Chemical Senses, 2021, Vol 46, 1–13
doi:10.1093/chemse/bjab007
Original Article
Advance Access publication 26 February 2021

OXFORD

Original Article

A System-Wide Understanding of the Human Olfactory Percept Chemical Space

Joel Kowalewski[1], Brandon Huynh[2] and Anandasankar Ray[1,2],

[1]Interdepartmental Neuroscience Program, University of California, 3401 Watkins Drive, Riverside, CA 92521, USA and [2]Department of Molecular, Cell and Systems Biology, University of California, 3401 Watkins Drive, Riverside, CA 92521, USA

Correspondence to be sent to: Anandasankar Ray, Department of Molecular Cell and Systems Biology, University of California Riverside, 3401 Watkins Drive, Riverside, CA 92521, USA. e-mail: anand.ray@ucr.edu

Editorial Decision 16 February 2021.

Abstract

The fundamental units of olfactory perception are discrete 3D structures of volatile chemicals that each interact with specific subsets of a very large family of hundreds of odorant receptor proteins, in turn activating complex neural circuitry and posing a challenge to understand. We have applied computational approaches to analyze olfactory perceptual space from the perspective of odorant chemical features. We identify physicochemical features associated with ~150 different perceptual descriptors and develop machine-learning models. Validation of predictions shows a high success rate for test set chemicals within a study, as well as across studies more than 30 years apart in time. Due to the high success rates, we are able to map ~150 percepts onto a chemical space of nearly 0.5 million compounds, predicting numerous percept–structure combinations. The chemical structure-to-percept prediction provides a system-level view of human olfaction and opens the door for comprehensive computational discovery of fragrances and flavors.

Key words: flavors, fragrances, machine learning, olfaction, prediction

Introduction

Human perceptual descriptions for olfactory stimuli are less stereotypic than for vision or auditory stimuli where perception can be predicted by clearly defined properties such as wave frequency. In fact, olfactory perception may vary without an apparent relationship to the physicochemical properties of an odorant nor the molecular and cellular organization of the olfactory system (Buck and Axel 1991; Vassar et al. 1993; Mombaerts et al. 1996; Mombaerts 1999, 2001). Yet general neuroanatomical olfactory pathways are well conserved across species and the olfactory capabilities of humans appear to be close to that of species that rely heavily on olfaction for survival and mating (McGann 2017). Genetic variation in olfactory receptors also explains a significant amount of variability in basic perceptual qualities like intensity as well as more complex perceptual qualities (McRae et al. 2012; Mainland et al. 2014; Trimmer et al. 2019). Although culture and language also affect olfactory perception (Majid and Kruspe 2018), individuals often show significant similarities in perceptual descriptions for the same chemical (Dravnieks 1985; Keller and Vosshall 2016), implying an underlying physicochemical basis for human olfactory perception. In fact, predicting percepts from physicochemical features is becoming increasingly plausible (Khan et al. 2007; Haddad et al. 2010; Snitz et al. 2013; Nozaki and Nakamoto 2016; Keller et al. 2017; Kepple and Koulakov 2017; Licon et al. 2019). However, the breadth and complexity of the human olfactory perceptual space and its physicochemical correlates remain poorly understood except for a select few (<10) perceptual descriptors (Keller et al. 2017). Moreover, because of the comparatively limited repertoire of olfactory receptors that have been functionally deorphanized (Saito et al. 2004; Keller et al. 2007; Mainland et al. 2014; Shirasu et al. 2014; de March et al. 2020; Hu et al. 2020; McClintock, Khan, et al. 2020; McClintock, Wang, et al. 2020; Pfister et al. 2020), predictive models with receptors are presently not as comprehensive in mapping to different perceptual qualities. There is subsequently an important role for computational modeling.

© The Author(s) 2021. Published by Oxford University Press. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

Previous attempts to predict ratings of odor perception from the physiochemical features of molecules have been successful to some degree, although these examples represent a small fraction of the perceptual descriptor space (Khan et al. 2007; Nozaki and Nakamoto 2016; Keller et al. 2017). In these previous efforts, several perceptual descriptors tested were hard to predict and these descriptors may have been difficult to evaluate by study volunteers or lack a strong physicochemical basis. Nevertheless, a natural language processing approach could successfully predict perceptual descriptors across studies, suggesting that descriptions of olfactory perceptual content are likely structured and not totally subjective (Gutiérrez et al. 2018). These earlier studies create an underlying framework that points to the intriguing possibility that the perceptual descriptions humans select to characterize odorants are associated with key physicochemical features, even those that are seemingly abstract and currently not well defined. Although prior structure–activity studies predate modern machine learning, indicating features enriched among chemicals with shared perceptual qualities (Rossiter 1996), some exceptions arise.

A few recent studies have modeled odor perception using large databases, training deep neural networks to predict perceptual descriptors from chemical features (Sanchez-Lengeling et al. 2019; Tran et al. 2019). These studies have suggested that many complex perceptual descriptors are predictable. The chemical representation or input in these studies undergoes modification, in which the predictive features are the weights of the neurons in the neural network, making them challenging to interpret. We have established a pipeline to clarify the physicochemical properties that best predict diverse perceptual descriptors and to rigorously test using different metrics and controls that ensure the machine-learning models are consistent with biological expectations. We find that chemical feature models can address many complex, biologically relevant tasks. As this suggests the important or predictive features that we identify are a resource for further research, we finally annotate a large commercially available chemical database with predicted odor qualities. These predictions reveal enriched structural motifs that help interpret the machine-learning models.

Materials and Methods

Psychophysical data

Keller (2016) study

We used data from 55 general public volunteers (Keller and Vosshall 2016) for external validation (Figures 1 and 2, Supplementary Figures 1–5). Due to limited diversity in the selection of odor descriptors supplied by naive volunteers and evidence indicating experience with odor language improves the quality of perceptual data (Lawless 1984; Dubois and Rouby 2002; Olofsson and Gottfried 2015), we primarily considered a sample of industry professionals as reported in the atlas of odor character profiles (Dravnieks 1985). Notably, the semantic descriptors (odor characters or perceptual descriptors) were sparsely used in some cases among the general public volunteers, suggesting that averaged ratings for a given descriptor (odor character) might represent a very small proportion of the respondents. This becomes particularly important for generating predictive models since missing data points (e.g., chemicals or odorants that are not rated by some participants) must be dealt with such as by averaging ratings for the nearest neighboring (k) odorants or filling-in with the median/mean rating across all odorants. Although these approaches are valid in predictive modeling, they are a significant modification of the respondent data; the failure to provide a rating is a potentially important source of information. We maintained, as a result, the 0–100 scale for the general public volunteer data but converted ratings to a % usage metric instead. Dilution was not considered, averaging % usage over the different dilutions. In preliminary analyses, there was however some evidence that models might benefit from training on data from a single dilution. Similarly, a small number of replicates that were performed in this study were not included in the final training and testing data sets.

Although with the % usage each odorant is assigned numeric values more naturally, this modification was also in line with the Dravnieks (1985) study data. The % usage therefore provided a means to compare 2 sources that to a first approximation appear very different. Dravnieks (1985) also reports a percent applicability metric. The percent applicability is the sum of the ratings for a chemical or odorant over all participants divided by the maximum possible sum. This was not used for our cross-study comparisons as ratings from an experienced participant panel might scale differently and the sample size between the 2 studies is very different. Because cross-study comparisons are not well defined, we opted for the simplest possible metric, the % usage.

Atlas of odor character profiles, Dravnieks (1985)

Dravnieks (1985) summarizes odor profiles for 180 odorants, replicates, and mixtures, with the latter not being used for predictions, from 507 industry professionals in total across 12 organizations. Each chemical was rated by between 120 and 140 participants. The participants scored a set of replicates, which were used to provide an index of discriminability for the data as the inverse of the squared correlation coefficient between replicates (RV). For this study, RV = 0.11. The scoring metric was on the range of 1–5 with 1 being slightly and 5 being extremely relevant. Raw scores were subsequently processed into 2 numeric values summarizing the participants' responses. We only focused on the % usage; the fraction of participants providing any response, 1–5 because it is the simplest metric to interpret and relate to other studies. The perceptual

Software and data resources

Type	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	R 3.5.1	https://github.com/tidyverse/ggplot2		ggplot2 (R)
Software, algorithm	R 3.5.1	https://github.com/igraph/igraph		igraph (R)
Software, algorithm	R 3.5.1	https://github.com/tgirke/ChemmineR		ChemmineR (R)
Software, algorithm	R 3.5.1	https://github.com/topepo/caret		caret (R)
Data	Keller (2016)	https://doi.org/10.1186/s12868-016-0287-2		
Data	GoodScents	http://www.thegoodscentscompany.com/index.html		
Data	DREAM	https://github.com/dream-olfaction/olfaction-prediction		407 Train, 69 Test chemical IDs for Keller (2016) data
Data	Dravnieks (1985)	http://doi.org/10.1520/DS61-EB		Data from the original 1985 edition

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

descriptor (or character) set available for the Dravnieks (1985) study was extensive but empirically driven. Recommendations from the American Society for Testing and Materials (ASTM) sensory evaluation committee winnowed an initial set of 800 possible odor characters (perceptual descriptors) for sensory analyses down to 160. Prompted by additional research, this figure was later revised to 146 relevant perceptual descriptors, a final set that addressed concerns in which clear perceptual differences could result in identical descriptor usage from study participants. This final set of 146 perceptual descriptors and the percent usage was subsequently prepared for machine-learning analyses.

