

## RESEARCH

# Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features

Hongyang Li<sup>1</sup>, Bharat Panwar<sup>1</sup>, Gilbert S. Omenn<sup>1,2</sup> and Yuanfang Guan<sup>1,\*</sup>

<sup>1</sup>Department of Computational Medicine and Bioinformatics and <sup>2</sup>Departments of Internal Medicine and Human Genetics and School of Public Health, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

\*Correspondence address: Yuanfang Guan, Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109. Tel: 734-764-0018; Fax: 734-615-6553; E-mail: [gyuanfan@umich.edu](mailto:gyuanfan@umich.edu)

## Abstract

**Background:** The olfactory stimulus-percept problem has been studied for more than a century, yet it is still hard to precisely predict the odor given the large-scale chemoinformatic features of an odorant molecule. A major challenge is that the perceived qualities vary greatly among individuals due to different genetic and cultural backgrounds. Moreover, the combinatorial interactions between multiple odorant receptors and diverse molecules significantly complicate the olfaction prediction. Many attempts have been made to establish structure-odor relationships for intensity and pleasantness, but no models are available to predict the personalized multi-odor attributes of molecules. In this study, we describe our winning algorithm for predicting individual and population perceptual responses to various odorants in the DREAM Olfaction Prediction Challenge. **Results:** We find that random forest model consisting of multiple decision trees is well suited to this prediction problem, given the large feature spaces and high variability of perceptual ratings among individuals. Integrating both population and individual perceptions into our model effectively reduces the influence of noise and outliers. By analyzing the importance of each chemical feature, we find that a small set of low- and nondegenerative features is sufficient for accurate prediction. **Conclusions:** Our random forest model successfully predicts personalized odor attributes of structurally diverse molecules. This model together with the top discriminative features has the potential to extend our understanding of olfactory perception mechanisms and provide an alternative for rational odorant design.

**Keywords:** olfactory perception; structure-odor relationships; random forest; chemoinformatics

## Background

Olfactory perception is the sense of smell in the presence of odorants. The odorants bind to and activate olfactory receptors (ORs), which transmit the signal of odor to the brain [1]. The existence of a large family of olfactory receptors enables humans to perceive an enormous variety of odorants with distinct sensory attributes [1]. An olfactory receptor can respond to multiple

odor molecules; conversely, an odorant may interact with many olfactory receptors with different affinities [2]. Unlike the well-defined wavelength of light in vision and frequency of sound in hearing, the size and dimensionality of the olfactory perceptual space is still unknown [3]. It is not clear how the numerous physicochemical properties of a molecule relate to its odor, or how mammals process and detect the broad range of the olfactory spectrum. Some structurally similar compounds display

Received: 12 April 2017; Revised: 6 November 2017; Accepted: 7 December 2017

© The Author(s) 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

distinct odor profiles, whereas some dissimilar molecules exhibit almost the same smell [4–6]. Even for an identical molecule, the perceived quality varies immensely between individuals due to genetic variation [7]. Therefore, accurate prediction of personalized olfactory perception from the chemical features of a molecule is highly challenging.

In the past, many attempts have been made to establish structure-odor relationships and predict the odor from the physicochemical properties of a molecule [8]. An early study showed that volatile and lipophilic molecules fulfill the requirements to be odorants [9]. The correlation of odor intensities with different structural, topological, and electronic descriptors was calculated for 58 different odorants; molecular weight, partial charge on most negative atoms, quantum chemical polarity parameter, average distance sum connectivity, and a measure of the degree of unsaturation were particularly important descriptors [10]. Multidimensional scaling and self-organizing maps were used to produce 2-dimensional maps of the Euclidean approximation of olfactory perception space [11]. A principal component analysis identified the latent variables in a semantic odor profile database of 881 perfume materials with semantic profiles of 82 odor descriptors and classified odors into 17 different classes [12]. Although it is not possible to predict the odor profile of a molecule, some progress has been achieved for predicting the intensity [13] and pleasantness of an odorant. Methods for predicting the perceived pleasantness of an odorant have utilized the most correlated physical features of molecular complexity [14] and molecular size [15, 16]. A major challenge is that different individuals perceive odorants with different sets of odorant receptors [17, 18], and perception is also strongly shaped by learning and experience [19]. Different cultures have different linguistic descriptions of smells [20–22], so generating olfaction datasets is tedious work. Many computational methods have been developed to relate chemical structure to percept [4, 10, 15, 16, 23–26], but most of them are based on single and very old psychophysical datasets [27]. Therefore, a rigorous quantitative structure-activity relationship (QSAR) model [28, 29] of personalized olfactory perception is needed for accurate predictions.

The Dialogue on Reverse Engineering Assessment and Methods (DREAM) organized the olfaction prediction challenge [30]. DREAM is a leader in organizing crowdsourcing challenges to evaluate model predictions and algorithms in systems biology and medicine [31]. Here we describe our winning algorithm, the best performer of subchallenge 1, for predicting individual responses and the second best performer of subchallenge 2 for predicting population responses. As olfactory perception is inherently a complex nonlinear process, decision tree-based algorithms are well suited to this problem. Particularly, a random forest (RF) consisting of multiple decision trees addresses the overfitting issue when the feature space is much larger than the sample space. Moreover, random forest is relatively robust to noise and outliers [32], especially when a large variability of individual perceptual responses is observed. To further reduce the effects of large variability, noise, and outliers, we integrated the average rating of individuals (population response) into our model. Our final model succeeds in predicting olfactory perception using only a small set of chemical features. These features are likely to be low- and nondegenerative molecular descriptors, indicating that traditional simple descriptors like functional groups are less effective in distinguishing the odor profiles of structurally similar molecules. Meanwhile, our model potentially provides useful insights on the basic molecular mechanisms of olfactory perception. Together with new scaffolds

of odorants observed in the dataset and top discriminative chemoinformatic features, our model offers an alternative for rational odorant design.

