PNAS
www.pnas.org

Supplementary Information for
Transport Features Predict If a Molecule is Odorous

Emily J. Mayhew, Charles J. Arayata, Richard C. Gerkin, Brian K. Lee, Jonathan M. Magill, Lindsey L. Snyder, Kelsie A. Little, Chung Wen Yu, Joel D. Mainland

*Emily J. Mayhew
Email: mayhewem@msu.edu

This PDF file includes:

Supplementary text (Materials and Methods)
Table S1
Figures S1 to S4
Legends for Datasets S1 to S3

Other supplementary materials for this manuscript include the following:

Datasets S1 to S3

Materials and Methods

To build models that can predict odorous/odorless status for any molecule, we generated a large dataset with sufficient chemical diversity to learn generalizable rules relating chemical structure to odor status. We used two strategies to gather information on odor status: scraping odor publicly available classifications from the literature and websites (low cost, lower confidence) and testing compounds with human subjects (high cost, higher confidence). We classified 128 molecules as odorous or odorless through human subject testing and 1796 additional molecules through web and literature searches; in total, our dataset includes 1924 unique molecules. We used chemoinformatic softwares OpenBabel, Dragon, and EPISuite to generate chemical features for each molecule. More detailed information on the construction of this dataset follows. Code used to generate models and figures can be found at https://github.com/emayhew/OlfactorySpace.

Gathering odor classifications from literature sources and websites. Molecules with stated odorous/odorless classifications were gathered from the literature (1–6) and from databases of odorous molecules (Sigma Aldrich Flavors & Fragrances, www.sigmaaldrich.com/industries/flavors-and-fragrances.html; The Good Scents Company database, www.thegoodscentscompany.com). Additionally, we collected and classified molecules with odor information from the websites Available Chemicals Directory (psds.ac.uk/acd; keyword: odorless), Wikipedia (wikipedia.org; keyword: odorless), and PubChem (pubchem.ncbi.nlm.nih.gov; keywords: odor, odour, smell, fragrance, aroma, sense of smell, no odor, no odour, no smell, no fragrance, odorless, odourless). The source of the odorous/odorless classification for each molecule in our dataset is given in Dataset S1.

Selecting compounds for human psychophysics testing. In selecting chemical compounds to be evaluated by human subjects, we were constrained by the availability of safety data. We gathered candidate molecules from four chemical libraries in which the majority of compounds are considered safe by experts: Generally Recognized as Safe database (GRAS, www.fda.gov/food/generally-recognized-safe-gras/gras-substances-scogs-database), Prestwick chemical library (www.prestwickchemical.com/screening-libraries/prestwick-chemical-library/), The Good Scents Company database (www.thegoodscentscompany.com), and Arctander's Perfume and Flavor Chemicals database (Arctander 1969). From these databases, we gathered a total set of 6009 unique compounds and calculated structural and physicochemical descriptors (Dragon v6, Talete; www.talete.mi.it).

Machine learning (ML) models can only learn rules and patterns present in the training set. In order to build well-performing and generalizable models, it is critical to generate a training set that represents the full space that the model should describe and is drawn from the same population as the validation and test sets. To choose a set of compounds that can best represent the totality of chemical space, we performed k-means clustering on the centered and scaled Dragon descriptors with thirty clusters and aimed to pick four compounds from each cluster. By including molecules from every cluster in the training, validation, and test sets, we ensured that patterns learned by our models applied to the full range of chemical compounds and could be tested on validation and test sets with a similar composition. A total of 128 compounds (Dataset S2) were selected to span the maximum range of chemical space regardless of their perceptual properties. All 128 compounds evaluated by human subjects for our study have no known risk to humans at the delivered concentrations, and have been approved by authoritative agencies such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMEA).

Determining odor classification using human psychophysics. We prepared samples for human subject testing by aliquoting neat material (1 mL of liquid compounds or 1 g of solid compounds) into triple-washed amber jars (2 acidic and 1 neutral wash cycle). Compounds with an extremely strong odor or compounds that were not available in large quantities (e.g. reserpine) were presented at lower concentrations or in dilution; information on compound dosing is included in Dataset S2. Between experimental sessions, we stored jars containing chemical compounds in a 3-shelf acrylic cabinet supplied by a low flow of continuous carbon-filtered air to ensure the stored odorants would not be contaminated by outside ambient air or other jars. During testing sessions, experimenters wore lab coats and odorless gloves to prevent contamination of the jars.

Ninety unique normosmic participants (64 females) aged 18-54, (median 27.9) who consented to our study were tested, and the protocol for the present study was approved by University of Pennsylvania Institutional Review Board. Approximately 15 participants were recruited for each set of 5 compounds (hereby referred to as “blocks”) but were permitted to participate in multiple blocks. Twenty-six blocks were run in total (130 compounds).

