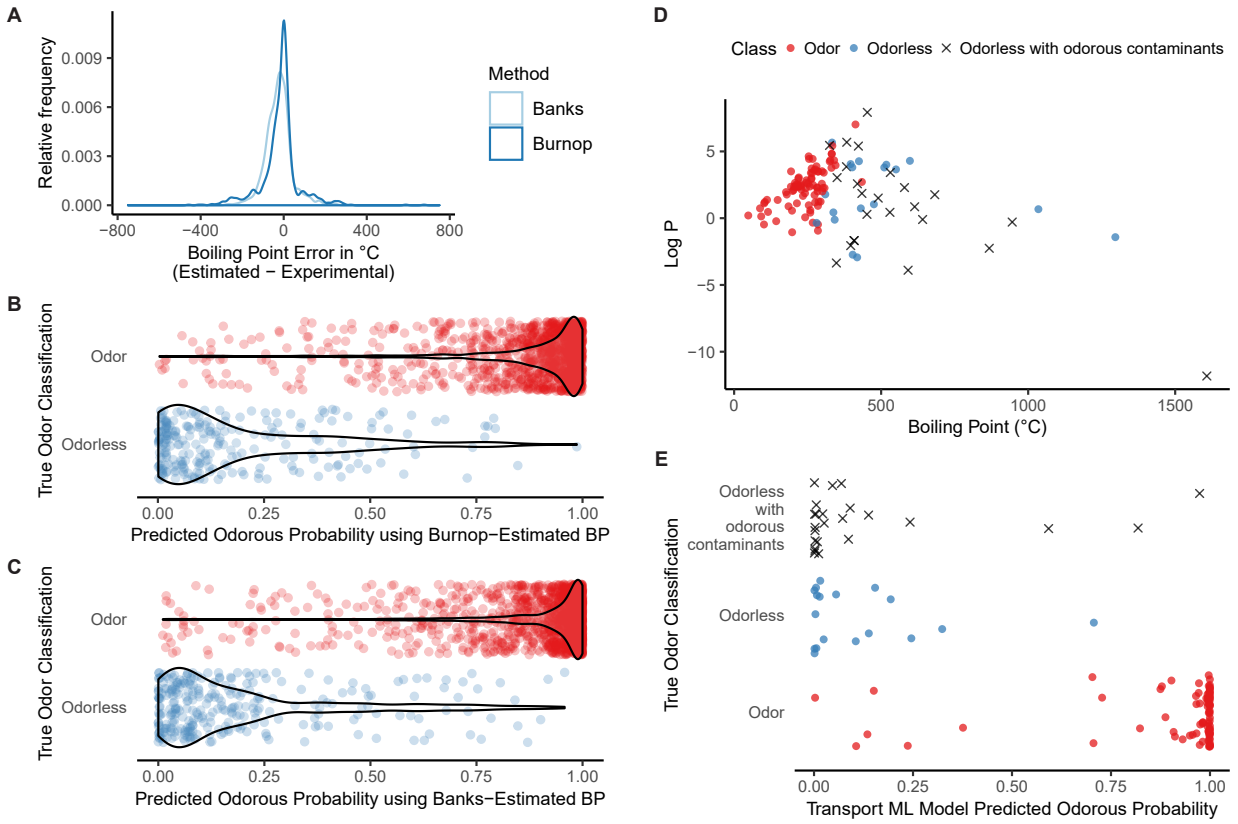

Figure 2: Inaccuracies in data impact model performance. (**A**) Difference between experimentally determined boiling point (BP) values and BP values calculated using the Burnop (*8*) and Banks (*9*) methods. (**B** - **C**) Odor classification predictions by transport-feature ML models using only estimated BP values calculated by the (B) Burnop or (C) Banks method. (**D**) Human subject-classified molecules in transport space defined by BP and log P. Many clearly non-volatile molecules were initially classified as odors due to odorous contaminants. (**E**) Transport-feature ML model odor predictions for human subject-classified molecules. Chemical compounds that are odorless but had odorous contaminants are correctly predicted to be odorless by the model.

Building a high performing transport model required correction of two major sources of error: boiling point values and odor classifications. Boiling point (BP) is commonly estimated from chemical formula, but these estimates are error-prone. We compared experimental BP (EPI Suite, U.S. EPA; PubChem, pubchem.ncbi.nlm.nih.gov) to estimates generated using the Burnop (*8*) and Banks (*9*) methods and found high median errors of 28 °C (Burnop) and 37 °C (Banks) even within these methods' domains of applicability (Fig. 3A). Using estimated BP values limits model accuracy; the classification performance of a ML model trained with identical parameters but only estimated BP values is worse (Fig. 2B-C: cross-validation AUROC = 0.92 (B), 0.95 (C); error rate = 5.8% (B), 5.6% (C)) than our final ML transport model (Fig. 1B: CV AUROC = 0.99, error rate = 2.4%). Experimental values should be used to generate high confidence odorous/odorless predictions, though estimated values may be adequate where a higher error rate is acceptable.