## Data Description

### Psychophysical dataset

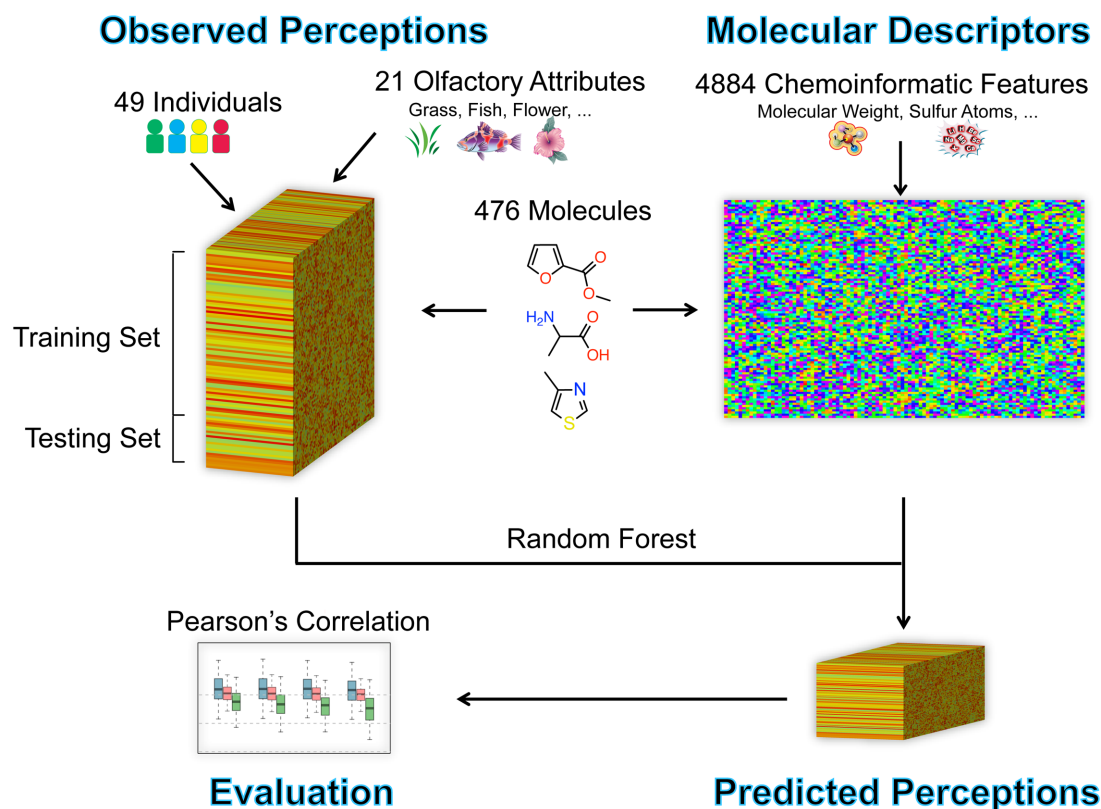
The DREAM organizers provided psychophysical data that were originally collected between February 2013 and July 2014 as part of the Rockefeller University Smell Study [33]. The data were collected from 61 ethnically diverse healthy men and women between the ages of 18 and 50. These subjects volunteered and gave their written informed consent to smell the stimuli used in this study [33]. They were naïve and didn't receive any kind of olfaction training. In the DREAM olfaction prediction challenge, the data of only 49 subjects were provided because some subjects didn't give permission to use their data. The perceptual ratings of 476 different molecules were assigned by these 49 subjects at 2 different concentrations (high and low); in addition, 20 molecules were tested twice. Each subject rated the perception of 992 stimuli (476 plus 20 replicated molecules at 2 different concentrations). Twenty-one perceptual attributes (intensity, pleasantness, and 19 semantic attributes) were used to describe the odor profile of a molecule. The semantic attributes are bakery, sweet, fruit, fish, garlic, spices, cold, sour, burnt, acid, warm, musky, sweaty, ammonia/urinous, decayed, wood, grass, flower, and chemical. Subjects used a scale from 0 to 100 where 0 is "extremely weak" and 100 is "extremely strong" for intensity; 0 is "extremely unpleasant" and 100 is "extremely pleasant" for pleasantness; and 0 is "not at all" and 100 is "very much" for semantic attributes. This dataset of 476 chemicals was divided into 3 subsets by the organizers: 338 for the training set, 69 for the leaderboard, and 69 for the test set. We combined the 338 training and 69 leaderboard molecules (407 molecules in total) as our final training set.

### Chemoinformatic features of molecules

A total of 476 structurally diverse odorant molecules were used in this study, including 249 cyclic molecules, 52 organosulfur molecules, and 165 ester molecules (Supplementary Fig. S1). The participating investigators were encouraged to use any kind of chemical and physical properties of the molecules for developing prediction models. By default, the organizers provided 4884 different chemical features for each of the 476 molecules, calculated by a commercial chemoinformatics software package known as Dragon (version 6) [34]. Features were divided into 29 different logical molecular descriptor blocks including constitutional descriptors, topological indices, 2D autocorrelations, etc. These chemoinformatic features are useful in establishing structure-odor relationships and further developing machine learning prediction models. The compound identification number (CID) for each molecule was also provided so participating investigators could obtain more information about the molecules from other resources (e.g., PubChem) [35].

## Results

The overall workflow of the olfaction prediction is shown in Fig. 1. The organizers provided an unpublished large psychophysical dataset of 476 structurally and perceptually diverse molecules sensed by 49 different individuals [33]. Twenty-one perceptual attributes were collected, including odor intensity



**Figure 1:** The overview of the olfaction prediction. The observed perceptions form a 3-dimensional array, where the 3 dimensions are 476 molecules, 49 individuals, and 21 olfactory attributes. The input chemoinformatic features form a 2-dimensional matrix, where the rows are 476 molecules and columns are 4884 molecular descriptors. Our random forest model is built on the training set (407 molecules), and the individual responses for the test set (69 molecules) are predicted. The final evaluation is based on the Pearson's correlation between observed and predicted perceptions.

and pleasantness and 19 semantic descriptors. A subset of 407 molecules (338 training and 69 leaderboard molecules) was used as the final training set in our random forest model, and the other 69 held-out molecules formed the test set. The organizers provided the Dragon software [34]-based large-scale molecular descriptors, containing 4884 chemical features for each molecule. Models were evaluated based on the Pearson's correlation between the observed and predicted perceptions.