Each test session consisted of 28 trials: 5 target compounds were presented 5 times each alongside two blank jars in a 3-alternative forced choice test. Participants were asked to choose the jar that contained an odorant, as well as rate their confidence in choosing the correct jar, ranging from 0 (Completely Unsure) to 10 (Extremely Confident). The remaining 3 trials were presentations of a control compound, linalool (1 mL, neat), to screen participant olfactory function. All participants were blindfolded during the entire test session, and the jars were delivered to the nose of participants by the experimenter. Participants also performed an auditory task during 30-second breaks between each trial to maintain subject attention while minimizing olfactory adaptation. The order of presentations of target and blank jars were randomized and responses were recorded using E-prime 3.0 (Psychology Software Tools, Pittsburgh, PA). The University of Pennsylvania Institutional Review Board approved this research protocol, and all human research participants gave informed consent.

After each block was completed, the probability of detection of each compound within the block was calculated. The scale of probability ranges from p=0.0 (undetectable) to p=1.0 (perfect detection), with a chance detection probability of p=0.33. Based on the binomial probability distribution for n=75 trials and chance p=0.33 (the cumulative, one-tailed probability of recording 33 or more correct trials out of 75 is p=0.035), compounds correctly discriminated from blanks in ≥ 33 trials were determined to be odorous (α=0.05). Once ≥ 33 correct identifications were made for all compounds in a block, we terminated data collection for that set of compounds and moved on to the next block. The number of completed and correct trials for each compound is included in Dataset S1, and the proportion of correct trials for each molecule is plotted in Fig. S4.

Correcting for odorous impurities in odorless compounds. We performed headspace extraction using StableFlex 2CM solid-phase microextraction (SPME) fibers (Supelco), exposing the SPME fibers for 5 minutes to the headspace of a jar prepared as described above. We then inserted the SPME fibers (60 s desorption time) into Thermo Scientific ISQ single-stage quadrupole GC-MS and Thermo-Fisher Trace GC Ultra GC-O instruments fitted with identical Restek 30-meter Stabilwax columns (1 μm coating) so that a given molecule should elute with the same retention time (RT) on both instruments. We used Xcalibur software (Thermo Electron Corp.) to analyze the GC-MS spectra and identified all compounds present above 10% relative abundance. In the case of a few compounds for which the volatility was too low for SPME, we performed direct injection of analytes (100 ppm). In direct injection experiments, the most abundant compound is guaranteed to be the nominal compound, and RT for the target molecule was determined by FID. We recorded the RT and odor quality notes for all odors perceived during each GC-O experiment.

Compounds which had either no peak in the GC-MS spectra or no corresponding GC-O odor within 60s of the GC-MS(RT) were reclassified as odorless. In total, 23 of the 111 compounds tested (21%) were reclassified, highlighting the importance of controlling for impurities in olfactory research. Results of QC on odor classification are shown in Fig. S4, and information on the cause of GC reclassifications is included in Dataset S2.

Preparing data for model-building. We pooled data on 128 compounds tested in our human psychophysics experiment with 1796 additional molecules for which an odor classification was available from literature sources or websites. Of these 1924 molecules, 1615 were classified as odorous and 309 were classified as odorless. We used Open Babel (7) to generate an energy-minimized 3-dimensional structure file (.sdf) for each molecule from the Simplified Molecular-Input Line-Entry System (SMILES) string. Next, we used the chemoinformatic software Dragon (Talete v6; http://www.talete.mi.it) to calculate 4885 structural and physicochemical descriptors based on the 3-D structure. Additionally, we gathered data on the boiling point and vapor pressure of the molecules. We calculated an approximate boiling point for each molecule using two published methods (8, 9). Due to large observed errors between experimental values and estimates generated with these methods, we also generated estimated and experimental boiling point and vapor pressure values using the program EPI Suite (U.S. EPA). Finally, we collected experimental boiling point values from PubChem. We used experimental data in all cases where it was available (1270 experimental boiling point values, 1122 experimental vapor pressure values) and estimates only where it was not. Boiling point estimates calculated using the Banks method were used only for comparison with experimental data and to make predictions on GDB molecules where no experimental values were available. Importantly, while we used multiple sources to generate chemical descriptors, the three features used by our transport ML model (boiling point, vapor pressure, octanol/water partition coefficient) are all available through open-access sources (EPI Suite, U.S. EPA).

Prior to any data analysis or modeling, we removed 60 compounds from the data set to form a test set and validation set of 30 compounds each (Dataset S1). Previous studies show that while ML approaches can learn well from noisy training sets, it is critical to have high confidence in the classification of examples used to measure model performance (10). Both the validation and test set molecules were drawn exclusively from the pool of 128 molecules tested in the lab. Because we wanted to ensure that validation and test sets were representative of the full data set, both datasets were composed of 1 molecule from each of 30 clusters and had an odorous:odorless ratio of 80:20 (the full lab-tested data set ratio at that time). The remaining 1867 compounds (68 tested in lab, 1796 gathered from the literature and websites) were pooled to form our training set (Dataset S1). Once the model training parameters were finalized, the 30 validation molecules were added to the training set.