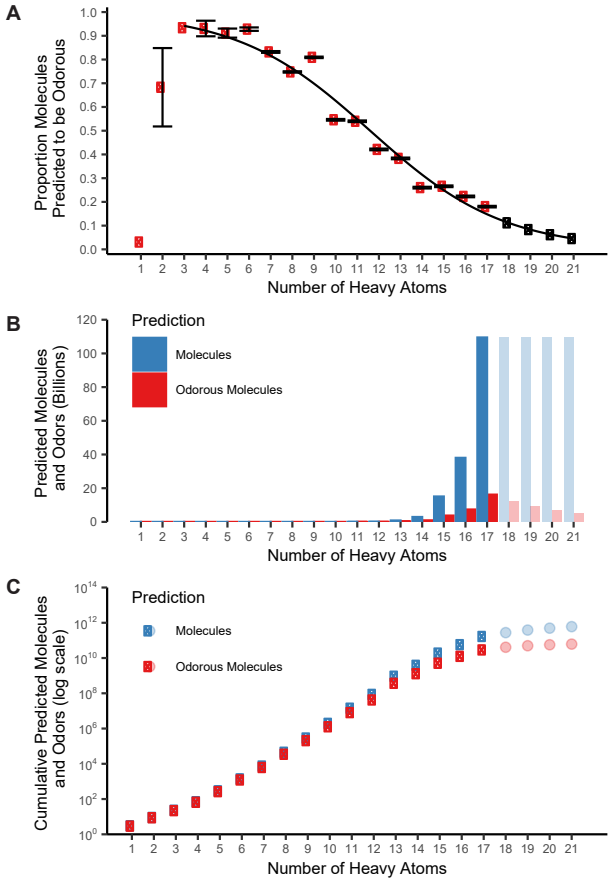

Figure 3: The transport model can be used to predict the population of odor space. (**A**) Proportion of molecules predicted by transport ML model to be odorous as a function of heavy atom count (HAC). Red circles show the mean probability generated for HAC tranches from the GDB database (*12*) with standard error bars; hollow circles show the projected odorous probability generated by the logistic regression fit plotted in black. (**B**) Estimated number of possible molecules and predicted odorous molecules from the GDB databases as a function of HAC. Lighter shaded bars indicate HAC values beyond GDB-17; as a conservative estimate, the number of possible molecules with HAC 18-21 was set to equal that of HAC 17. (**C**) Cumulative estimates of possible molecules and odorous molecules with increasing HAC on a logarithmic scale. The red datapoint at HAC 17 reflects our conservative estimate of 33 billion odorous molecules; the light red datapoint at HAC 21 reflects our less-conservative estimate of 62 billion odorous molecules.

The second major source of error that we rectified is the odor classifications themselves. ML models can tolerate some noise in the training data, but inaccuracies in the validation set used to tune models and especially in the test set used to measure final model performance can be more costly; specifically, predictive performance is bounded from above by mis-labeled data in the test set (*10*). To ensure our validation and test set were composed of accurately classified molecules, we supplemented our literature and web-scraped data with 128 molecular odor classifications (111 odorous, 17 odorless) generated through human psychophysics experiments. We analyzed all 111 compounds with a human-detectable odor using paired GC-MS and GC-O as a quality control (QC) measure to identify cases in which impurities, rather than the target compound, were responsible for the odor detected by human subjects. We found that 22% of molecules classified as odorous were actually odorless compounds contaminated with odorous compounds, despite high nominal purity ratings from vendors (Fig. 2D). Had we not performed this QC, we would have falsely believed that model performance was poor (pre-corrected Transport ML AUROC = 0.75). In fact, most disagreements between our model's predictions and pre-QC classifications were due to the model correctly identifying misclassifications in the dataset (Fig. 2E).

Chemical compounds are common stimuli in olfaction research, but the impact of impuri-

ties on data is rarely discussed. Odor detection thresholds vary by many orders of magnitude across molecules, so even high purity (e.g. 99%) is insufficient grounds to consider odor to be driven entirely by the dominant molecular species (*11*). Addressing the impact of contaminants on our data was thus vital for accurately measuring model performance.