### Variability of olfactory perception among individuals

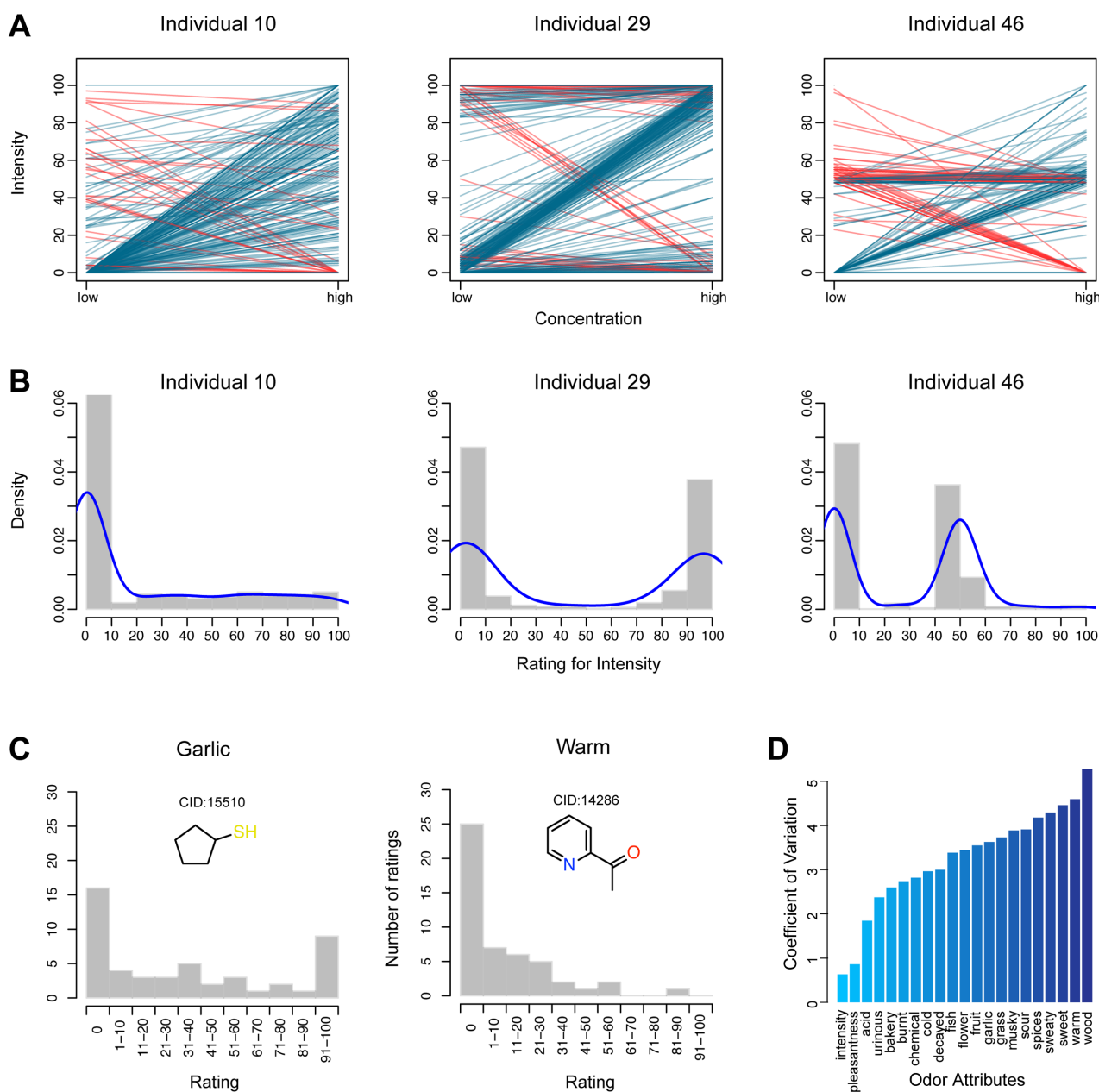
The intensity perceptions of 476 molecules at high and low concentrations vary tremendously among individuals. For example, individuals 10, 29, and 46 exhibit entirely different perceptual profiles for intensity (Fig. 2). Ideally the perceptual rating for intensity should increase as the measuring concentration rises (blue lines in Fig. 2A), while it is commonly observed that the intensity rating of some molecules decreases (red lines in Fig. 2A and Supplementary Fig. S2). In fact, the 49 subjects were lacking any kind of professional training, and they were biased in assigning the perceptual rating value between 0 and 100 (Fig. 2B and Supplementary Fig. S3). Except for molecules without odor rated near 0, individual 10 tended to assign ratings uniformly, whereas individual 29 preferred to rate at 100 and individual 46 was inclined to rate around 50.

In addition to intensity, the other perceived attributes were rated differently among individuals (Fig. 2C). Even for the same molecule, cyclopentanethiol (CID: 15 510), 16 subjects did not apply the descriptor “garlic,” whereas 9 subjects rated

it 100. Similarly, 2-acetylpyridine (CID: 14 286) showed great variability in “warm” ratings—about half of the subjects rated it 0, and the other half perceived “warm” from it. The large differences may result from the relative ambiguity of the word “warm” to describe odor. The average variance of 21 attribute ratings across all individuals is shown in Fig. 2D. Compared with intensity and pleasantness, the 19 semantic qualities display much larger coefficients of variation. Thus, the diversity of perceptual ratings between subjects considerably complicates the prediction challenge.

### Strategies for accurate personalized olfaction predictions

Considering the large variability of the perceived ratings, we propose that random forest could be an excellent choice as a base learner because it applies the strategy of training on different parts of the dataset and averaging multiple decision trees to reduce the variance and avoid overfitting. We compared different machine learning algorithms (linear, ridge, support vector regression, random forest) using 5-fold cross-validations and found that random forest outperforms other base learners in predicting individual responses for “intensity,” “pleasantness,” and 19 semantic descriptors (Fig. 3A). Given a small sample size of 407 training molecules, random forest identifies and utilizes the most discriminative features out of 4884 molecular descriptors to make decisions. Clearly, a simple linear regression model fails when the dimension of the feature space is too large. Therefore, random forest was selected as our base-learner and used in the follow-up improvements.