We dropped any features with more than 10% missing values and then performed k-nearest neighbor imputation for remaining missing values (k=5, R package bnstruct v1.0.8). We normalized the raw physicochemical descriptors in the training set by scaling each descriptor from 0 to 1, then applied the same normalization factors to descriptors in the validation and test sets. Consequently, the model was blind to the range of descriptors outside of the training set (11, 12). We then trimmed the dataset by dropping descriptors that were highly correlated (r > 0.99) or had negligible variation between molecules. All data processing and analysis was conducted using the open-source statistical software R (version 3.5.3), and preprocessing steps were done using the preProcess function in the R package caret (version 6.0.81) (13).

Training and evaluating models. Selecting the best algorithm for a problem is often an empirical process. We compared the performance of five ML algorithms that have been successfully applied in related research to generate odor classification models: logistic regression, support vector machine (SVM) (14, 15), random forest (RF) (16–18), stochastic gradient boosting (GB) (19, 20), and extreme gradient boosting (XGB) (21, 22). Models were optimized during training to maximize the area under the receiver operating characteristic curve (AUROC) on cross-validation splits. We chose to optimize models based on AUROC because it more heavily penalizes false positives, making it a better metric than accuracy in cases with imbalanced classes. All models were trained using the caret package for R (13) (training control parameters: method = repeatedcv, number = 5, repeats = 2). We addressed this significant class imbalance in our dataset (1619 odorous:310 odorless) by applying a Synthetic Minority Over-sampling TEchnique (SMOTE) to our training set using the R package DMwR (version 0.4.1) (23). We tested several ratios of odorous:odorless training sets and chose the ratio (62 odorous:38 odorless) which optimized cross-validation AUROC.

The XGB algorithm produced the best-performing models as assessed through cross-validation (CV) AUROC. We applied regularization during model training. We used a tuning grid of L1 and L2 weights ranging from 1E-5 to 1 and selected the weights that resulted in the highest CV AUROC (L1: alpha = 0.4, L2: lambda = 0.2; CV AUROC = 0.9954); however, we found that CV AUROC did not vary much with changing weights (range: 0.9941 – 0.9954). Our final ML models were generated using the following parameters:

	ML transport model	ML many-feature model
Algorithm	eXtreme Gradient Boosting	eXtreme Gradient Boosting
nrounds	100	100
lambda	0.2	0.2
alpha	0.3	0.4
eta	0.3	0.3

We tested the performance of all models on our held-out test set of 30 molecules. ROC curves for models trained with each algorithm and feature set are plotted in Fig. S2.

To better measure the uncertainty in model performance, we subsequently tested on 25 random draws of 30 lab-tested molecules, training on all remaining molecules. Model tuning parameters and AUROC mean and median values across the 25 splits are reported in Table S1. We report these performance statistics in the manuscript, but the original model (trained on all molecules except those in the held-out test set) is used to produce subsequent figures and analyses.

Comparing model performance across chemical classes. To evaluate the performance of our models across chemical classes, we generated 80 randomized train-test splits (80:20 ratio) from our full dataset of 1924 molecules. We trained each model and evaluated its AUROC performance on the relevant subset of the test split. We report average AUROCs across the 80 randomized train-test splits. Each chemical class subset was constructed by filtering for molecules matching the SMARTS queries below.

Chemical class	**SMARTS query used**
Benzene	c1ccccc1
Ester	[CX3](=O)O[#6]
Carboxylic Acid	[CX3](=O)[OX2H1]
Aldehyde or ketone	[$([#6][CX3](=O)[#6]),$([#6][CX3H1](=O))]
Alkyl (4+ carbon chain)	[CH2X4][CH2X4][CH2X4][CH2X4]
Amine	[NX3;H2,H1,H0;!$(NC=O);!$(NO)]
Organohalide	[#6][#9,#17,#35,#53]
Hydroxyl	[C;!$(C=O)][OX2H]
Ether	[!$(C=O);$([C&!a])][OX2&H0][!$(C=O);#6]
Inorganic	[#6] (Query results were inverted as SMARTS does not support negative matches)
Organosulfide	[SX2][#6]

Estimating number of possible odorants. In order to generate an estimate of the total number of possible odorants, we applied an XGB-trained transport model based on estimated boiling point and log P to a representative set of molecules from the GDB chemical database (24, 25). Estimated boiling point values were used because experimental values are not available for the majority of molecules in the GDB database. The database was binned by heavy atom count (HAC), and each bin was randomly downsampled to contain no more than 10,000 molecules. The final GDB subset included 107,086 total molecules (Dataset S3).