GoodScents test data

GoodScents is a database of 2000+ chemicals, containing basic physicochemical information as well as perceptual descriptor labels, if available, from published reference materials. Since it is not possible to predict a descriptor for which there is no Dravnieks (1985) equivalent, we had to define exclusionary criteria to properly evaluate the predictions. This included in addition to removing nonunique chemicals those without descriptor labels matching or similar to Dravnieks (1985), leaving 2525 chemicals for test set validation. Examples of similar descriptors in GoodScents include "weedy" and "nutty," which correspond with "crushed weeds," and "walnut" and "peanut butter in Dravnieks (1985), respectively." The 146 Dravnieks (1985) descriptor models assigned a probability score. Receiver operating characteristic (ROC) curves were subsequently computed using the observed descriptor labels for each of the 2525 chemicals. A chemical described simply as "nutty," for example, is expected to have high probabilities for "peanut butter" and "walnut," but not for "orange" and "chemical." Cases where descriptors were correlated in Dravnieks (1985) (>0.85) were also defined as a set to avoid overly penalizing the assignment of redundant descriptors to new chemicals. We identified earlier that models with this level of correlation are often interchangeable, with only a non-significant reduction in prediction performance. Namely, machine-learning assignment of "Chemical" to an odorant described as "Varnish" was not incorrect given the data. The ROC assesses that high probabilities are correctly assigned to the observed descriptors. When the machine-learning models predict descriptors that are unlike those observed, the area under the ROC curve decreases. An independent *t*-test comparison was made between actual areas under the ROC curve (AUCs) and those using random probability scores.

Selecting optimally predictive chemical features

Optimizing chemical structures

Chemical features were computed with DRAGON 6 for Dravnieks (1985) (Figures 1A and 2A). Chemical structures were optimized and 3D coordinates computed with OMEGA. Molecular or chemical features were precomputed and made publicly available for the DREAM study, and these data files were used as is for analysis of the 55 public volunteers reported in the Keller's (2016) study.

Chemical feature ranking and importance

Cross-validated recursive feature elimination

Recursive feature elimination (RFE) iteratively selects subsets of features to identify optimal sets. The algorithm is a "wrapper" and therefore relies on an additional algorithm to supply predictions and quantify importance. Often this is a decision tree such as random forest, which was used here, since the algorithm computes feature importance internally. This distinction between internal and external simply means that although any arbitrary algorithm can supply the prediction error—here, the error in predicting the % usage value—many lack a well-defined method for quantifying feature importance. Feature importance and ranking must, in these instances, be supplied externally such as by nonlinear regression models for each predictor and outcome compared with a constant.

Including cross-validation with the RFE partitions the training data into multiple folds. This step avoids biasing performance estimates but results in lists of top predictors over the cross-validation folds such that importance of a predictor is based on a selection rate.

Random forest

Random forest is an extension of basic decision trees that overcomes the often-poor generalizability of these models by aggregating the predictions from multiple trees trained on bootstrap samples and different predictor sets, effectively limiting redundancy between trees. Rows that are excluded as part of bootstrapping process are used to estimate prediction performance on new data. This also provides a method for assigning importance to features through randomization; the % increase in prediction error after randomizing a feature is accordingly the ranking metric that was used for tabulating chemical feature importance (shown in Supplementary File 1).

Selection bias

Selecting features or predictors on the same dataset used for cross validation results in models that have already "seen" possible partitions of the data and therefore performance metrics will be biased. Selection bias (Ambroise and McLachlan 2002) was addressed by bootstrapping and cross-validation, which ensure some separation between predictor/feature selection and model-fitting/validation. In addition to these methods, we used hidden test sets and also showed that the models could be used to predict perceptual responses from a completely different experiment, removing methodological biases arising from odorant preparation and presentation or any unforeseen regularities that machine-learning algorithms could exploit but that are fundamentally task irrelevant for the analyst or researcher interested understanding rather than predicting.

Selecting optimal machine-learning algorithms

The support vector machine (SVM) with the radial basis function (RBF) kernel outperformed random forest, regularized linear models (ridge and lasso), and linear SVM, tuning over L1 versus L2 regularization (Figure 1). However, gradient boosted decisions trees and tree ensembles such as random forest nevertheless approximated performance of RBF SVMs on the public volunteer data (Keller 2016), which was used in part for the DREAM analysis, and in certain cases outperformed it. This emphasizes that the optimal algorithm is context dependent. To ensure consistency in our analysis of different psychophysical data sources, we did not report the results in this manner, that is, fitting the best-performing algorithm each time. We instead aggregated multiple SVM models to improve generalizability. Algorithm selection and training was done using the R package, caret (classification and regression training) (Kuhn 2008; R Development Core Team 2016).

Cross-study predictions

For cross-study predictions, models were fit as shown in the Figure 1a pipeline with Dravnieks (1985) data (Figure 1 and Supplementary Table 2). Multiple SVM models were fit with slightly different chemical features and their predictions were aggregated. This ensemble approach limits the tendency to overfit during the training phase.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

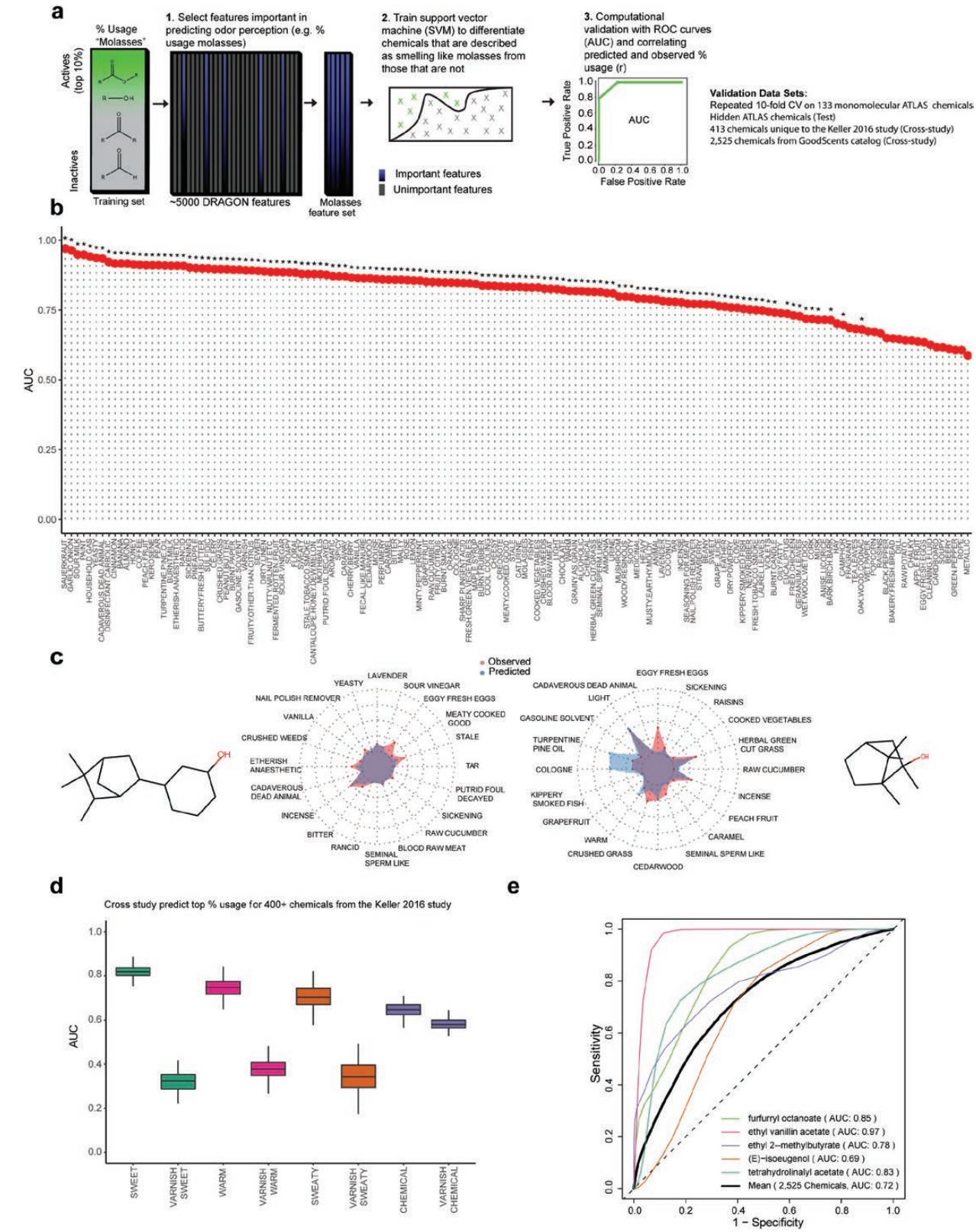

Figure 1. Predicting perceptual descriptors from physicochemical features using machine learning. (**a**) Pipeline for predicting Dravnieks (1985) ratings (% Usage) for perceptual descriptors, an example is provided for the descriptor, "molasses." The important chemical features are detected that predict "Molasses." An SVM is fit, and the predictions are assessed by different methods such as the AUC from ROC plots. (**b**) Chemicals within the top 10% of ratings (% Usage) are labeled as "Active." The AUC quantifies the relationship between sensitivity to the actives (chemicals in the top 10% ratings) versus false positives. Bars in the plot represent the average AUC from 3 models with different chemical features. The AUC is computed on chemicals excluded from training (30 times, 10-fold cross-validation repeated 3 times). Significance (*) is determined by one-sided *t*-test, comparing the AUC to an identical model trained on shuffled "Active" labels.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