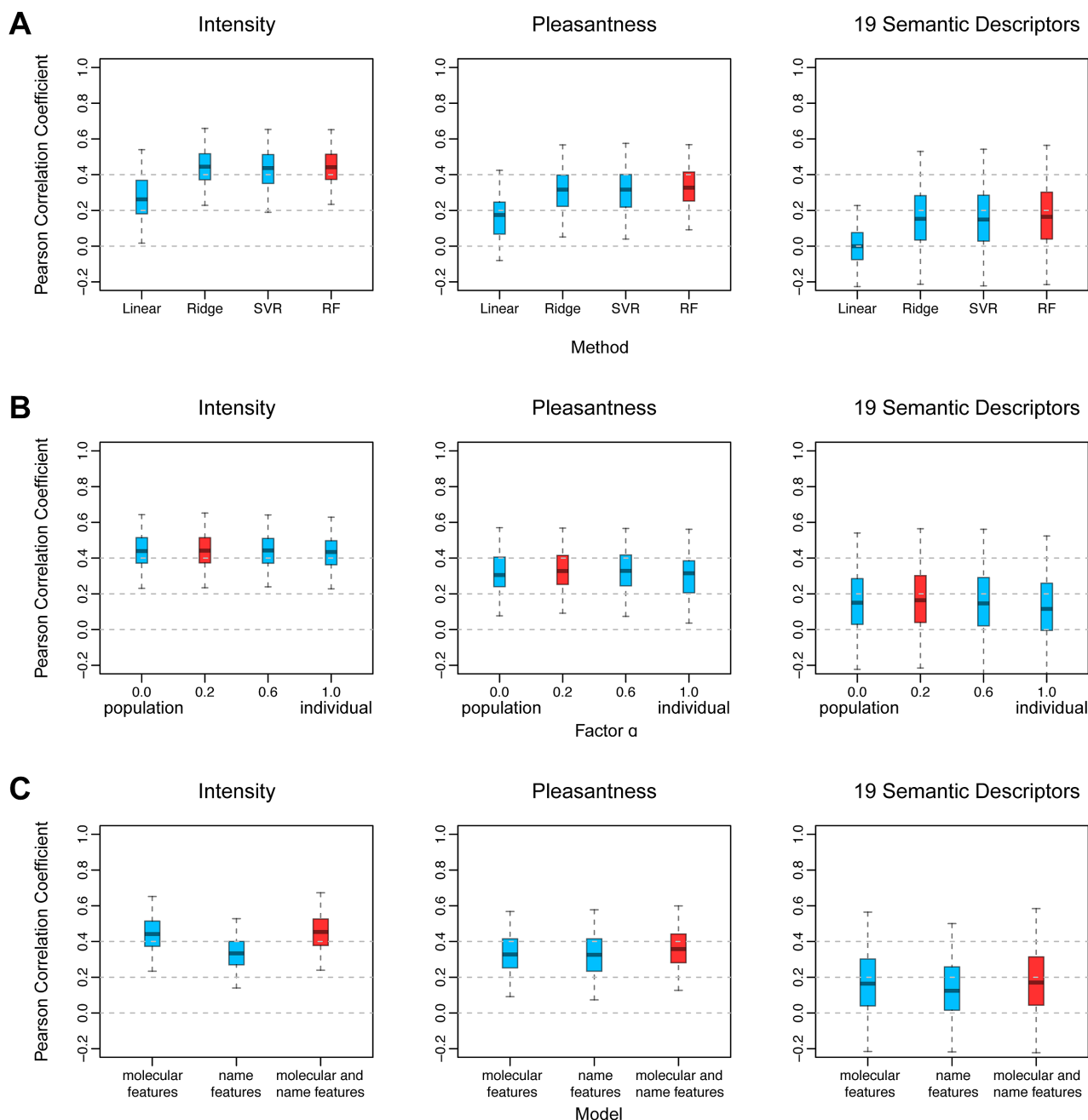


**Figure 2:** Variability of olfactory perception among individuals. **(A)** The intensity ratings for all molecules at low and high concentrations from individuals 10, 29, and 46. Blue lines represent the ideal cases, in which the rating values increase as the concentration becomes higher. Conversely, red lines represent decreased rating values at high concentration. **(B)** The density distributions of the intensity ratings from these 3 individuals. Blue lines are the fitting curves of the density distribution. The intensity ratings and density distributions from all individuals are shown in Supplementary Figs S1 and S2, respectively. **(C)** The “garlic” and “warm” rating distributions among 49 individuals for 2-acetylpyridine and cyclopentanethiol, respectively. **(D)** The coefficients of variation of 21 perceptual attributes in increasing order.

Recognizing that the population responses (the average perceptions of all individuals) are more stable compared with individual responses, we overcome the variability of individual ratings by introducing a weighting factor  $\alpha$ . This parameter serves as a balance between individual and population ratings. When  $\alpha$  equals 0, only population ratings are considered. Conversely, when  $\alpha$  equals 1, only individual ratings are used (see the “Methods”). Surprisingly, a small  $\alpha = 0.2$  achieves the largest Pearson’s correlation coefficient (Fig. 3B). Without population information ( $\alpha = 1.0$ ), the correlation of predicting the 19 semantic descriptors is the lowest. This reveals that population percep-

tions play a crucial role when individual responses display large fluctuations.

To further improve the performance, we applied a sliding window of 4-letter size to each molecule name, generating a total of 11 786 binary name features (see the “Methods”). These name features are very similar to the molecular fingerprints, providing extra information about the similarities of molecules. The 4-letter window is selected to efficiently capture the chemical similarity. Larger sliding window sizes greatly increase computational cost, as the size of the feature space is an exponential function of window size. Although the random forest model



**Figure 3:** The performance of different models and strategies. From left to right, the Pearson's correlation coefficients of intensity and pleasantness and 19 semantic descriptors from 5-fold cross-validations are shown as a boxplot. The red base-learners or strategies are used in our final model. (A) The performance of 4 different base-learners: linear, ridge, SVM, and random forest. (B) The performance of using different values of weighting factor  $\alpha$ . (C) The performance of using molecular features alone, name features alone, and both molecular and name features.

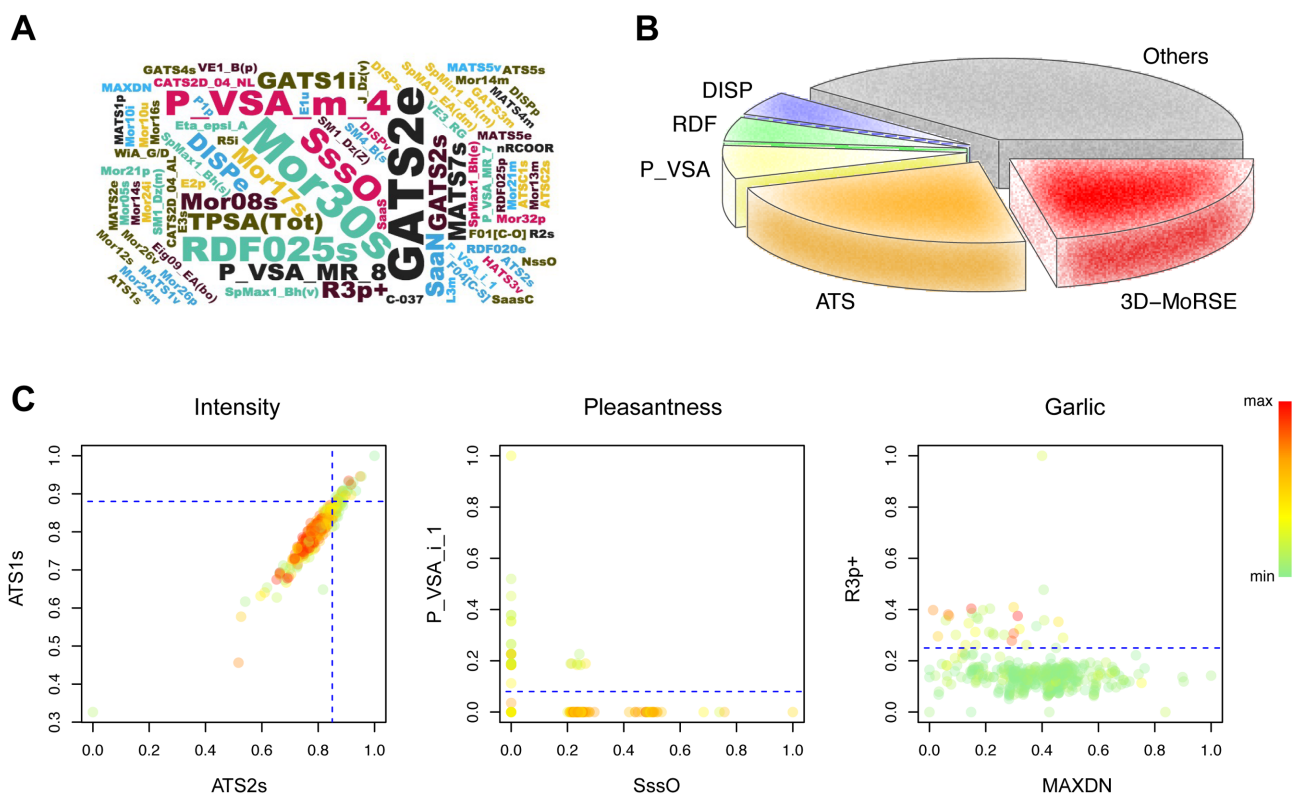
using only the name features has relatively low correlations, the ensemble model aggregating both molecular and name features performs the best, ranking first in predicting the 21 personalized perceptual attributes (Fig. 3C and Supplementary Fig. S4).

### Discriminative chemoinformatic features for olfaction prediction

Our random forest model evaluates the importance of each molecular descriptor in prediction. It is well known that sulfur-containing organic molecules tend to have “garlic” odor,

whereas esters are often smelled as “fruity.” Many models have been built to correlate molecular size and complexity with the “pleasantness” of a compound. However, it remains unclear what chemical features of a molecule decide its multiple odor attributes. Random forest enables us to estimate the importance of each chemical feature by permuting the values of a feature across samples and computing the increase in prediction error. We calculate the increased delta error of each chemical feature for all 21 olfactory qualities. Interestingly, top-ranking features used by random forest do not necessarily have high linear Pearson's correlations with observed ratings. For





**Figure 4:** Top discriminative features used in random forest. (A) The word cloud of top 5 features used in predicting 21 perceptual attributes. (B) The pie chart of molecular descriptor categories in the top 5 features. (C) Projection of all molecules onto selected discriminative feature spaces. The color of each spot represents the relative strength of the perceived rating, averaged among 49 individuals. The dashed lines display the possible decision boundaries created by random forest.