To estimate the possible number of odorous molecules, we need to estimate both the possible number of unique molecules and the proportion of those molecules that will be odorous. The GDB-17 database generated by Ruddigkeit et al (25) includes 166.4 billion molecules with 17 or fewer heavy atoms composed of only C, H, N, O, S, and halogens; this dataset provides a conservative estimate of the number of possible, relevant molecules. In a later study, Ruddigkeit et al (26) found that known fragrance molecules have up to 21 heavy atoms, so our estimate of possible odorants with up to 17 heavy atoms is likely to be an undercount of total possible odorants.

The proportion of molecules predicted to be odorous is heavily dependent on HAC. We calculated an average odorous probability for each HAC from 1 to 17. In our conservative estimate, we simply multiplied the average odorous probability by the number of molecules of that HAC in GDB-17 and summed these values. To show the variability in these estimates, we subsequently used transport model-predicted odorous probabilities to generate 10 binary odorous/odorless classifications for each of the 107,086 molecules in our GDB subset. The standard error bars in Fig. 3a represents the variation in the proportion of odorous molecules by HAC across the 10 simulations. To show the trend in the proportion of predicted odorants with increasing HAC, we fit a logistic regression model (glm function, R) to the binary odor classifications as a function of HAC (Fig. 3a).

Because our model is based on principles of physical transport which apply to all molecules, we anticipate that our model will produce accurate predictions for all GDB-17 molecules. However, to provide a more cautious estimate, we also segmented our down-sampled set of GDB-17 molecules based on the distance to their nearest neighbor in our training data. For each molecule in our training data and GDB-17 subset, we generated bit-based Morgan fingerprints (radius=2, nBits=2048; RDKit) and calculated Tanimoto similarity between each GDB-17 molecule and its nearest training set neighbor. In Fig. S3, we show the cumulative number of predicted novel odorants as a function of ML Transport model-generated odorous probabilities for several Tanimoto similarity cut-offs. Considering only GDB-17 molecules with structurally similar molecules in our training data (Tanimoto similarity > 0.4), we calculate a lower bound estimate of 10 million predicted odorants in GDB-17 with structurally similar known odorants.

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. However, the poor model performance with extreme training set imbalance suggests that balanced training sets like ours gives better estimates of size of odor space.

Visualizing odors in chemical space. A two-dimensional semi-supervised embedding of odor space was created by applying the UMAP algorithm (27) to a random sample (n=107,086 molecules) of the GDB-17 database along with an additional n=8,366 molecules drawn from the experimental olfaction literature. All molecules containing atoms other than H, C, Cl, F, I, N, O, S, P (n=420 molecules) were excluded. N=4,864 physicochemical features were computed using Dragon 6.0 software. Missing feature values (0.06%) were median imputed and then all features were scaled between 0 and 1. The UMAP algorithm was also shown the model-predicted odorous probabilities of a random 25% of the odorants to assist in clustering molecules of similar odor probability. An interactive version of is available at http://odormap.pyrfume.org.

Table S1. Model performance across algorithms and feature sets. For each algorithm, tuning parameters (determined using cross-validation on only the literature-classified molecules) are given, and resampled test set (using only the lab-classified molecules) area under the receiver operating characteristic curve (AUROC) is reported for models trained on either transport features (molecular weight, vapor pressure, boiling point, log P) or many features (>3700 chemoinformatic features calculated using Dragon software). Reported AUROC mean and median values are calculated from 25 random test set draws of 30 lab-tested molecules each; models were trained on all molecules not drawn into the test set.

Algorithm	Tuning Parameters	Feature Set	Test Set Performance AUROC mean ± std (median)
eXtreme Gradient Boosting	Learning Rate=0.1; Number of Estimators=100	Transport	0.975 ± 0.028 (0.985)
		Many-feature	0.974 ± 0.024 (0.977)
Gradient Boosting classifier	Learning Rate=0.1; Number of Estimators=100	Transport	0.980 ± 0.026 (0.994)
		Many-feature	0.974 ± 0.026 (0.976)
Random Forest	Max Depth=3; Number of Estimators=100	Transport	0.990 ± 0.011 (0.995)
	Number of Estimators=100	Many-feature	0.977 ± 0.028 (0.986)

Support Vector classifier	Kernel=rbf;
	C=0.1
		Transport	0.987 ± 0.020 (0.995)

		Many-feature	0.988 ± 0.14 (0.990)
Logistic Regression	C=0.1	Transport	0.981 ± 0.021 (0.986)
	C=0.001	Many-feature	0.963 ± 0.033 (0.973)