Notably, chemicals do overlap between the 2 studies. Removing these chemicals (58) from Dravnieks (1985) significantly reduces the available training data. We instead removed the overlap from the Keller 2016 data set, leaving 413 chemicals as a test set. Although theoretically all 146 perceptual descriptors could be assessed, the choice of "warm," "sweaty," "sweet," and "chemical" depended on key differences in the perceptual descriptors available for the 2 studies, Keller (2016) and Dravnieks (1985). For instance, although Dravnieks (1985) used word strings in many cases such as "putrid, foul, decayed" to provide greater context, Keller 2016 opted for "decayed." It is unclear what affect this difference might have and if it is nontrivial. The interpretation of the cross-study prediction becomes ambiguous as a result. Identically presented descriptors, like "chemical," "warm," "sweaty," and "sweet" are well-defined cases for testing models across studies.

Network analyses and visualizations

Matrices for network

Chemical and perceptual descriptor relationships were modeled as bipartite graphs from an incidence matrix with perceptual descriptors as rows and columns the combined, unique optimal chemical feature sets (Figure 2). The optimal feature sets are from iteratively fitting a random forest model on 100 different partitions of the Dravnieks (1985) training data. We ranked the features based on the random forest importance over the partitions. Several different perceptual descriptor–chemical feature matrices were assembled by varying the number of ranked features per descriptor (e.g., top 3, 5, 10). Incidence matrices from the top 3, 5, or 10 chemical features are therefore identical except for the number of columns (unique chemical features). Factor analysis was performed to reduce the number of perceptual descriptors for clarifying network plots as in Figure 2b. This was run using the factanal function in addition to functions in the nFactors (Raiche 2010) R package for factor extraction.

Specifically, values in the incidence matrices are 1 or 0; the optimal chemical features for each perceptual descriptor are 1, otherwise 0. This amounts to a sparse matrix with the nonzero values, indicating relationships among the optimal physicochemical features and the perceptual descriptors. Collectively, these binary strings are likened to a set of combinatorial chemical feature codes for the Dravnieks (1985) perceptual descriptors. We subsequently separated the bipartite graph for clarity into its constituent, adjacency matrices, which are symmetrical, $m \times m$ and $n \times n$, matrices, with m denoting rows (perceptual descriptors) and n the columns (chemical features) in the original incidence matrix. An adjacency matrix can be obtained by multiplying an incidence matrix by its transpose.

Clustering networks

Several methods are available for identifying modules, communities, or clusters in networks assembled from adjacency matrices. We tested several, selecting the Louvain algorithm based on its higher modularity score for Dravnieks (1985) data. Actual or observed network properties were in turn compared to 10 000 random network simulations (Erdos-Renyi) of approximately identical size and density. The actual network properties differed from those generated through the random simulation.

Tools for network analysis and visualization

Graph analyses were done using the igraph package (Csardi and Nepusz 2006) in R, plots with ggplot2 (Wickham 2016), and functions from the ggnetwork package for visualizing the networks.

eMolecule predictions and network representation

The eMolecule predictions are from Dravnieks descriptor models trained on the % usage (0–100 ratings), with detailed performance in Supplementary File 2, Supplementary Figure 1a, Supplementary Figure 2b, and Figure 3. The regression-based models predict or estimate these ratings for the eMolecules chemicals. Because the Dravnieks training set is not structurally exhaustive, we applied 2 filters to further sort the predictions. These include (1) an atom pair fingerprint based on commonly occurring feature sets in biologically active compounds (Cao et al. 2008) and (2) the % usage values of the chemicals at the top end of the distribution (% usage). Initially, the % usage values for the top chemicals (exemplars) per descriptor were applied to filter the predictions. For each descriptor, the reduced set was then compared with the physicochemical features of the exemplar chemicals using atom pair fingerprints. Since atom pairs are a coarse representation of complex 3D molecules, we applied a Tanimoto similarity coefficient threshold of 0.25. This ensured that predictions per descriptor displayed basic 2D features that overlapped with the Dravnieks exemplar chemicals, while exploring new structural patterns or motifs that are potentially missed in 2D comparisons. Notably, projecting from a small chemical training set to a larger chemical set potentially amplifies noise in the training data. We would therefore recommend the table of top chemical features in Supplementary File 1 if interested in a less exploratory resource.

Enriched substructures/cores

Enriched cores were analyzed using RDKit through Python (Van Rossum and Drake 1995; Landrum 2006) (Figure 4). The algorithm performs an exhaustive search for maximum a common substructure among a set of chemicals. In practice, larger sets often yield less substantive cores. To remedy this, the algorithm includes a threshold parameter that relaxes the proportion of chemicals containing the core. We used a threshold of 0.5, requiring that half of the chemicals from the top 10 contained the core.

Natural language processing (Supplementary File 3)

Though key differences exist across computational or machine-learning studies of odor perception, it is particularly important to identify the challenges and strengths of different approaches as well as to provide an overview of the perceptual qualities that appear easy or difficult to predict. Natural language processing libraries, while not optimized for odor language, do offer an initial bridge between studies including diverse and different perceptual descriptors. To that end, we used the spaCy library (Van Rossum and Drake

The number of "Active" labels remains unchanged. Significance threshold set at $P \leq 0.05$ after adjusting for false discovery rate. (**c**) Predicted versus observed % usage for select test chemicals. For clarity, only a selection of perceptual descriptors is shown. See Supplementary File 2 for additional detail and chemicals. (**d**) Dravnieks (1985) trained models of "Sweet," "Warm," "Sweaty," and "Chemical" predict ratings for these same descriptors from a study of public volunteers for 413 test chemicals (Keller and Vosshall 2016). Cutoffs to convert the public volunteer data into actives are from the Dravnieks (1985) study (top 10% usage). Significance is determined by one-sided *t*-test, comparing the perceptual descriptor models with a nonidentical but top-performing Dravnieks (1985) model, "Varnish" over 100 bootstrap samples. Public volunteer perceptual data are averaged over dilution. (**e**) Average prediction performance (AUC) when assigning 1–146 Dravnieks (1985) perceptual descriptor labels to 2525 test chemicals with known labels in the GoodScents database. Performance metrics other than classification (AUC) in Supplementary File 2. Formal definitions for the performance metrics and the SVM algorithm are provided in Materials and methods.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