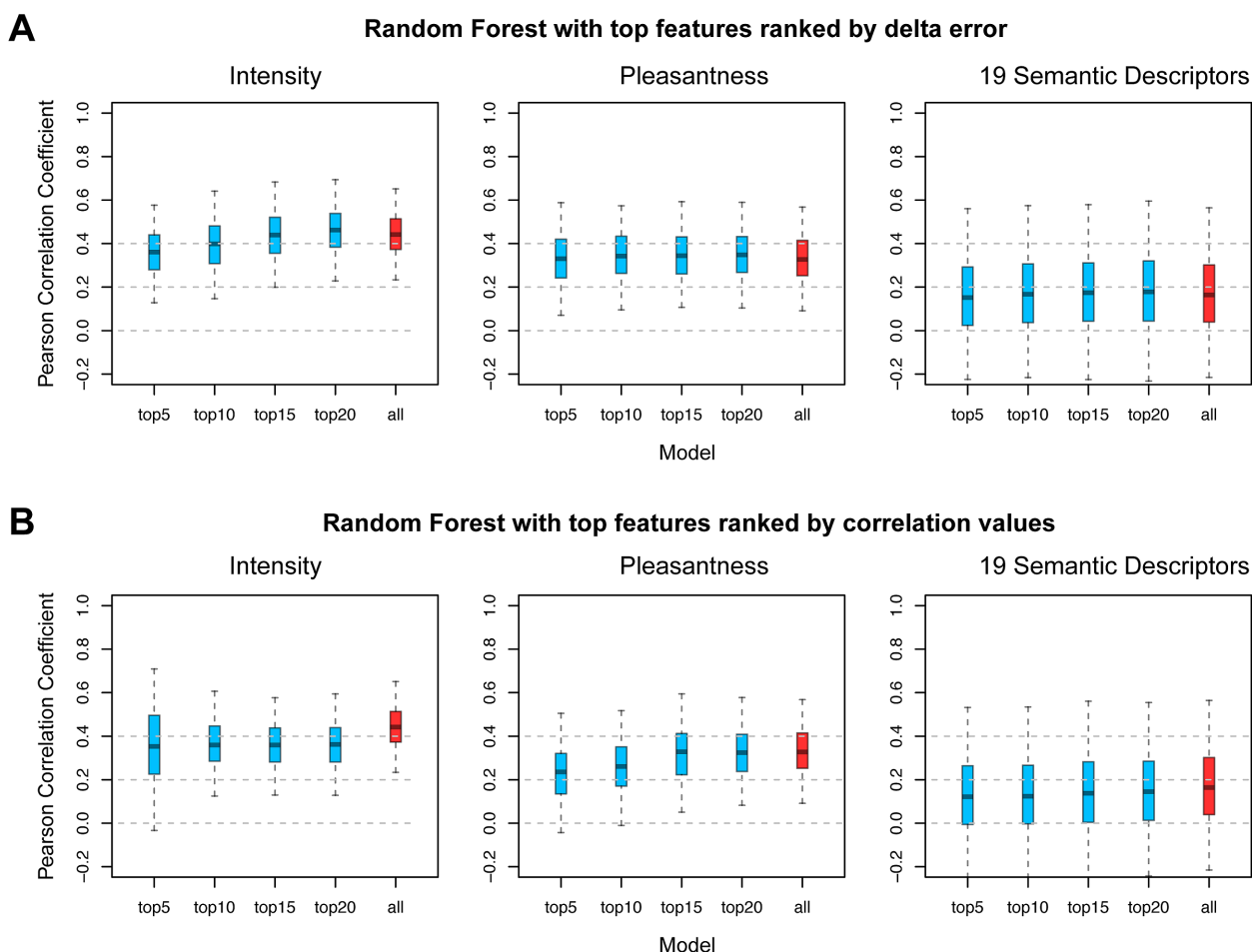
example, the correlation coefficients of the top 5 features for “decayed” prediction are listed in Supplementary Table S1. The molecular feature P.VSA\_m\_4 (interpreted as the presence of sulfur atoms) ranks first and has the largest correlation. However, the second and third features in random forest have nearly no correlations with observed perceptions. Upon inspecting the top 5 features that have the largest correlation values (yellow columns in Supplementary Table S1), we notice that they are all related to sulfur atoms, leading to high redundancy and intercorrelation.

By analyzing all the top 5 features ranked by the delta error, we find that discriminative molecular features are more likely to be low- or nondegenerative. The complete lists of top 20 features ranked by delta error or Pearson’s correlation are shown in Supplementary Tables S2 and S3, respectively. Interestingly, simple chemical features (molecular weight, number of sulfur atoms, presence of a functional group, etc.) are not very powerful in prediction because they display high degeneracy—different molecules may have identical or similar values [36–38]. Features with low- or nondegeneracy more likely play an essential role in our random forest model. The frequency of all top 5 molecular features is represented as the size of words in Fig. 4A. We find that autocorrelation of a topological structure (ATS) and 3D-MoRSE descriptors occur 24 and 23 times, respectively, whereas simple descriptors such as N% (percentage of N atoms), nRCOOR (number of aliphatic esters), and NssO (number of ssO atoms) are used only once (Fig. 4B).

To understand how the random forest model works, we projected all molecules onto selected important feature spaces (Fig. 4C). The color of each molecule represents the strength

of the perceived rating. For example, ATS1s and ATS2s are the most important features in predicting the “intensity” of a molecule. They can be interpreted as the combined information of molecule size and the intrinsic state of all atoms. Molecules with large ATS1s and ATS2s values tend to have low intensity (top right green spots in Fig. 4C, left panel). Another example is the “pleasantness” rating of a molecule, for which SssO (presence of ester or ether) and P.VSA\_i\_1 (presence of sulfur or iodine atom) are crucial. Clearly, molecules containing sulfur or iodine atoms have lower “pleasantness” values (green spots above the dashed line in Fig. 4C, middle panel). And it is widely known that ester has a characteristic pleasant odor and lower ethers can act as anesthetics, whereas presence of sulfur atom leads to unpleasant “garlic” and “decayed” odor. Therefore, key features of “garlic” odor include MAXDN (presence of ketone or ester) and R3p+ (presence of sulfur atom). Molecules containing sulfur atoms are more likely to be “garlicky,” whereas ketones and esters seldom have such smells (red spots above the dashed line in Fig. 4C, right panel).

Rebuilding the random forest model with the top 5, 10, 15, or 20 key features, we find that a small set of chemical features is sufficient for accurate prediction. These top features selected by random forest may have very low linear Pearson’s correlations with perceived qualities, yet they are powerful in discriminating different odorants. This is because the relationship between molecular features and olfactory perception is inherently nonlinear. Intriguingly, random forest with only the top 5 features achieves similar performance as random forest with all 4884 features for almost all olfactory qualities (Fig. 5A and Supplementary Figure S5). The only exception is “intensity,” for



**Figure 5:** The performance of random forest using top features. From left to right, the Pearson's correlation coefficients of intensity and pleasantness and 19 semantic descriptors from 5-fold cross-validations are shown as boxplots. The red model is the random forest using all chemoinformatic features. (A) The performance of random forest using the top 5, 10, 15, and 20 features ranked by delta error. (B) The performance of random forest using the top 5, 10, 15, and 20 features ranked by Pearson's correlation.