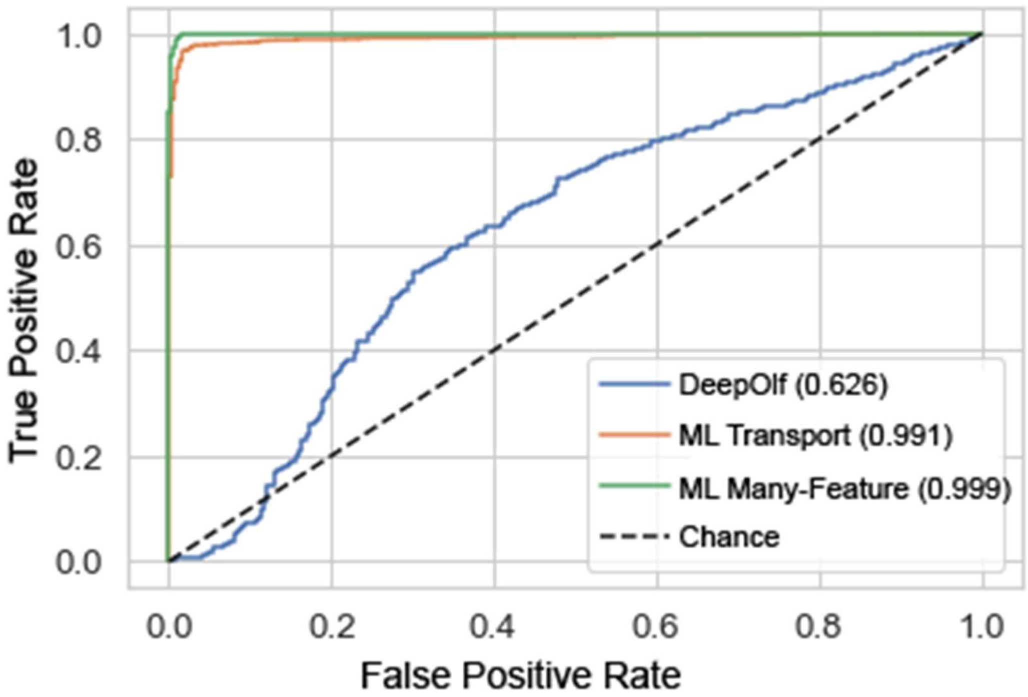

Fig. S1. DeepOlf model underperforms on cleaned dataset. We measured the performance of the DeepOlf model published by Sharma et al (2020)[6] on our full cleaned dataset (n=1924); this plot of true positive rate versus false positive rate shows the receiver operating characteristic (ROC) curves for the DeepOlf, ML Transport, and ML Many-Feature models. Chance classification accuracy is represented as a dotted black line. The area under the ROC curve for each model is reported in parentheses next to the model name.

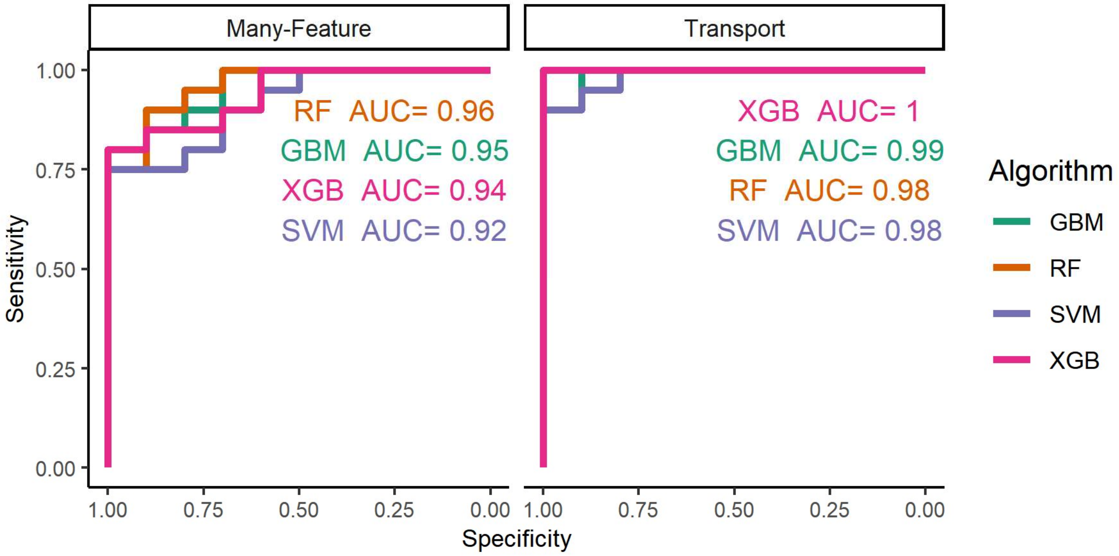

Fig. S2. Performance on the held out test set by various models. The receiver operating characteristics (ROC) curve is plotted for models trained with many features (left) or transport features (right) by four machine learning algorithms (GBM – gradient boosting machine, RF – random forest, SVM – support vector machine, XGB – eXtreme Gradient Boosting), indicated by line color. The area under the curve (AUC) is labeled for each model; models trained with only transport features outperform models trained with many features regardless of algorithm when tested on the original test set (30 molecules: 10 odorless, 20 odorous).