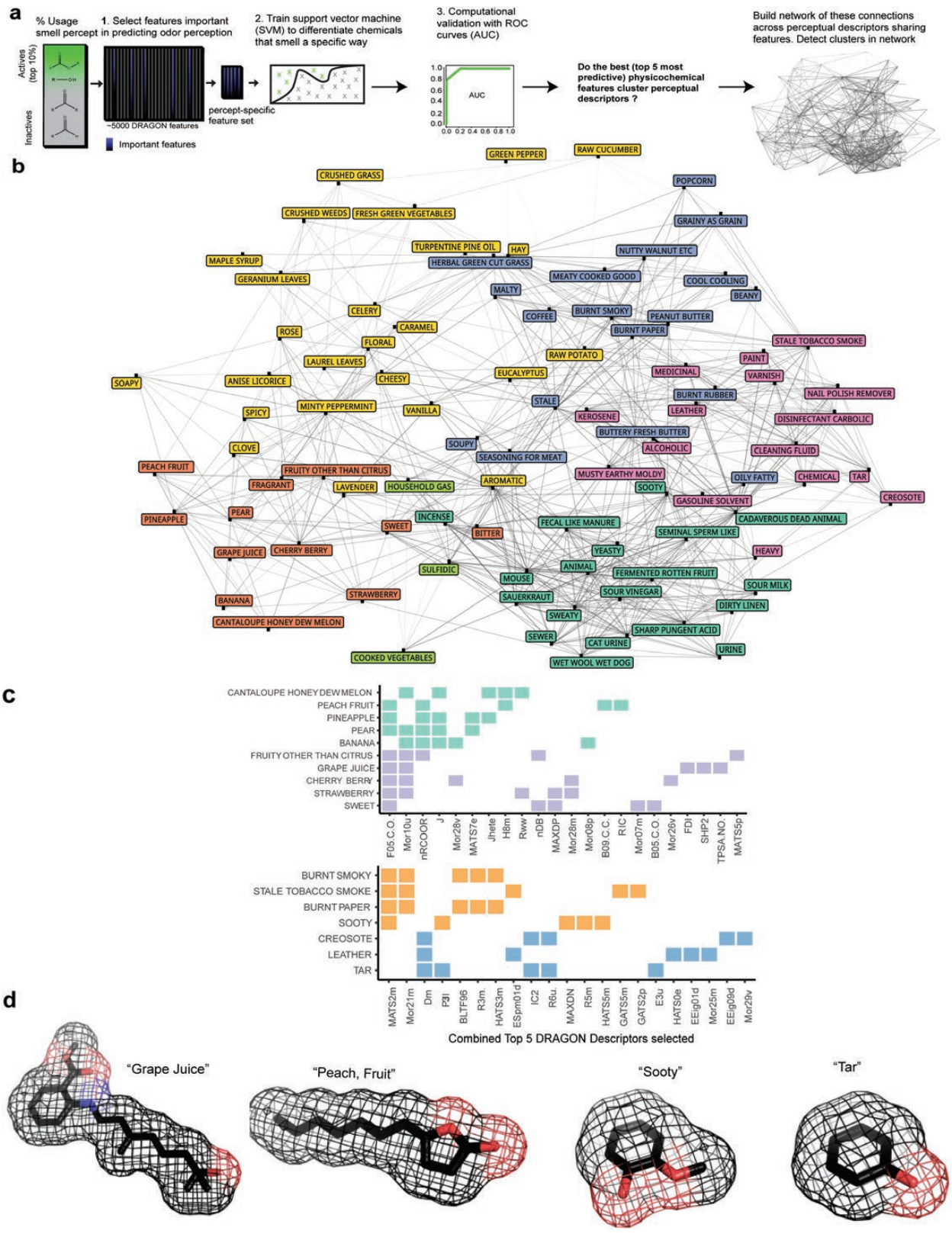

Figure 2. Building perceptual descriptor networks from few physicochemical features. (**a**) Pipeline summarizing methods for selecting the most important chemical features for predictions of Dravnieks (1985) perceptual descriptors, followed by the construction of networks that help visualize relationships among these descriptors when considering physicochemical information alone. (**b**) Assembled network from the top 5 chemical features per descriptor. Descriptors with shared top 5 chemical features are connected in the network. Similar perceptual descriptors are color-coded based on the Louvain algorithm. (**c**) Two sets of correlated descriptors are analyzed based on the chemical features that are important (among the top 5) for predicting them. (Top) Matrix 1: "fruity" descriptors. (Bottom) Matrix 2: "sooty" descriptors. Louvain clustering (square color) shows the similar descriptors are separable into 2 subgroups. Filled-in squares, regardless of color, represent the importance of the labeled chemical feature. (**d**) Exemplar chemicals from the computationally inferred clusters.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

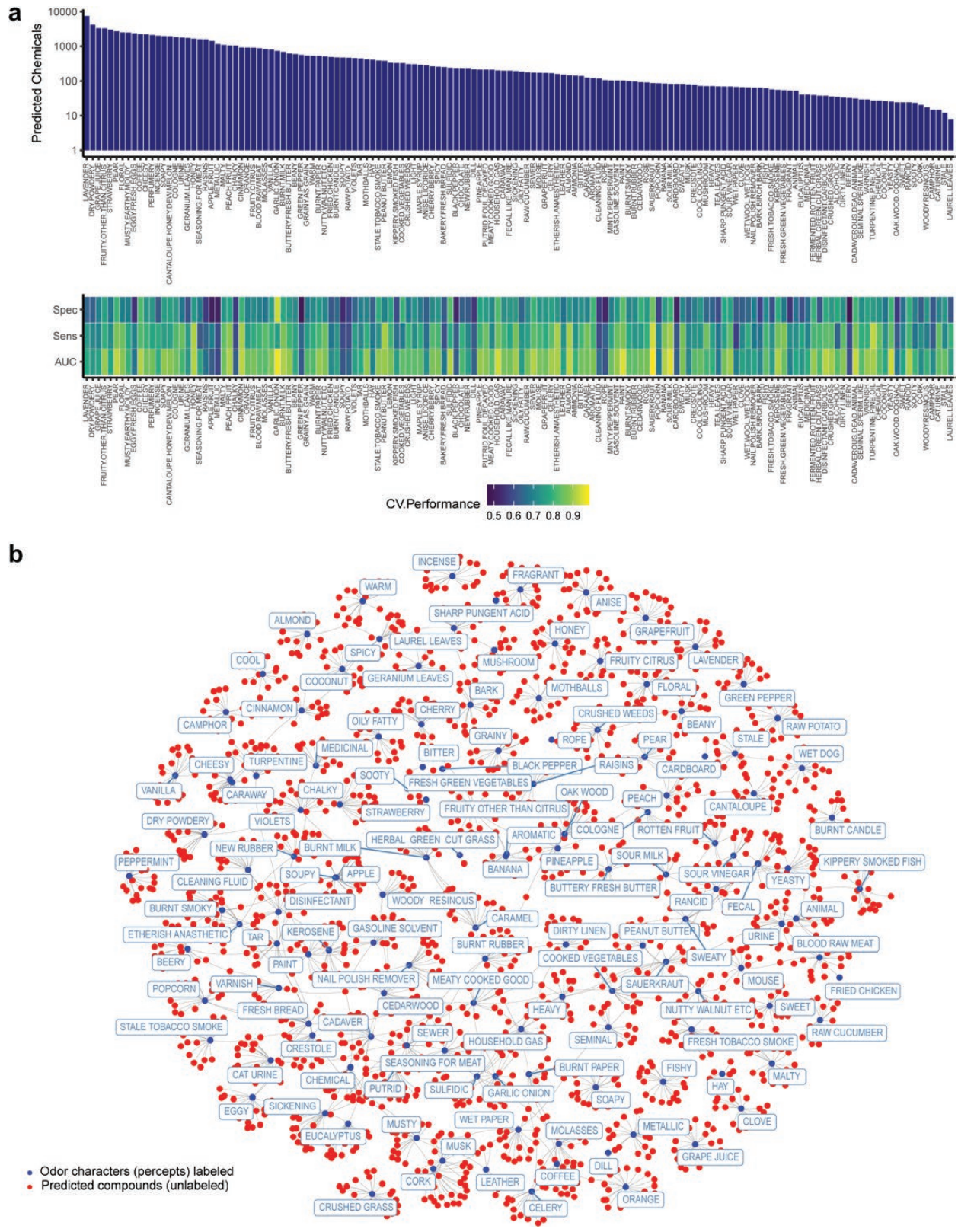

Figure 3. Predicting and mining large commercially available chemical spaces. (**a**) The machine-learning models are used to predict perceptual descriptors from ~440 000 compounds. (Top) predicted chemical counts are based on optimal thresholds from the ROC curves and structural similarity (atom pair similarity > 0.25) to training actives. An optimal threshold is the point on the curve that minimizes false positives and maximizes true positives. (Bottom) Detailed validation for the models ordered with respect to the number of predicted chemicals. (**b**) A 2D representation of predictions for 15 hits for each perceptual descriptor (or all chemicals that exceed the % usage threshold for actives), with edges connecting compounds that are predicted for multiple descriptors. The newly predicted chemicals are indicated as unnamed red dots, and each descriptor as blue dots and labeled in rectangles. Predictions are from the SVM algorithm with a radial basis function (RBF) kernel. See Materials and methods for additional information.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