The dispute over the size of odor space has endured in the field because the criteria for odorous molecules were not rigorously defined. We now have a tool to address this debate: a quantitative model that makes highly accurate odor classification predictions and generalizes to new molecules. Chemical space is vast, but it can be enumerated. Applying graph theory and rational chemical constraints (e.g. bonds per atom, bond angles, ring strain), Ruddigkeit et al. (*12*) developed GDB-17, a database of 166 billion unique molecules with 17 of fewer heavy atoms. Due to the limited range of atoms included (C, H, N, O, S, halogens) and careful elimination of unstable structures, the size of this database can be thought of as a conservative estimate for the number of possible molecules with heavy atom count (HAC) ≤ 17.

We applied our model to a random sample of over 100,000 molecules from GDB-17 (*12*), including all molecules with HAC 1 to 7 and approximately 10,000 molecules each with HAC 8 to 17, to generate odor classification predictions. Almost 50,000, or about half, of the molecules we sampled were predicted to have an odor. The proportion of molecules predicted to be odorous is highly dependent on HAC; more than 75% of molecules with between 3 and 9 heavy atoms are predicted to be odorous, but the proportion declines with increasing HAC to approximately 5% predicted odorous for HAC 17 (Fig. 3A). By multiplying the proportion of predicted odors by the number of molecules for each HAC in GDB-17, we calculate that there are at least 33 billion possible odorous molecules (Fig. 3B-C).

This estimate is conservative because we know the odorous range extends beyond HAC 17 and that GDB-17 does not include all possibly odorous chemical structures. We cannot confidently estimate the number of possible molecules with greater than 17 heavy atoms, but we can safely assume there are at least as many possible unique molecules with 18 heavy atoms as there are with 17. Similarly, we do not know what proportion of molecules above HAC 17 will be odorous, but we can assume that the proportion will approach 0 with increasing HAC. We set an upper HAC bound of 21 because odorous molecules with up to 21 heavy atoms have been identified (*6*). We fit a logistic regression model to transport ML model predictions on our GDB subset (Fig. 3A) to predict the rate of decline in the proportion of odorous molecules with increasing HAC. To generate a less conservative estimate, we then applied these extrapolated odorous proportions to the number of molecules with 17 heavy atoms in GDB-17, a conservative estimate for the number of molecules with HAC 18-21. This approach yields an estimate of 62 billion odorous molecules.

Our estimate is sensitive to the composition of our training set: increasing the ratio of odorous molecules in the dataset increases the proportion of predicted odors.

There is significant uncertainty in our estimates, both of the proportion of molecules that are odorous and the number of possible molecules, but all variants of our approach