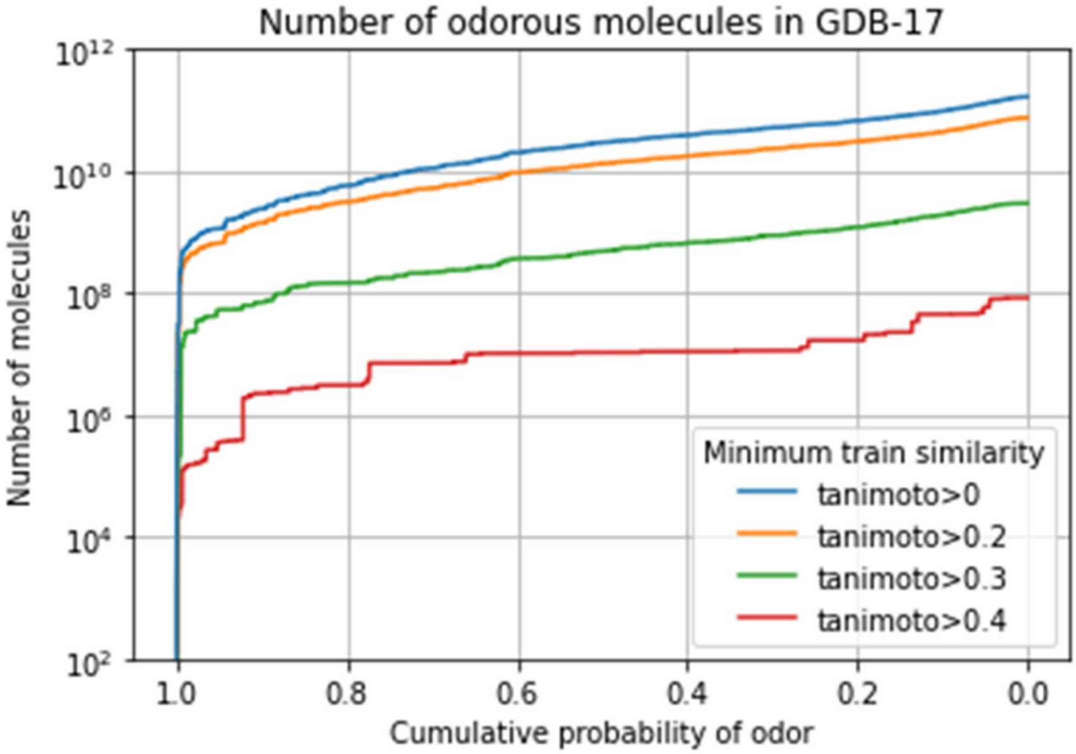

Fig. S3. Possible odorant estimates as a function of nearest training set example. Our transport model predicts 40 billion molecules in GDB-17 will be odorous (probability of having an odor > 0.5). By segmenting GDB-17 molecules by the minimum Tanimoto (structural) similarity to a training set molecule, we show a conservative lower bound estimate of 10 million predicted odorous molecules that have a structurally similar neighbor (Tanimoto similarity > 0.4) in our training data.