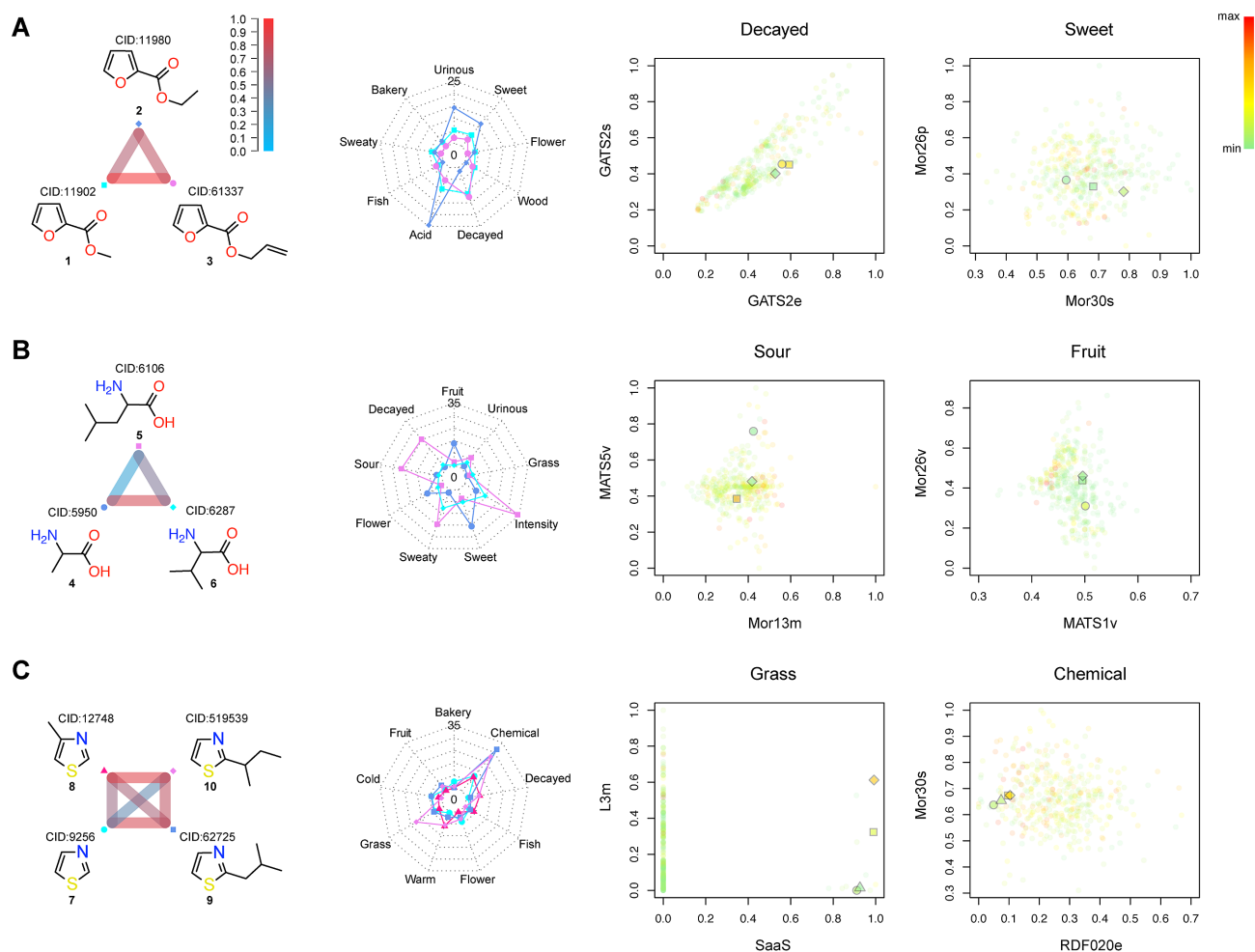
which the top 15 features are adequate. This result indicates that a small set of chemical features is often sufficient to predict the odor of a molecule. We also test the performance of random forest using features ranked by Pearson's correlation (Fig. 5B). The predicting power of these features is lower due to collinearity and redundancy. For example, the top 10 features for "garlic" quality are all related to the number of sulfur atoms, although they display very high correlation values (Supplementary Table S3).

### Deciphering the divergent multi-odor profiles of structural analogs

Structurally similar compounds with distinct odor profiles were observed in triads and a tetrad of molecules. The first example is 3 furoate esters (Fig. 6A). If we compare the functional groups of these 3 molecules, methyl 2-furoate (1) and ethyl 2-furoate (2) are more similar, while allyl 2-furoate (3) has a unique alkenyl group. Intriguingly, the pairwise correlations between them across 21 perceived olfactory qualities reveal that 2 is the odd one in terms of odor. It is clearly shown in the radar chart of selected odor qualities. Compound 2 has intense "sweet," "acid," and "urinous" characters, whereas 1 and 3 display more

"decayed" odor. The second triad of molecules comprises common L-amino acids: alanine (4), leucine (5), and valine (6) (Fig. 6B). Their odor profiles differ a lot, especially between 4 and 5; 4 has "fruit," "sweet," and "flower" odors, whereas 5 is characterized by "sour," "decayed," "sweaty," and "intensity"; 6 has a relatively similar odor profile as 4. The odd one in structural terms is 4 as it has the smallest side chain, whereas the odd one in terms of odor is 5. The last group of molecules includes thiazole (7) and its derivatives (8–10) (Fig. 6C); 9 stands out because of its signature "grass" odor, whereas both 9 and 10 have a very high "chemical" odor.

Our random forest model distinguishes the multi-odor profiles of structural analogs using complex molecular features. Although these analogs are extremely similar in terms of chemical structure and functional group, the values of their 2- and 3-dimensional molecular descriptors are distinct. The average rating of each molecule is represented by its color, and the structural analogs mentioned above are shown in a larger size (Fig. 6, right panels). The top features used by our random forest model clearly separate the structurally similar molecules with dissimilar odor attributes. For example, 10, with a strong "grass" odor (top right orange diamond in Fig. 6C, the 3rd panel), has large SaaS and L3m values, whereas 7 and 8, with a weak "grass" odor



**Figure 6:** Distinguishing different odor profiles of structurally similar molecules by random forest. The odor profiles of (A) 3 furoate esters, (B) 3 amino acids, and (C) 4 thiazole derivatives. The left panel shows the pairwise correlations between structurally similar molecules. The color of each edge represents the correlation value across 21 perceptual attributes. The middle panel shows the radar charts of selected odor attributes. The symbol and color correspond to the molecule on the left. The right panel displays the projections of all molecules onto selected discriminative feature spaces. The color of each spot represents the relative strength of the perceived rating, averaged among 49 individuals. The larger symbols correspond to the molecules on the left.

(bottom-right green circle and triangle), have relatively small values; 9 (middle right yellow square), with medium “grass” odor, has around average values among the tetrad.