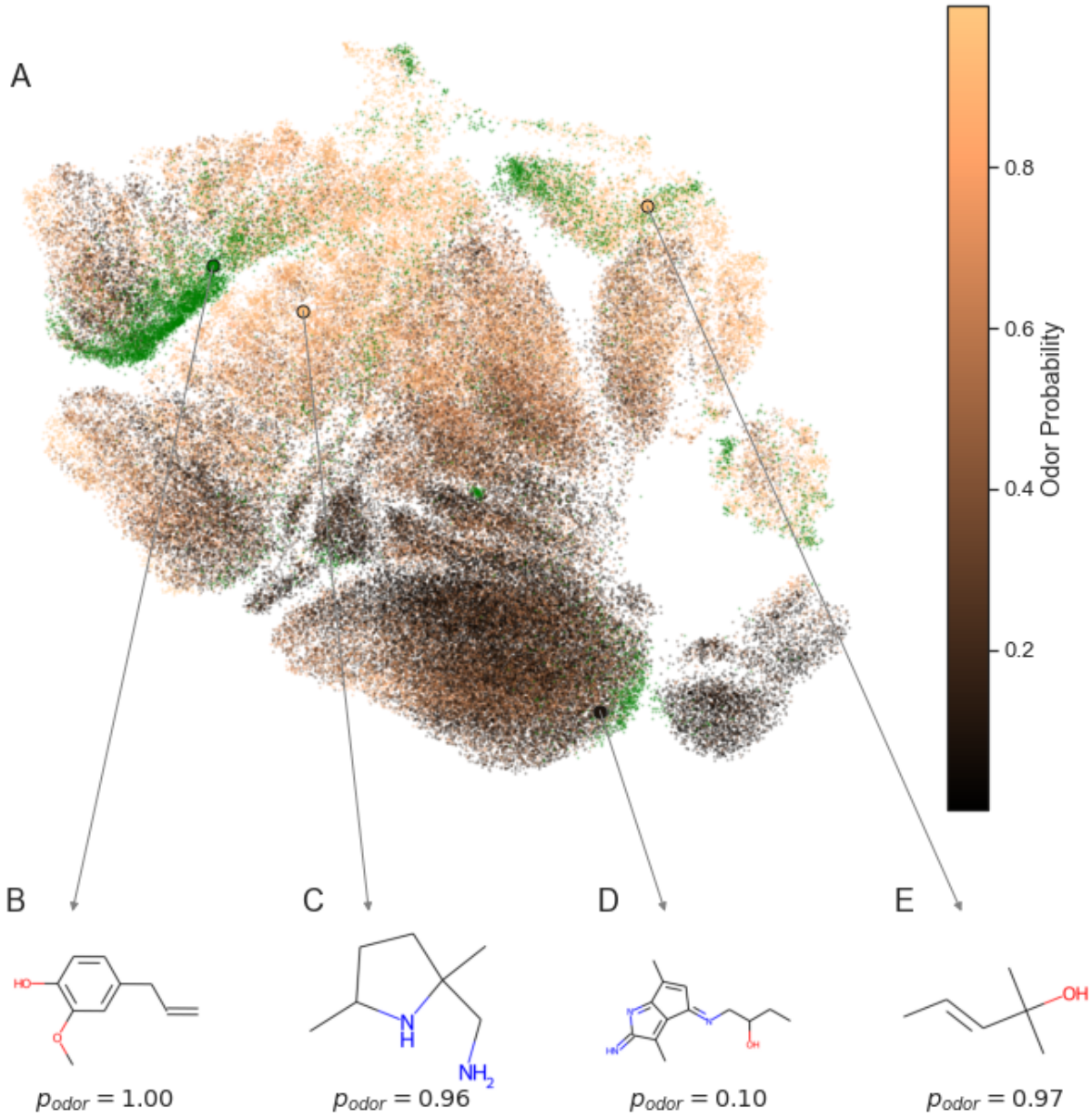

Figure 4: Visualization of olfactory space highlights understudied regions. (**A**) Uniform Manifold Approximation and Projection (UMAP) plot of known odorous molecules (green) and possible molecules from GDB-17 colored by their transport ML-predicted odorous probability. Many regions dense with probable odors are sparsely represented by known odors. (**B**) Eugenol, a known odorant. (**C-E**) Example molecules from GDB-17 and their transport ML-predicted probability of being odorous (p_{odor}).

have yielded estimates between 10 and 200 billion. Therefore, we predict that there are at least tens of billions of possible odorous molecules, a value 6 orders of magnitude larger than the commonly cited estimate of 10,000 odorants (*1*).

Any database or catalog of purchasable odorous molecules is dwarfed by the scale of odor space. For example, the Sigma-Aldrich Flavor and Fragrance Catalog typically lists only 1,600 molecules. Many densely populated regions of odorous chemical space (Fig. 4A) are sparsely represented by known odorants (Fig. 4B). Our map of the odorous region of chemical space can identify new likely odorants (Fig. 4C, 4E) and be used to filter out unlikely odorants (Fig. 4D). The sheer number of as-yet-unsynthesized odorous molecules is striking and there are whole classes of odorous molecules that have not been synthesized and whose odor characteristics are to date unknown.

Drawing the borders of olfactory space is largely a matter of understanding physical transport. Nearly all molecules that can physically enter a receptor binding pocket have an odor. Although there are some clear exceptions, for some individuals, such as androstenone (*13*), few molecules appear to be odorless simply because there is no sufficiently sensitive receptor. Another piece of evidence suggesting that the receptor repertoire is generally broadly tuned enough to detect most odors is that methane, which has a high vapor pressure and so does not sorb into the mucosa at standard pressure, is odorless. However, when divers were given methane at 13 atmospheres, a pressure at which it can dissolve into mucosa, it smelled of camphor (*14*). The receptor repertoire over the course of human evolution could not have evolved to detect this molecule, yet this repertoire maintains enough general sensitivity to – once given access to it – detect it nonetheless.

Our model answers the question of what makes a molecule odorous, but many questions remain. Our estimate of the number of possible odorous molecules cannot resolve the number of discriminable odor percepts (*15 –17*); many of these predicted odorants may have indistinguishable odor percepts, and odorant mixtures may produce percepts that are distinct from that of any single odorant. Everything we know about odorants is derived from a tiny subset of all volatiles—a catalog of all volatiles present in foods represents less than 0.000002% of the molecules we can smell. This model invites researchers into the unknown, providing a map to new regions of odor space and the means to representatively sample it.