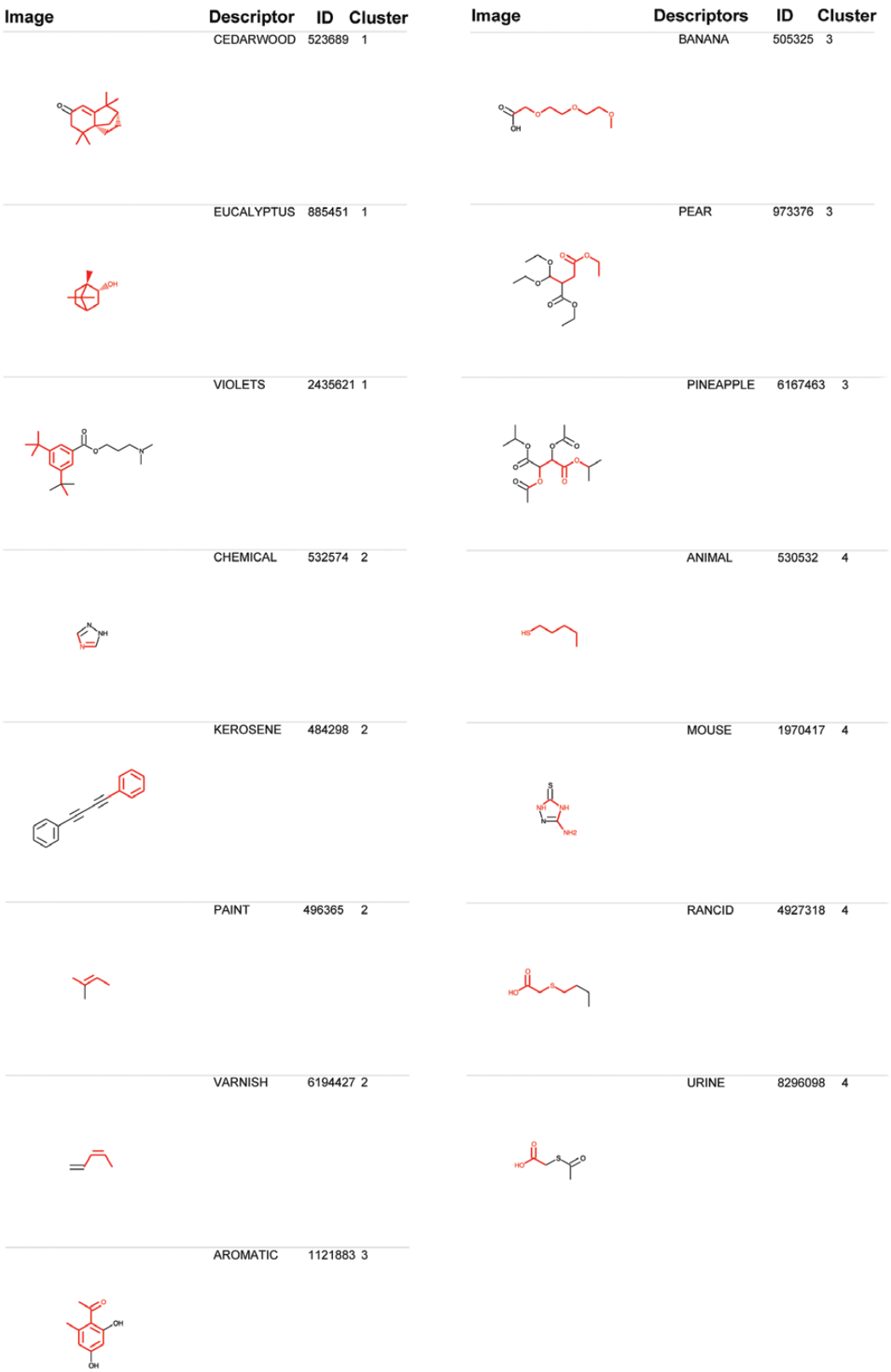

Figure 4. Enriched chemical features among predictions. (**a**) Top predicted chemicals in eMolecules from the Figure 3 network are clustered and analyzed for common structural features (substructures or cores). These are highlighted (red) in images of representative chemicals from the predictions. The ID is the eMolecules identifier. Simple structural features are common among predicted chemicals, enabling basic comparisons between different perceptual descriptors based on chemical structure. Accordingly, this is an example of how a large network of predictions can offer additional insight. See Materials and methods for details on the maximum common substructure algorithm for identifying the enriched features.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

1995; Honnibal et al. 2020) and a convolutional neural network previously trained on GloVe Common Crawl (Pennington et al. 2014) and OntoNotes 5. The training set comprised more than 1 million English text words. The network uses the high dimensional training data to learn a lower dimensional space that represents syntactic and semantic associations in the training texts or documents. New words are vectors that are projected into this space, enabling estimates of semantic or syntactic features. Here, we compared pairs of different odor or perceptual descriptors, generating all pairwise similarities. The similarity coefficient for word vectors is the Cosine similarity. The inverse of the coefficient is a distance; the resulting distance matrix was hierarchically clustered using the Ward D2 method in R.

Support vector machine

Training the SVM involves identifying a set of parameters that optimize a cost function, where cost 1 and cost 0 correspond to training chemicals labeled as "Active" and "Inactive," respectively (Figure 1). θ^T is the scoring function or output of the support vector machine. If the output is ≥0, the prediction is "Active." The function (f) is a kernel function.

$$\text{SVM cost} = \min_{\theta} C \sum_{i=1}^{m} y^{(i)} cost_1(\theta^T f^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

The kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The RBF or Gaussian kernel enables the learning of more complex, nonlinear boundaries. It is therefore well suited for problems in which the biologically active chemicals cannot be properly classified as a linear function of physicochemical properties. This kernel computes the similarity for each chemical (x) and a set of landmarks (l), where σ^2 is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals ("Active" vs. "Inactive").

$$\text{Gaussian} \quad \text{kernel} = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

Model performance metrics

The AUC assesses the true positive rate (TPR or sensitivity) as a function of the false-positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score (x) is greater than the threshold (T), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at AUC = 0.5.

$$TPR\,(T) = \int_{T}^{\infty} f_1\,(x) \quad dx$$

$$FPR(T) = \int_{T}^{\infty} f_0\,(x) \quad dx$$

where T is a variable threshold and x is a probability score.

However, we generated classifiers that are more authentic than theoretical random classification, shuffling the chemical feature values in the models and statistically comparing the mean AUCs across multiple partitions of the data. This controls against optimally tuned algorithms predicting well simply because of specific predictor attributes (e.g., range, mean, median, and variance) or models that are of a specific size (number of predictors) performing well even with shuffled values. Additionally, biological data sets are often small, with stimuli or chemicals that—rather than random selection—reflect research biases, possibly leading to optimistic validation estimates without the proper controls. We used the AUC with classification-based training, such as to predict binary labels (Active/Inactive). For classification-based training we initially converted the % usage into a binary label (Active/Inactive) using the top 10% of the distribution as the cutoff. To provide additional context, we showed performance estimates varying the cutoff as well. The basis for a classification-based performance metric was the often top-heavy distribution of the % usage. It is for instance possibly not as relevant for models to accurately predict chemicals with minimal % usage. Rather, it is preferable for models to accurately predict whether a chemical will smell "Sweet" or not.

To provide further clarity, we also reported multiple performance metrics including the correlation between the predicted and observed % usage, the root mean squared error (RMSE), and mean absolute error (MAE): RMSE: It is the square root of the mean difference between predicted values and those observed (% usage). It is the average prediction error on the same scale as the target or outcome being predicted. We supplied this metric because the correlation coefficient (R) is not always an accurate representation of model performance and classification of exemplar chemicals required an arbitrary cutoff (e.g., 90th percentile). We reported the correlation coefficient, R, between the predicted and observed % usage due to its previous use with human perceptual data. MAE: It is the mean of the absolute difference between predicted and observed (% usage). It thus assigns equal weight to all prediction errors, whether large or small.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y - \hat{y})^2}{N}}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}|,$$

where, $\hat{y}$ = predicted and y = observed

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where, TP = true positive and FN = false negative

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

where, TN = true negative and FP = false positive

Results

To better clarify the physicochemical basis of diverse perceptual descriptors, we designed a pipeline that begins with the identification of chemical features that contribute most to perceptual descriptors, followed by training machine-learning models to predict percepts from these features and evaluating their predictions (Figure 1a, Materials and methods). We used perceptual data from 2 human studies, Dravnieks (1985) and Keller (2016), conducted at different

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

times and with different participant demographics (Dravnieks 1985; Keller and Vosshall 2016). In the study by Dravnieks (1985), fragrance industry professionals rated 137 individual volatile chemicals for 146 different odor qualities (perceptual descriptors). We identified ~30 predictive physicochemical features (from DRAGON) for each of these perceptual descriptors (Supplementary File 1) (see Materials and methods for details). Machine-learning models that were trained with the physicochemical features successfully predicted most of the perceptual descriptors as seen by the computational validation (Figure 1b and Supplementary Figure 1) (average area under curve [AUC] = 0.81, average shuffle AUC = 0.62; t = 24.17, $P < 10^{-55}$; top 50 models average AUC = 0.90, average shuffle AUC =0.62; t = 55.54, $P < 10^{-75}$). We also observed that altering the general classification cutoff from the top 10% usage to the top 15% or 25% changes the AUC value determined for different percepts (Supplementary Table 1). Specifically, of note is the increase in performance as the cutoff is lowered, suggesting these descriptors in the study data set have fewer high scoring (% usage) examples for training and the high scoring chemicals may not be as physicochemically distinct as lower scoring (% usage) chemicals. To remove bias because of differences in the score distribution, we next evaluated other metrics for the validation such as RMSE, MAE, and correlation between predicted and observed % usage (R) (Supplementary File 2; see details in Materials and methods, "Model performance metrics"). Since each chemical has a complex perceptual profile, we analyzed the correlation between predicted and observed % usage over the validation for the full (146 descriptor set), which suggested good results (Supplementary Figure 1a). Next, for a set of hidden test chemicals, the predicted olfactory profile over all 146 perceptual descriptors also correlated well with the known human ratings (average r = 0.72; best predicted chemical: r = 0.86; worst predicted chemical: r = 0.67) (Figure 1c; Supplementary Figure 1).