## Discussion

The complex and sophisticated signaling of diverse odorants has fascinated scientists for many decades, yet the molecular mechanisms of olfactory perception are still not fully understood. One odorant interacts with a broad range of olfactory receptors, and each olfactory receptor recognizes multiple odorants, leading to the complicated tuning of olfactory perception [39, 40]. In addition, neuron firing is intrinsically nonlinear in nature, requiring the membrane potential to be raised above threshold. Therefore, a nonlinear random forest model is well suited to the olfactory prediction and avoids overfitting, given a comparatively small sample size and much larger feature spaces. Moreover, random forest is relatively robust to label noise and outliers [32], considering the vast variability of odor ratings among individuals.

The linguistic descriptions of smells vary among individuals, especially when they lack experience and training [41]. This

finding suggests that using a low-variant dataset of odorants rated by professional perfumers may further improve the performance of predictive models. Besides, using semantic descriptors itself introduces biases, and alternative approaches such as perceptual similarity rating of odorants should be considered [26]. Recognizing that extra Morgan-NSPDK features created by matching target molecules against reference odorants increase the predicting performance [30], a larger training set of diverse molecules, including natural odorant products, will be helpful to build more accurate models.

Our random forest model potentially provides an alternative for rational odorant design [42, 43]. In addition to modifications of a natural odorant product, the perceptual dataset used in this study consists of many untested molecules, providing new odorant scaffolds of different semantic qualities. Moreover, a small set of top-ranking features estimated by the random forest model is sufficient to accurately predict human olfactory perception, largely reducing the input feature spaces. This model is potentially useful for evaluation of new molecules, and modification of these discriminative features provides an alternative for rational odorant design. Like the association of functional groups with certain odors, this study may link complex



chemoinformatic features to a broader range of odors, providing a useful perspective for understanding olfactory perception mechanisms.

## Methods

### Random forest model

Random forest is an ensemble learning algorithm for regression and classification [32]. In a random forest, each decision tree is built from a random sampling with replacement (bootstrap samples). Furthermore, a random set of features are used to determine the best split at each node during the construction of a tree. As a result of averaging many trees (100 is used in our model), overfitting is avoided and the effects of outliers and noises are reduced.

We use perceptual ratings as targets and chemical descriptors as features to train random forest models. For the combination of 49 individuals and 21 perceptual attributes, we have 1029 (49\*21) models in total. We further consider the average ratings among 49 individuals as a population rating and combine it with the individual rating as prediction targets to train our models (see the “Integrating individual and population ratings” section below).

### Preprocessing of the dataset

There were many cases in which subjects indicated that they smelled nothing, so the intensity rating was automatically set to “0” and the ratings for other perceptual attributes were left blank (NaN); therefore, we have removed all the “NaN” entries. For the intensity attribute, we used the target values at “1/1000” dilution. For pleasantness and 19 semantic attributes, we used target values at “high” concentration as a set of examples, and the average value at both “high” and “low” concentrations as another set of examples. As the original number of the sample (407) is relatively small, combining high and low concentrations doubles the sample size, and this step is crucial to achieve high performance. There are 20 replicated molecules, which were selected in the original Rockefeller University Smell Study. In general, the ratings were consistent between the 2 replicates [33]. In our study, we treat replicates as separate examples as the fraction is small (20/407 = 0.049), which doesn’t affect the results. The input molecular features were scaled to values between 0 and 1. The scaling formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where  $x$  is the original value and  $x'$  is the scaled value.

### Selection of base-learner

To address the large variability of perceived odor qualities among individuals, we tried a range of different machine learning algorithms (linear, ridge, SVM with rbf kernel, and random forest with 100 trees) to find the best-performing base-learner. The regularization  $\alpha$  in the ridge is 10. The penalty parameter  $C$  and coefficient  $\gamma$  of the SVM rbf kernel are 1000 and 0.01, respectively. All other parameters are the default ones. We applied a 5-fold cross-validation to the training data (407 molecules) and evaluated the performance based on the correlations of the 21 perceptual attributes between the predicted and observed ratings. Random forest outperformed other base-

learners and was used in the follow-up improvement of our model.

### Integrating individual and population ratings

The perceptual rating of attributes varies greatly. To reduce the effects of noise and outliers, we introduce a weighting factor,  $\alpha$ , as the weight for individual ratings and  $(1 - \alpha)$  as the weight for population ratings. The reweighted target value  $y$  is given as:

$$y = \alpha \times y_{\text{individual}} + (1 - \alpha) \times y_{\text{population}},$$

where  $y_{\text{individual}}$  is the rating from an individual and  $y_{\text{population}}$  is the average rating from 49 individuals. Different values of  $\alpha$  were tested and evaluated by the correlation of 21 perceptual attributes.  $\alpha = 0.2$  had the best performance and was used in our final model.

### Creating name features of molecules

In the past, sliding window-based (overlapping patterns) strategies were applied successfully to develop residue-level predictions [44, 45]. We used a sliding window of 4-letter size to extract features from the molecule names. For example, 4-letter indexing generated a total of 7 sliding windows from “acetic acid” (ACET, CETI, ETIC, TIC., IC.A, \_ACI, ACID). We created 11 786 binary name features from all molecule names using this sliding window approach. If a window pattern is present in the molecule name, “1” was assigned to that feature; otherwise, “0” was used while creating input name features.

### Evaluation of the importance of each feature by random forest

The importance of each feature was evaluated by permuting the values across observations and computing the increase in prediction error by random forest. The increased delta error of each chemical feature for all 21 olfactory attributes was calculated and ranked. Larger delta error implies that the feature is more important and discriminative in prediction.

### Availability of supporting data

The DREAM olfaction challenge dataset, model details, and source code are available at [https://github.com/Hongyang449/olfaction\\_prediction\\_manuscript](https://github.com/Hongyang449/olfaction_prediction_manuscript).

Snapshots of the code, molecular descriptors, and the olfactory perception data and chemoinformatic features of odorant molecules are also available from the GigaScience database, GigaDB [46].

### Additional files

Table S1. The top 5 features ranked by random forest delta error or Pearson’s correlation.

Table S2. The top 20 features of 21 perceptual attributes ranked by random forest delta error.

Table S3. The top 20 features of 21 perceptual attributes ranked by Pearson’s correlation.

Table S4. The Pearson’s and Spearman’s correlations of top 500 features ranked by Pearson’s correlation.

Figure S1. The 2D chemical structures of 476 odorant molecules.

Figure S2. The intensity ratings for all molecules at low and high concentrations from 49 individuals.

Figure S3. The density distributions of the intensity ratings for all molecules from 49 individuals.

Figure S4. The performance of random forest using chemical and name features.

Figure S5. The performance of random forest using top features ranked by delta error.

## Completing interests

The authors declare that they have no competing interests.

## Funding

This work is supported by National Science Foundation 1452656 and Alzheimer's Association BAND-15-367116 (Biomarkers Across Neurodegenerative Diseases Grant 2016).

## Author Contributions

Y.G. conceived and designed the prediction algorithm. Y.G. and H.L. performed computational analysis of the observed and predicted data. H.L. analyzed the discriminative chemoinformatic features and prepared figures. H.L., B.P., G.O., and Y.G. contributed to the writing of the manuscript. All authors read and approved the final manuscript.