References

1. L. Buck, R. Axel, A novel multigene family may encode odorant receptors: A molecular basis of odor recognition. *Cell* .**65** , 175–187 (1991).

2. J.-N. Jaubert, C. Tapiero, J.-C. Dore, The Field of Odors : Toward a Universal Language for Odor Relationships. *Perfum. Flavorist*(1995), pp. 1–16.

3. I. Hahn, P. W. Scherer, M. M. Mozell, A mass transport model of olfaction. *J. Theor. Biol.* **167** , 115–128 (1994).

4. H. Boelens, Structure-activity relationships in chemoreception by human olfaction. *Trends Pharmacol. Sci.* **4** , 421–426 (1983).

5. E. Poivet, N. Tahirova, Z. Peterlin, L. Xu, D. J. Zou, T. Acree, S. Firestein, Functional odor classification through a medicinal chemistry approach. *Sci. Adv.* **4** , eaao6086 (2018).

6. L. Ruddigkeit, M. Awale, J. L. Reymond, Expanding the fragrance chemical space for virtual screening. *J. Cheminform.* **6** , 27 (2014).

7. A. Sharma, R. Kumar, R. Semwal, I. Aier, P. Tyagi, P. Varadwaj, DeepOlf: Deep neural network based architecture for predicting odorants and their interacting Olfactory Receptors. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* (2020), doi:10.1109/tcbb.2020.3002154.

8. V. C. E. Burnop, Boiling point and chemical constitution. Part 1. A additive function of molecular weight and boiling point. *J Chem Soc* , 826–829 (1938).

9. W. H. Banks, Considerations of a vapour pressure-temperature equation, and their relation to Burnop's boiling-point function. *J Chem Soc* , 292–295 (1939).

10. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J. Am. Med. Assoc.* **316** , 2402–2410 (2016).

11. M. Paoli, D. Münch, A. Haase, E. Skoulakis, L. Turin, C. G. Galizia, Minute impurities contribute significantly to olfactory receptor ligand studies: Tales from testing the vibration theory. *eNeuro* .**4** (2017), doi:10.1523/ENEURO.0070-17.2017.

12. L. Ruddigkeit, R. Van Deursen, L. C. Blum, J. L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52** , 2864–2875 (2012).

13. A. Keller, H. Zhuang, Q. Chi, L. B. Vosshall, H. Matsunami, Genetic variation in a human odorant receptor alters odour perception.*Nature* . **449** , 468–472 (2007).

14. P. Laffort, C. Gortan, Olfactory properties of some gases in hyperbaric atmosphere. *Chem. Senses* . **12** , 139–142 (1987).

15. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans can discriminate more than 1 trillion olfactory stimuli. *Science (80-.).* **343** , 1370–1372 (2014).

16. R. C. Gerkin, J. B. Castro, The number of olfactory stimuli that humans can discriminate is still unknown. *Elife* . **4**(2015), doi:10.7554/eLife.08127.

17. M. Meister, On the dimensionality of odor space. *Elife* .**4** (2015), doi:10.7554/eLife.07865.

Acknowledgements

The authors acknowledge support from the George Preti Research Support Core for Analytical Chemistry and would like to thank George Preti, Katherine A. Prokop-Prigge, Bruce A. Kimball, Kai Zhao, and Gary K. Beauchamp for comments and discussions. **Funding:** This research was supported in part by grants from the NIH (R01DC013339, U19NS112953, R01DC018455, and R01DC017757) and the Ajinomoto Co. Innovation Alliance Program. E.J.M. was supported by T32DC000014 and F32DC019030. **Author contributions:** C.J.A., J.M.M., L.L.S., and C.W.Y. collected human behavioral data. L.L.S., K.A.L., and E.J.M carried out quality control using GC-O. E.J.M., C.W.Y., K.A.L., R.C.G., and J.D.M. curated external data. B.K.L. validated the models and analyzed the domain applicability. C.W.Y and C.J.A conducted initial data cleaning and analysis, and E.H.M and R.C.G. wrote the finalized models and visualized the data. E.J.M. prepared the initial draft of the manuscript. J.D.M. conceived and supervised the project, and edited the manuscript. J.D.M. and E.J.M. acquired funding for the project. All authors discussed and made contributions to the final version. **Competing interests:** J.D.M. received research funding from Ajinomoto Co., Inc. **Data and materials availability:** All data and code are available in the supplementary materials.