The study by Dravnieks (1985) used experienced human raters, and to generalize the utility of our approach, we next applied it to the more recent study by Keller (2016) of general public volunteers (Keller and Vosshall 2016). As with the study by Dravnieks (1985), perceptual descriptors for a set of 69 hidden test chemicals (Keller et al. 2017) were also well predicted from physicochemical features (Supplementary Figure 1b) or with multiple train/test sets from all 476 chemicals (Supplementary Figure 1c).

The 2 studies, though differing significantly in methodology, evaluated a small number of identical perceptual descriptors. It was therefore possible to test whether models from the 1985 study could predict equivalent perceptual descriptors in the 2016 study (cross-study). Prior work has performed this analysis on a small number of overlapping chemicals using an approach involving semantic similarity and chemical features (Gutiérrez et al. 2018). We focused on 413 nonoverlapping chemicals and more traditional modeling methods to evaluate across studies. Models for "Sweet," Warm," "Sweaty," and "Chemical" trained on the study by Dravnieks (1985) were successful at classifying the 413 chemicals unique to the study by Keller (2016) (Dravnieks 1985; Keller and Vosshall 2016) (Figure 1d) (average cross-study AUC = 0.73 ± 0.07, maximum AUC = 0.82 ± 0.03 for "Sweet"). As a control, we compared the cross-study predictions with the Dravnieks (1985) model for a distinct percept, "Varnish," which achieved good accuracy in Figure 1b and is similar to "Chemical" but expected to differ from the rest. Consistent with expectation, the overall average AUC using the Dravnieks (1985) "Varnish" model cross-study was 0.41 ± 0.12. When we trained the Dravnieks (1985) models on randomly shuffled labels before the cross-study predictions, the overall average AUC was 0.52 ± 0.07 (Supplementary Table 2. These results suggest that identical perceptual descriptors across studies are predictable from a set of physicochemical features, despite differences in study sample demographics and odor diversity.

We next analyzed if the descriptors within each study could be predicted equally well by a different descriptor model with good classification accuracy. For the study by Keller (2016), "Bakery," which is similar to the many food-related descriptors in the study but differs from the rest, did not classify the 69 test chemicals as well as the percept-specific models (Supplementary Figure 2a). Of the 146 Dravnieks (1985) study descriptors, ~96% were better predicted by the percept-specific model vs "Varnish" (average Varnish AUC = 0.51; t = 21.65, $P < 10^{-59}$) (Supplementary Figure 2b). However, the "Varnish" model was indistinguishable from "Chemical," "Paint," and "Etherish," implying chemical features are redundant in some cases. As this also suggested some descriptors in an arbitrarily large descriptor space might be predicted equally well by semi or even unrelated chemical feature models, we studied this exhaustively (Supplementary Figure 3a). Overall, predictions with the actual descriptor model were often statistically better, even for some seemingly similar descriptors. However, this is not always the case, suggesting some descriptors may simply lack quality exemplar chemicals. We also tested additional, alternative methods to evaluate the descriptor models, with similar results summarized in Supplementary File 2.

Apart from these 2 semiquantitative psychophysical studies (Dravnieks (1985; Keller 2016), a large amount of perceptual data is available as text at various databases, some using identical or similar perceptual descriptors. Although these databases are not quantitative or methodical, we tested each of our 146 Dravnieks (1985) perceptual descriptor models on a unique set of 2525 chemicals from one such database maintained by the GoodScents company. The predicted perceptual scores of each chemical were evaluated against the known textual data using ROC analysis (Materials and methods). Although this task differed dramatically from previous test data sets, on average, the predictions compared favorably to the observed percepts (AUC = 0.72, t = 48.53, $P < 10^{-15}$) (Figure 1e). Collectively, these examples of predictive success within and across data sets establish that many perceptual descriptors, even those that are seemingly abstract, have a physicochemical basis that can be identified.

To get an overview of the physicochemical basis of odor perception, we created network representations of the relationship between the percepts and the most predictive chemical features (Bullmore and Sporns 2009; Meunier et al. 2010; Koulakov et al. 2011; Zhou et al. 2018). For example, we expected that similar descriptors ("Fruity, Citrus," "Lemon," "Grapefruit") were best predicted by similar chemical features and they would cluster together in the network (Figure 2a). Initially, we performed simple hierarchical clustering to compare the distances between the perceptual descriptors based on the % usage (Supplementary Figure 4a) and then based on chemical feature sets in the machine-learning models for comparison. Although some chemical features were selected for multiple descriptor models, resulting in unconventional pairings in the hierarchical tree relative to perceptual ratings, we observed many similarities (Supplementary Figure 4b).

We next turned our attention to the network-based visualizations, reducing the chemical features down to the top 3 for 93 of the most distinct perceptual descriptors in the study by Dravnieks (1985) (117 features in total). Despite the limited information, distinct clusters were detectable. In general, networks using more

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

chemical features (top 5 or 10) were better connected (Figure 2b, Supplementary Figure 5a). Interestingly, these networks relate well to those assembled only from human participant ratings rather than physicochemical information (Castro et al. 2013). Taken together, these analyses suggested that perceptual descriptors with highly correlated % usage (e.g., descriptors that are fruit-like) may be subtly different in terms of the most important or predictive chemical features.

The human olfactory system discriminates similar smelling chemicals and does so presumably by detecting minor differences in key physicochemical features using an array of odorant receptors. To understand how a machine-learning algorithm might achieve such discrimination, we selected 2 groups of closely correlated perceptual descriptors, fruit-like and soot-like, and performed a network analysis as before. As expected, many top physicochemical features were shared among these similar descriptors, and yet separate subclusters were present (Figure 2c, top and bottom). Representative compounds with descriptors such as "Grape Juice" and "Peach, Fruit" are subtly different from each other, as are ones for "Sooty" and "Tar" (Figure 2d). When examining these differences in physicochemical features, it is evident how slight variations in structurally related chemicals could result in distinct perceptual responses. We also observed this in an additional analysis (Supplementary Figure 5b). This suggests that physicochemical information in machine-learning models can address a complex challenge, similar to the biologically relevant discriminatory task.

An analysis of the chemical features selected for all the perceptual models suggested that the 3D structure of a chemical contributed significantly to predictions of odor perception, particularly the 3D-MoRSE (Schuur et al. 1996) and GETAWAY (Consonni et al. 2002) chemical features (DRAGON), which are 3D representations weighted by physicochemical properties that are possibly without precise structural interpretations (Supplementary Figure 5c). Simpler 2D features and functional group counts were less important but still among the top 10 features for some of the perceptual descriptors (Supplementary Figure 5c).

Only a miniscule portion of the odor–chemical space has been evaluated for perceptual information, and this in part reflects the low throughput and high cost of human studies. One approach to overcome this is to extend small experimental data sets to large, unexplored chemical spaces. Subsequently, we predicted the 146 Dravnieks (1985) study perceptual descriptors for a ~440 000 chemical library (Boyle, Guda, et al. 2016; Boyle, McInally et al. 2016) (Figure 3a, top and bottom). We evaluated ~68 million descriptor–chemical combinations and predicted numerous (hundreds to thousands) new chemicals that smell like each descriptor. These chemicals represent a massive expansion (>3000 times) of the previously known chemical space with perceptual descriptors, which is likely to cover a substantial fraction of putative volatile chemicals with odorant properties. Ultimately, the predictions allowed us to create, for the first time, a comprehensive chemical space of all 146 Dravnieks (1985) perceptual descriptors.

Visualizing this massive chemical space in a 2D image is difficult, so we represented only a fraction of the top predictions in the form of a network (Figure 3b). We next clustered similar perceptual descriptors, highlighting the frequently occurring chemical features among the top predictions. Though the machine-learning models incorporate potentially abstract chemical features, this type of analysis can help visualize structural features that may contribute to a certain percept (Figure 4).

Discussion

In this study, we provided a comprehensive analysis of odor perception prediction from physicochemical features of volatile chemicals and have supplied important groundwork to understand optimal methods, metrics, and approaches in modeling diverse perceptual descriptors. We do so with an additional focus on transparency and interpretability.

Of note is the finding that most perceptual descriptors are best predicted by chemical features that describe 3D geometries. The value of 3D information was anticipated however when considering structurally similar odorants share many 2D features. To successfully discriminate odorant percepts, machine-learning models utilize additional physicochemical properties, particularly 3D shape. In data sets with an arbitrarily large number of perceptual descriptors, the important chemical features could be redundant and cross-descriptor predictions overlap. However, we found that, although important chemical features overlap, the set of descriptors for a percept and the models themselves were indeed largely distinct. This would be consistent with evidence that perceptual descriptors appear highly structured and are not arbitrary (Gutiérrez et al. 2018).