## References

- Gaillard I, Rouquier S, Giorgi D. Olfactory receptors. *Cell Mol Life Sci* 2004;61(4):456–69.
- Buck LB. Olfactory receptors and odor coding in mammals. *Nutr Rev* 2004;62(11):184–188.
- Read JCA. The place of human psychophysics in modern neuroscience. *Neuroscience* 2015;296:116–29.
- Sell CS. On the unpredictability of odor. *Angew Chem Int Ed* 2006;45(38):6254–61.
- Laska M, Teubner P. Olfactory discrimination ability for homologous series of aliphatic alcohols and aldehydes. *Chem Senses* 1999;24(3):263–70.
- Boesveldt S, Olsson MJ, Lundström JN. Carbon chain length and the stimulus problem in olfaction. *Behav Brain Res* 2010;215(1):110–3.
- Keller A, Zhuang H, Chi Q et al. Genetic variation in a human odorant receptor alters odour perception. *Nature* 2007;449(7161):468–72.
- Chastrette M. Trends in structure-odor relationship. *SAR QSAR Environ Res* 1997;6(3–4):215–54.
- Boelens H. Structure-activity relationships in chemoreception by human olfaction. *Trends Pharmacol Sci* 1983;4:421–6.
- Edwards PA, Jurs PC. Correlation of odor intensities with structural properties of odorants. *Chem Senses* 1989;14(2):281–91.
- Mamlouk AM, Chee-Ruiter C, Hofmann UG et al. Quantifying olfactory perception: mapping olfactory perception space by using multidimensional scaling and self-organizing maps. *Neurocomputing* 2003;52:591–7.
- Zarzo M, Stanton DT. Identification of latent variables in a semantic odor profile database using principal component analysis. *Chem Senses* 2006;31(8):713–24.
- Mainland JD, Lundström JN, Reisert J et al. From molecule to mind: an integrative perspective on odor intensity. *Trends Neurosci* 2014;37(8):443–54.
- Kermen F, Chakirian A, Sezille C et al. Molecular complexity determines the number of olfactory notes and the pleasantness of smells. *Sci Rep* 2011;1(1):206.
- Zarzo M. Hedonic judgments of chemical compounds are correlated with molecular size. *Sensors* 2011;11(12):3667–86.
- Khan RM, Luk C-H, Flinker A et al. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *J Neurosci* 2007;27(37):10015–23.
- Menashe I, Man O, Lancet D et al. Different noses for different people. *Nat Genet* 2003;34(2):143–4.
- Keydar I, Ben-Asher E, Feldmesser E et al. General olfactory sensitivity database (GOSdb): candidate genes and their genomic variations. *Hum Mutat* 2013;34(1):32–41.
- Perez M, Nowotny T, d'Etterre P et al. Olfactory experience shapes the evaluation of odour similarity in ants: a behavioural and computational analysis. *Proc R Soc B* 2016;283(1837):20160551.
- Chrea C, Valentin D, Sulmont-Rossé C et al. Culture and odor categorization: agreement between cultures depends upon the odors. *Food Qual Preference* 2004;15(7–8):669–79.
- Ayabe-Kanamura S, Schicker I, Laska M et al. Differences in perception of everyday odors: a Japanese-German cross-cultural study. *Chem Senses* 1998;23(1):31–38.
- Levitán CA, Ren J, Woods AT et al. Cross-cultural color-odor associations. *PLoS One* 2014;9(7):e101651.
- Haddad R, Khan R, Takahashi YK et al. A metric for odorant comparison. *Nat Methods* 2008;5(5):425–9.
- Koulakov AA, Kolterman BE, Enikolopov AG et al. In search of the structure of human olfactory space. *Front Syst Neurosci* 2011;5:65.
- Castro JB, Ramanathan A, Chennubhotla CS. Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS One* 2013;8(9):e73289.
- Snitz K, Yablonka A, Weiss T et al. Predicting odor perceptual similarity from odor structure. *PLoS Comput Biol* 2013;9(9):e1003184.
- Dravnieks A. Odor quality: semantically generated multidimensional profiles are stable. *Science* 1982;218(4574):799–801.
- Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *CCHTS* 2006;9(3):213–28.
- Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasitikul V. Advances in computational methods to predict the biological activity of compounds. *Expert Opin Drug Discovery* 2010;5(7):633–54.
- Keller A, Gerkin RC, Guan Y et al. Predicting human olfactory perception from chemical features of odor molecules. *Science* 2017;355(6327):820–6.
- Saez-Rodriguez J, Costello JC, Friend SH et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* 2016;17(8):470–86.
- Breiman L. Machine Learning. 2001;45:5. <https://doi.org/10.1023/A:1010933404324>, Accessed 7 April, 2017.
- Keller A, Vosshall LB. Olfactory perception of chemically diverse molecules. *BMC Neurosci* 2016;17(1):55.
- Todeschini R, Consonni V, eds. *Molecular Descriptors for Chemoinformatics*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2009. <http://doi.wiley.com/10.1002/9783527628766>, Accessed 22 March, 2017.
- The complete list of molecular descriptors calculated by Dragon 6, [http://www.talete.mi.it/products/dragon-molecular\\_descriptor\\_list.pdf](http://www.talete.mi.it/products/dragon-molecular_descriptor_list.pdf), Accessed 7 April, 2017.

36. Godden JW, Bajorath J. Shannon entropy—a novel concept in molecular descriptor and diversity analysis. *J Mol Graph Model* 2000;**18**(1):73–6.
37. Godden JW, Bajorath J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J Chem Inf Comput Sci* 2002;**42**(1):87–93.
38. Godden JW, Bajorath J. An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb Sci* 2003;**22**(5):487–97.
39. Zhao H. Functional expression of a mammalian odorant receptor. *Science* 1998;**279**(5348):237–42.
40. Malnic B, Hirono J, Sato T et al. Combinatorial receptor codes for odors. *Cell* 1999;**96**(5):713–23.
41. Livermore A, Laing DG. Influence of training and experience on the perception of multicomponent odor mixtures. *J Exp Psychol Hum Percept Perform* 1996;**22**(2):267–77.
42. Turin L, Yoshii F. Structure-odor relations: a modern perspective. *Handbook of Olfaction and Gustation* 2003; 11–25. [https://www.researchgate.net/profile/Fumiko\\_Yoshii/publication/241197841\\_Structure-odor\\_relations\\_A\\_modern\\_perspective/links/02e7e52c762d501279000000.pdf](https://www.researchgate.net/profile/Fumiko_Yoshii/publication/241197841_Structure-odor_relations_A_modern_perspective/links/02e7e52c762d501279000000.pdf), Accessed 19 January, 2018.
43. Turin L. *Chemistry and Technology of Flavors and Fragrances*. Rowe DJ, ed. Oxford, UK: Blackwell Publishing Ltd.; 2004. <http://doi.wiley.com/10.1002/9781444305517>, Accessed 22 March, 2017.
44. Panwar B, Gupta S, Raghava GP. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinformatics* 2013; **14**:44.
45. Panwar B, Raghava GP. Prediction of uridine modifications in tRNA sequences. *BMC Bioinformatics* 2014;**15**:326.
46. Li H, Panwar B, Omenn GS et al. Supporting data for “Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features.” *GigaScience Database* 2017. <http://dx.doi.org/10.5524/100384>, Accessed 15 January, 2018.