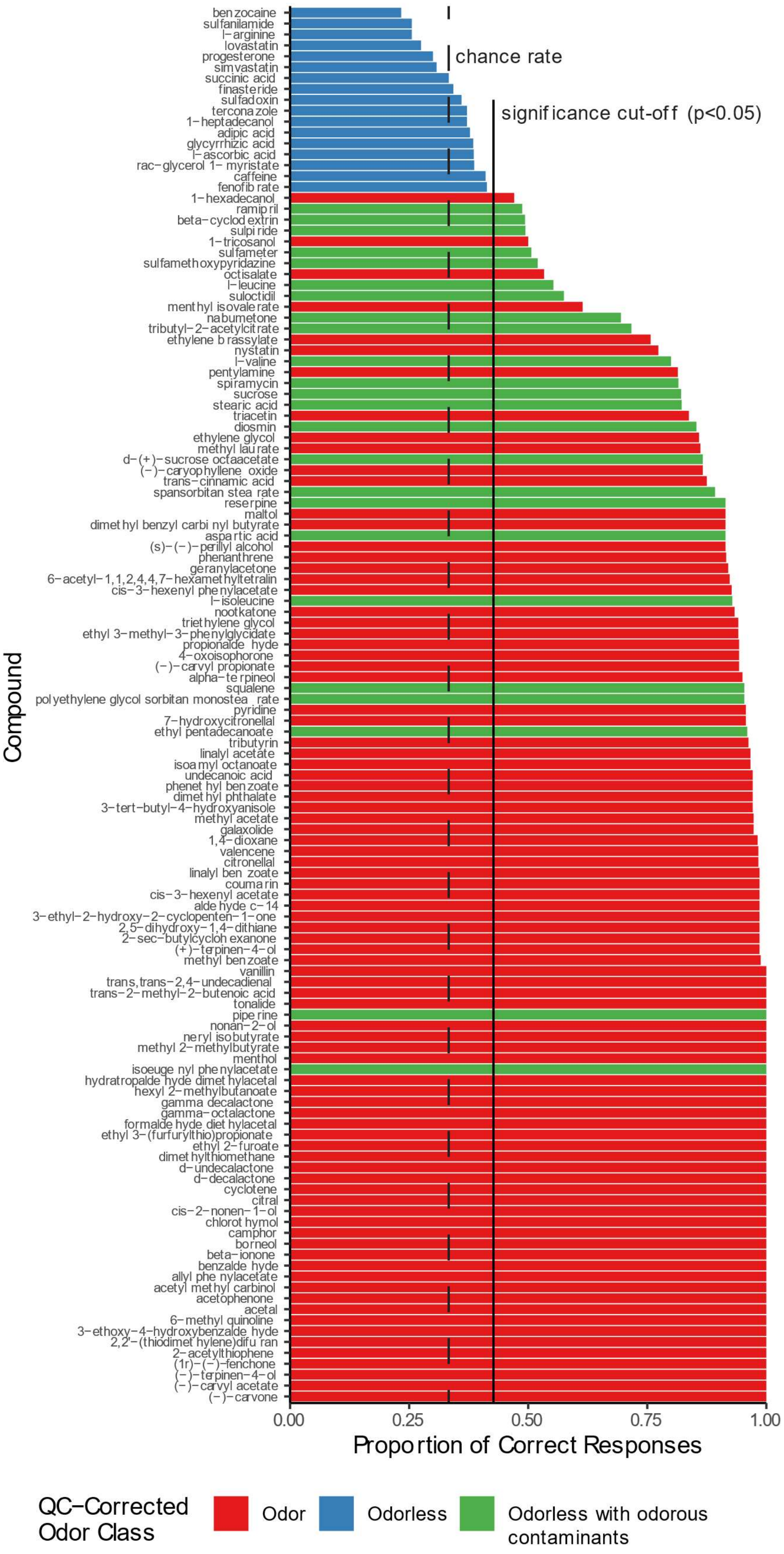benzocaine
sulfanilamide
l−arginine
lovastatin
progesterone
simvastatin
succinic acid
finasteride
sulfadoxin
terconazole
1−heptadecanol
adipic acid
glycyrrhizic acid
l−ascorbic acid
rac−glycerol 1− myristate
caffeine
fenofibrate
1−hexadecanol
ramipril
beta−cyclodextrin
sulpiride
1−tricosanol
sulfameter
sulfamethoxypyridazine
octisalate
l−leucine
suloctidil
menthyl isovalerate
nabumetone
tributyl−2−acetylcitrate
ethylene brassylate
nystatin
l−valine
pentylamine
spiramycin
sucrose
stearic acid
triacetin
diosmin
ethylene glycol
methyl laurate
d−(+)−sucrose octaacetate
(−)−caryophyllene oxide
trans−cinnamic acid
spansorbitan stearate
reserpine
maltol
dimethyl benzyl carbinyl butyrate
aspartic acid
(s)−(−)−perillyl alcohol
phenanthrene
geranylacetone
6−acetyl−1,1,2,4,4,7−hexamethyltetralin
cis−3−hexenyl phenylacetate
l−isoleucine
nootkatone
triethylene glycol
ethyl 3−methyl−3−phenylglycidate
propionalde hyde
4−oxoisophorone
(−)−carvyl propionate
alpha−terpineol
squalene
polyethylene glycol sorbitan monostea rate
pyridine
7−hydroxycitronellal
ethyl pentadecanoate
tributyrin
linalyl acetate
isoamyl octanoate
undecanoic acid
phenethyl benzoate
dimethyl phthalate
3−tert−butyl−4−hydroxyanisole
methyl acetate
galaxolide
1,4−dioxane
valencene
citronellal
linalyl ben zoate
coumarin
cis−3−hexenyl acetate
aldehyde c−14
3−ethyl−2−hydroxy−2−cyclopenten−1−one
2,5−dihydroxy−1,4−dithiane
2−sec−butylcyclohexanone
(+)−terpinen−4−ol
methyl benzoate
vanillin
trans,trans−2,4−undecadienal
trans−2−methyl−2−butenoic acid
tonalide
piperine
nonan−2−ol
neryl isobutyrate
methyl 2−methylbutyrate
menthol
isoeugenyl phenylacetate
hydratropaldehyde dimethylacetal
hexyl 2−methylbutanoate
gamma decalactone
gamma−octalactone
formaldehyde diethylacetal
ethyl 3−(furfurylthio)propionate
ethyl 2−furoate
dimethylthiomethane
d−undecalactone
d−decalactone
cyclotene
citral
cis−2−nonen−1−ol
chlorothymol
camphor
borneol
beta−ionone
benzalde hyde
allyl phenylacetate
acetyl methyl carbinol
acetophenone
acetal
6−methyl quinoline
3−ethoxy−4−hydroxybenzalde hyde
2,2'−(thiodimethylene)difu ran
2−acetylthiophene
(1r)−(−)−fenchone
(−)−terpinen−4−ol
(−)−carvyl acetate
(−)−carvone
Compound
chance rate
significance cut-off (p<0.05)
0.00
0.25
0.50
0.75
1.00
Proportion of Correct Responses
QC−Corrected
Odor Class
🟥 Odor
🟦 Odorless
🟩 Odorless with odorous contaminants