Although caution is required in interpreting results from the Dravnieks (1985) or Keller (2016) data sets, which are small samples by typical machine-learning standards, our validations and control analyses establish that they are nevertheless rich sources of information for uncovering structure–odor percept relationships. The generalizability of physicochemical feature-based models across the differing sample demographics and the mostly distinct odor panels is further evidence. To that end, we have ultimately outlined a simple pipeline that can be applied to facilitate data-driven theories about the human olfactory perceptual space and its physicochemical origins on a considerably larger scale.

A handful of recent studies have used a variety of different computational approaches analyzing similar sets of perceptual study data (Keller et al. 2017; Gutiérrez et al. 2018; Nozaki and Nakamoto 2018; Sanchez-Lengeling et al. 2019). A direct comparison across these studies is somewhat limited due to differences in training and evaluation chemicals, metrics used, as well as differences in data processing. Although this study differs in several significant ways to others, we have attempted to place the results in context of diverse odor perception prediction efforts in a tabular form which shows the benefits of each approach (Supplementary File 3). To evaluate generalizability across different studies, we also expanded our analyses on cross-study validations from 2 separate sources, where training and testing are performed on data from different psychophysics studies (Figure 1d and e). These results suggested that models trained on the Dravnieks data could be successfully adapted to predict percepts of chemicals in the Keller study and a very different, nonexperimental data set in GoodScents. Although the size of the training set directly impacts success, and models trained on more data perform better, we find that the validation rates obtained in this study are quite good relative to the size of the Dravnieks training data. Some of the previous modeling efforts relied on open-source chemical feature representations including e-Dragon, a free web interface to an early version of Dragon and Mordred/RDKit. Analysis here used proprietary geometry optimization tools such as OMEGA alongside the full version of Dragon. When we compared different feature representations, it is evident that there are performance gains and losses depending on the perceptual descriptor set size (Supplementary File 3). Although there is no optimal approach, the tools used in this study appear to improve predictions, particularly when benchmarking using the same database, such as within the Keller 2016

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

data; that is, training (407) and testing (69) on chemicals from the study by Keller (2016) (Supplementary File 3). These comparisons across various modeling efforts with multiple performance metrics provide some insight into the top selected perceptual descriptors and their predictability for each perceptual descriptor.

The chemical features we report for the Dravnieks (1985) perceptual descriptors are potentially a valuable resource and will likely benefit researchers in identifying new chemicals that smell a specific way. Predicted compounds from the large computational screen are a rich source of information about the potential human olfactory chemical space. By applying machine learning alongside traditional chemoinformatic tools, we suggest it is now possible to extrapolate from the quality perceptual study data to large chemical spaces. We therefore anticipate that this study will provide a powerful approach and resource for the discovery of new flavors and fragrances.

Supplementary material

Supplementary data are available at *Chemical Senses* online.

Supplementary Figure 1. a) Average correlation (R) between the predicted and observed % usage for the full set of perceptual descriptors over cross validation. Dravnieks (1985) study chemicals (x-axis) are abbreviated as the CAS identifier. **b)** Evaluation of chemical (DRAGON) feature models trained on the Keller 2016 study data. Models classify 69 test chemicals (used in the DREAM analysis) as smelling like a given descriptor (top 10% Usage). These chemicals were excluded from training and chemical feature selection. The area under the ROC curve (AUC) compares predictions to the data observed from the general public volunteers in that study. Chance performance is defined by training models identically but on mislabeled chemicals (shuffle). Error is the standard deviation over 100 bootstrap samples. **c)** A similar analysis is done using an alternative validation method where all 476 chemicals in the Keller 2016 study are repeatedly divided into training and testing chemical sets (10-fold cross-validation, repeated 3 times). This covers more diversity than the 69 test chemicals. Chemical features for these models were selected using a subset of the data to minimize biased validation. The predictions are aggregated from the support vector machine (SVM) and regularized random forest algorithms. Additional information on AUC calculation and its interpretation are in Materials and Methods. Chemical feature selection methods and biases that affect validation are also defined in Materials and Methods. Source data supplied in Supplementary Figure 1.

Supplementary Figure 2. a) Area under the ROC curve (AUC) for classifying the top 10% of usage on 69 test chemicals with chemical (DRAGON) features across perceptual descriptors from the 55 Keller 2016 study participants, averaging over dilution. The 69 test chemicals are as reported in the DREAM analysis (Keller et al. 2017). AUCs computed from aggregated scores of a RBF SVM and a regularized random forest. Performance of each perceptual descriptor model is plotted alongside performance if replacing the predictions with the "Bakery" model. Chemical features selected and models fit on 407 training chemicals. Error (standard deviation) is over 100 bootstrap samples of the 69 test chemicals. **b)** Classification (AUC) of top 10% of usage for the 146 Dravnieks (1985) perceptual descriptors descriptor models (teal dots) compared to predictions using a top performing "Varnish" (purple dots) model. Perceptual descriptors colored in purple failed to outperform "Varnish," $p > .05$, adjusting for FDR (Benjamini-Hochberg). Plotted AUCs reflect the average of 3 RBF SVM models using different chemical features from a pool of ~70 over 30 cross validation folds (10- fold CV repeated 3 times) (RBF: Radial Basis Function; SVM: Support Vector Machine; FDR: False Discovery Rate). See Supplementary Figure 3 for exhaustive comparisons.

Supplementary Figure 3. a) Dravnieks (1985) study prediction performance over the cross validation where the percent usage of each perceptual descriptor is predicted by the models for the other descriptors. The color is the p value adjusted for FDR (T-test). Descriptor labels are colored (red) to distinguish the models that are of a lower quality rather than perceptual redundancy. These perceptual descriptors may fail for many reasons but notably most are not well represented among Dravnieks (1985) study chemicals (e.g. lack exemplars for classification training).

Supplementary Figure 4. a) Hierarchical clustering of the Dravnieks (1985) study data by % usage. The cluster (colors) number is determined by the gap statistic over bootstrap samples. The distance is Euclidean. **b)** Hierarchical clustering is instead performed based on chemical feature sets appearing in the machine learning models. The distance is 1-Jaccard index, where the Jaccard index here indicates the similarity of binary strings (1,0) specifying if a chemical feature is or is not in the perceptual descriptor model.

Supplementary Figure 5. a) The 10 most important chemical (DRAGON) features for accurate predictions of perception (% usage) are used to build a network representation that shows relationships among the perceptual descriptors in terms of their prospective physicochemical similarity. Connectivity in the network signifies shared chemical features among 93 distinct perceptual descriptors and is used to infer clusters of similar perceptual descriptors according to the Louvain algorithm. The large number of features leads to a densely connected network but clusters detected. **b)** *Left*, discriminating top chemicals that smell like "cherry" versus "tar," according to Dravnieks (1985) study respondents. The discrimination success is quantified by the average AUC across 30 cross validation folds (10-fold CV repeated 3 times) for models comprised of 1, 2, and 3 principal components (PC 1–3) that optimally retain information in the combined top 10 chemical (DRAGON) features (20 total). Error bars reflect the standard error. Note the 3 component model provides perfect classification. *Right*, exemplar chemicals for "cherry (berry)" and "tar" that are structurally similar but with subtly distinct chemical features. **c)** Counts of the chemical (DRAGON) features selected in bins from the top 1–10 (x–axis) for 146 perceptual descriptors with respect to the broad categories (y-axis) the features fall into.

Supplementary Table 1. Related to Figure 1b. The average AUC is shown for varying classification cutoffs. The % usage is transformed into active and inactive labels according to the top end of the % usage distribution (Top 10, 15, and 25), which changes the number of active and inactive chemicals.

Supplementary Table 2. Cross-Study classification performance. Dravnieks (1985) models predict the same perceptual descriptor in the Keller 2016 study for 413 chemicals unique to the study. The area under the curve (AUC) is averaged over 100 bootstrap samples. The perceptual descriptor is the model used for predictions. Each descriptor is appended with "Shuffle" or "Varnish," showing the performance when the Dravnieks (1985) study model is trained on shuffled labels for exemplar chemicals or, alternatively, the Dravnieks (1985) "Varnish" model.

Conflict of interest

J.K. and A.R. are listed as inventors in patent applications filed by UCR. A.R. is founder of Sensorygen Inc. that discovers novel insect repellents, flavors, and fragrances.

Data availability

Data used in the analyses are publicly available from the references cited, and the data generated in this manuscript are provided as Excel files in Supplementary Files 1–3. Any data associated with this manuscript is also available in other formats on request from the communicating author. Upon request authors will make available, any previously unreported custom computer code or algorithm used to generate results that are reported in the paper and central to its main claims.

References

Ambroise C, McLachlan GJ. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*. 99(10):6562–6566.

Boyle SM, Guda T, Pham CK, Tharadra SK, Dahanukar A, Ray A. 2016. Natural DEET substitutes that are strong olfactory repellents of mosquitoes and flies. *bioRxiv*, doi: 10.1101/060178.

Boyle SM, McInally S, Tharadra S, Ray A. 2016. Short-term memory trace mediated by termination kinetics of olfactory receptor. *Sci Rep*. 6:19863.

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

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022

Bullmore E, Sporns O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*. 10:186–198. doi: 10.1038/nrn2575.

Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. 2008. ChemmineR: a compound mining framework for R. *Bioinformatics*. 24(15):1733–1734.

Castro JB, Ramanathan A, Chennubhotla CS. 2013. Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS One*. 8(9):e73289.

Consonni V, Todeschini R, Pavan M. 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J Chem Inf Comput Sci*. 42(3):682–692.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *Complex Systems*, 1695. Available from: https://igraph.org.

de March CA, Titlow WB, Sengoku T, Breheny P, Matsunami H, McClintock TS. 2020. Modulation of the combinatorial code of odorant receptor response patterns in odorant mixtures. *Mol Cell Neurosci*. 104:103469.

Dravnieks A. 1985. Atlas of odor character profiles. In: Dravnieks A, editor. *West Conshohocken*, PA: ASTM International, 1992. doi: 10.1520/DS61-EB.

Dubois D, Rouby C. 2002. Names and categories for odors: the veridical label. In: Rouby C, Schaal B, Dubois D, Gervais R, Holley A, editors. *Olfaction, taste, and cognition*. Cambridge: Cambridge University Press. p. 47–66. doi: 10.1017/CBO9780511546389.009.

Gutiérrez ED, Dhurandhar A, Keller A, Meyer P, Cecchi GA. 2018. Predicting natural language descriptions of mono-molecular odorants. *Nat Commun*. 9(1):4979.

Haddad R, Medhanie A, Roth Y, Harel D, Sobel N. 2010. Predicting odor pleasantness with an electronic nose. *PLoS Comput Biol*. 6(4):e1000740.

Honnibal M, Montani I, Van Landeghem SA. 2020. spaCyspaCy: industrial-strength natural language processing in python. *Zenodo*. doi: 10.5281/zenodo.1212303.

Hu XS, Ikegami K, Vihani A, Zhu KW, Zapata M, de March CA, Do M, Vaidya N, Kucera G, Bock C, et al. (2020). Concentration-dependent recruitment of mammalian odorant receptors. *ENeuro*. 7(2). doi: 10.1523/ENEURO.0103-19.2019.

Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, Szalai B, Mainland JD, Ihara Y, Yu CW, Wolfinger R, et al.; DREAM Olfaction Prediction Consortium. 2017. Predicting human olfactory perception from chemical features of odor molecules. *Science*. 355(6327):820–826.

Keller A, Vosshall LB. 2016. Olfactory perception of chemically diverse molecules. *BMC Neurosci*. 17(1):55.

Keller A, Zhuang H, Chi Q, Vosshall LB, Matsunami H. 2007. Genetic variation in a human odorant receptor alters odour perception. *Nature*. 449(7161):468–472.

Kepple D, Koulakov A. 2017. Constructing an olfactory perceptual space and predicting percepts from molecular structure. *arXiv*. doi: 1708.05774.

Khan RM, Luk CH, Flinker A, Aggarwal A, Lapid H, Haddad R, Sobel N. 2007. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *J Neurosci*. 27(37):10015–10023.

Koulakov AA, Kolterman BE, Enikolopov AG, Rinberg D. 2011. In search of the structure of human olfactory space. *Front Syst Neurosci*. 5:65.

Kuhn M. 2008. caret package. *J Stat Softw*. 28(5): 1–26. Available from: http://www.jstatsoft.org/v28/i05/paper.

Landrum G. 2006. RDKit: open-source cheminformatics. Available from: Http://Www.Rdkit.Org.

Lawless HT. 1984. Flavor description of white wine by "expert" and nonexpert wine consumers. *J Food Sci*. doi: 49(1):120–123. doi: 10.1111/j.1365– 2621.1984.tb13686.x.

Licon CC, Bosc G, Sabri M, Mantel M, Fournel A, Bushdid C, Golebiowski J, Robardet C, Plantevit M, Kaytoue M, et al. 2019. Chemical features mining provides new descriptive structure-odor relationships. *PLoS Comput Biol*. 15(4):e1006945.

Mainland JD, Keller A, Li YR, Zhou T, Trimmer C, Snyder LL, Moberly AH, Adipietro KA, Liu WL, Zhuang H, et al. 2014. The missense of smell: functional variability in the human odorant receptor repertoire. *Nat Neurosci*. 17(1):114–120.

Majid A, Kruspe N. 2018. Hunter-gatherer olfaction is special. *Curr Biol*. 28(3):409–413.e2.

McClintock TS, Khan N, Alimova Y, Aulisio M, Han DY, Breheny P. 2020. Encoding the odor of cigarette smoke. *J Neurosci*. 40(37):7043–7053.

McClintock TS, Wang Q, Sengoku T, Titlow WB, Breheny P. 2020. Mixture and concentration effects on odorant receptor response patterns in vivo. *Chem Senses*. 45(6):429–438. doi: 10.1093/chemse/bjaa032.

McGann JP. 2017. Poor human olfaction is a 19th-century myth. *Science*. 356(6338):eaam7263. doi: 10.1126/science.aam7263.

McRae JF, Mainland JD, Jaeger SR, Adipietro KA, Matsunami H, Newcomb RD. 2012. Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, cis-3-hexen-1-ol. *Chem Senses*. 37(7):585–593.

Meunier D, Lambiotte R, Bullmore ET. 2010. Modular and hierarchically modular organization of brain networks. *Front Neurosci*. 4:200.

Mombaerts P. 1999. Molecular biology of odorant receptors in vertebrates. *Annu Rev Neurosci*. 22:487–509.

Mombaerts P. 2001. The human repertoire of odorant receptor genes and pseudogenes. *Annu Rev Genomics Hum Genet*. 2(1):493–510.

Mombaerts P, Wang F, Dulac C, Vassar R, Chao SK, Nemes A, Mendelsohn M, Edmondson J, Axel R. 1996. The molecular biology of olfactory perception. *Cold Spring Harb Symp Quant Biol*. 61:135–145.

Nozaki Y, Nakamoto T. 2016. Odor impression prediction from mass spectra. *PLoS One*. 11(6):e0157030.

Nozaki Y, Nakamoto T. 2018. Correction: predictive modeling for odor character of a chemical using machine learning combined with natural language processing. *PLoS One*. 13(12):e0208962.

Olofsson JK, Gottfried JA. 2015. The muted sense: neurocognitive limitations of olfactory language. *Trends Cogn Sci*. 19:314–321.

Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. In: EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference; Doha, Qatar. doi: 10.3115/v1/d14-1162.

Pfister P, Smith BC, Evans BJ, Brann JH, Trimmer C, Sheikh M, Arroyave R, Reddy G, Jeong HY, Raps DA, et al. 2020. Odorant receptor inhibition is fundamental to odor encoding. *Curr Biol*. 30(13):2574–2587.e6.

R Development Core Team. 2016. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing.

Raiche G. 2010. *nFactors: an R package for parallel analysis and non graphical solutions to the Cattell scree test*. R package version 2.3.3.

Rossiter KJ. 1996. Structure-odor relationships. *Chem Rev*. 96(8):3201–3240.

Saito H, Kubota M, Roberts RW, Chi Q, Matsunami H. 2004. RTP family members induce functional expression of mammalian odorant receptors. *Cell*. 119(5):679–691.

Sanchez-Lengeling B, Wei JN, Lee BK, Gerkin RC, Aspuru-Guzik A, Wiltschko AB. 2019. Machine learning for scent: learning generalizable perceptual representations of small molecules. *arXiv*. doi:1910.10685.

Schuur JH, Selzer P, Gasteiger J. 1996. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J Chem Inf Comput Sci*. 36(2):334–344.

Shirasu M, Yoshikawa K, Takai Y, Nakashima A, Takeuchi H, Sakano H, Touhara K. 2014. Olfactory receptor and neural pathway responsible for highly selective sensing of musk odors. *Neuron*. 81(1):165–178.

spaCy. (2016). *spaCy API documentation*.

Snitz K, Yablonka A, Weiss T, Frumin I, Khan RM, Sobel N. 2013. Predicting odor perceptual similarity from odor structure. *PLoS Comput Biol*. 9(9):e1003184.

Tran N, Kepple D, Shuvaev SA, Koulakov AA. 2019. Deepnose: using artificial neural networks to represent the space of odorants. *bioRxiv*, 464735. doi: 10.1101/464735.

Trimmer C, Keller A, Murphy NR, Snyder LL, Willer JR, Nagai MH, Katsanis N, Vosshall LB, Matsunami H, Mainland JD. 2019. Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proc Natl Acad Sci USA*. 116(19):9475–9480.

Van Rossum G, Drake Jr FL. 1995. Python reference manual. Amsterdam: Centrum voor Wiskunde en Informatica.

Vassar R, Ngai J, Axel R. 1993. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell*. 74(2):309–318.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York. Available from: https://ggplot2.tidyverse.org.

Zhou Y, Smith BH, Sharpee TO. 2018. Hyperbolic geometry of the olfactory space. *Sci Adv*. 4(8): eaaq1458. doi: 10.1126/sciadv.aaq1458.

Downloaded from https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjab007/6153471 by guest on 21 June 2022