Fig. S4. Classification of molecules as odorous or odorless through human psychophysics experiments and analytical quality control. Blindfolded subjects were presented with 1 compound-containing jar and 2 blank jars in each trial of a 3-alternative forced choice task; the proportion of correct responses is plotted for the 128 lab-tested molecules The chance rate of selecting the correct jar is ⅓, and the minimum statistically significant (a = 0.05) correct selection rate is 0.43 (> 32 correct selections). Molecules correctly differentiated from the blanks in more than 32 trials were initially classified as odorous, plotted in red or green, while molecules correctly selected below this rate were classified as odorless, plotted in blue. A paired gas chromatography-mass spectrometry/gas chromatography-olfactometry (GC-MS/GC-O) quality control (QC) procedure was applied to identify cases in which odorous contaminants, and not the nominal compound, were responsible for the human-detectable odor; these molecules were reclassified as odorless (plotted in green).

Dataset S1 (separate file). Complete dataset used in model training and testing. The dataset includes 1924 molecules, identified by a SMILES string; given for each molecule is an odor classification (odor/odorless), the source of that classification, the train/validation/test set assignment, model-predicted odorous probability, and physicochemical features used by models to generate predictions.

Dataset S2 (separate file). Information on the odor classification of 128 lab-tested molecules. Each molecule is identified by a common name and SMILES string. Table includes number of successful and total trials from the 3-alternative forced choice (3-AFC) experiment, the initial odorous/odorless classification resulting from the 3-AFC experiment, and final classification following GC-MS/GC-O QC.

Dataset S3 (separate file). Subset of molecules from GDB-17 used in enumeration calculations. Each molecule is represented by a SMILES string, and for each molecule is given the heavy atom count (HAC) and the transport ML model predicted odorous probability.

SI References

1. H. Boelens, Structure-activity relationships in chemoreception by human olfaction. *Trends Pharmacol. Sci.* **4**, 421–426 (1983).
2. M. H. Abraham, R. Sánchez-moreno, J. E. Cometto-muñiz, W. S. Cain, An algorithm for 353 odor detection thresholds in humans. *Chem. Senses* **37**, 207–218 (2012).
3. K. M. Hau, D. W. Connell, Quantitative structure-activity relationships (QSARs) for odor thresholds of volatile organic compounds (VOCs). *Indoor Air* **8**, 23–33 (1998).
4. S. Arctander, *Perfume and Flavor Chemicals* (Allured Publishing Corporation, 1969).
5. P. Laffort, C. Gortan, Olfactory properties of some gases in hyperbaric atmosphere. *Chem. Senses* **12**, 139–142 (1987).
6. L. J. van Gemert, *Odour Thresholds*, 2nd Editio (Oliemans Punter & Partners BV, 2011).
7. N. M. O'Boyle, *et al.*, Open Babel: An Open chemical toolbox. *J. Cheminform.* **3**, 1–14 (2011).
8. 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).
9. 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).
10. V. Gulshan, *et al.*, 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. S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in Data Mining: Formulation, Detection, and Avoidance. *Knowl. Discov. Data* **6** (2012).
12. A. Barla, *et al.*, Machine learning methods for predictive proteomics. *Brief. Bioinform.* **9**, 119–128 (2008).
13. M. Kuhn, Building predictive models in R using the caret package. *J. Stat. Softw.* **28** (2008).
14. Z. R. Yang, Biological applications of support vector machines. *Brief. Bioinform.* **5**, 328–338 (2004).
15. E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **43**, 1882–1889 (2003).
16. A. Keller, *et al.*, Predicting human olfactory perception from chemical features of odor molecules. *Science (80-.).* **355**, 820–826 (2017).
17. G. Cano, *et al.*, Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Syst. Appl.* **72**, 151–159 (2016).
18. V. Svetnik, *et al.*, Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
19. J. Lu, *et al.*, Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine. *Biochim. Biophys. Acta - Gen. Subj.* **1860**, 2664–2671 (2016).
20. V. Svetnik, *et al.*, Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **45**, 786–799 (2005).
21. R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, E. M. Gifford, Extreme gradient boosting as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **56**, 2353–2360 (2016).
22. I. Babajide Mustapha, F. Saeed, Bioactive molecule prediction using extreme gradient boosting. *Molecules* **21**, 1–11 (2016).
23. Nitesh V. Chawla, K. W. Bowyer, L. O. Hall, P. W. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
24. L. C. Blum, J. L. Reymond, 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
25. 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).
26. L. Ruddigkeit, M. Awale, J. L. Reymond, Expanding the fragrance chemical space for virtual screening. *J. Cheminform.* **6**, 1–12 (2014).

27. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).