Learning Olfactory Mixture Similarity: CWYK team

Vahid Satarifard [1], Wenjie Yin [2], Mårten Björkman [2], Kobi Snitz [3], Danica Kragic [2], Nicholas Christakis [1], Noam Sobel [3], Aharon Ravia [4*]

1 - **Y**ale Institute for Network Science, Yale University, New Haven, CT 06520

2 - **K**TH Royal Institute of Technology Stockholm, Sweden

3 - Department of Neurobiology, **W**eizmann Institute of Science, Rehovot, Israel

4 - **C**ornell Tech, Cornell University, New York, United States

* - Corresponding author: aharon.ravia@cornell.edu

Summary Sentence

We used XGBoost with selected Dragon chemical features and olfactory semantic descriptors of single molecules and binary mixtures of augmented data to train a model for olfactory mixture similarity predictions.

We will make our submission public as part of the challenge archive, the source code is available at: https://github.com/Satarifard/CWYK-Olfboost.

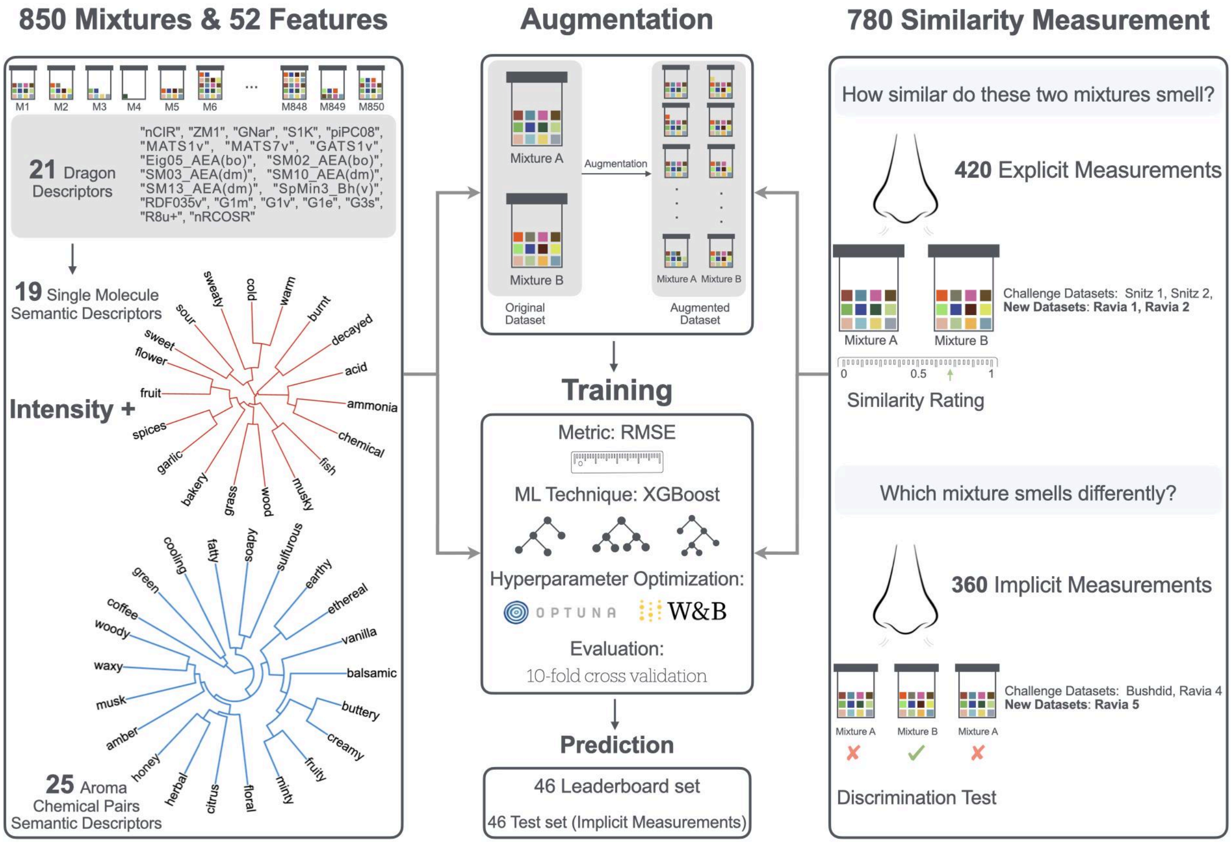
850 Mixtures & 52 Features
M1
M2
M3
M4
M5
M6
...
M848
M849
M850
21 Dragon Descriptors
"nCIR", "ZM1", "GNar", "S1K", "piPC08", "MATS1v", "MATS7v", "GATS1v", "Eig05_AEA(bo)", "SM02_AEA(bo)", "SM03_AEA(dm)", "SM10_AEA(dm)", "SM13_AEA(dm)", "SpMin3_Bh(v)", "RDF035v", "G1m", "G1v", "G1e", "G3s", "R8u+", "nRCOSR"
19 Single Molecule Semantic Descriptors
Intensity +
sweaty
cold
warm
burnt
decayed
acid
ammonia
chemical
fish
musky
wood
grass
bakery
garlic
spices
fruit
flower
sweet
sour
cooling
fatty
soapy
sulfurous
earthy
ethereal
vanilla
balsamic
buttery
creamy
fruity
minty
floral
citrus
herbal
honey
amber
musk
waxy
woody
coffee
green
25 Aroma Chemical Pairs Semantic Descriptors
Augmentation
Mixture A
Mixture B
Augmentation
Mixture A
Mixture B
Original Dataset
Augmented Dataset
Training
Metric: RMSE
ML Technique: XGBoost
Hyperparameter Optimization:
OPTUNA
W&B
Evaluation:
10-fold cross validation
Prediction
46 Leaderboard set
46 Test set (Implicit Measurements)
780 Similarity Measurement
How similar do these two mixtures smell?
420 Explicit Measurements
Mixture A
Mixture B
Challenge Datasets: Snitz 1, Snitz 2,
New Datasets: Ravia 1, Ravia 2
0
0.5
1
Similarity Rating
Which mixture smells differently?
360 Implicit Measurements
Mixture A
Mixture B
Mixture A
Challenge Datasets: Bushdid, Ravia 4
New Datasets: Ravia 5
Discrimination Test

Figure 1: This schematic workflow illustrates the complete workflow involved in our study, including model input, feature selection, data augmentation, and model training.The hierarchical clustering of semantic descriptors is conducted based on the cosine similarity of word embeddings derived from the Word2Vec model [9].

Background/Introduction

In the past, both chemical descriptors and olfactory semantic descriptors of single molecules were separately used to predict olfactory mixture similarity [1-3]. The chemical descriptors have been mapped to the semantic space of single molecules [4], and then used for training regression models such as lasso to predict the olfactory mixture similarity [3]. Such models implicitly assume there is no interaction between molecules in olfactory space. However, experimental results show that antagonistic interactions between odorants cause nonlinear responses of olfactory receptors to complex mixtures, which enhances the encoding capacity of olfactory information [5]. To account for such non-linearity, we used a chemical-aroma pair model that is trained on experimental measurements of semantic labels of binary mixtures [6]. By combining these sets of single molecule and binary mixture semantic labels along with the compound identifiers (CID), we trained a machine learning model to predict the olfactory mixture similarity. We used XGBoost as the machine learning technique that has been shown to be advantageous for smaller datasets such as the olfactory mixture dataset.

Regarding feature selection, we used 21 Dragon chemical descriptors that have been shown to be more relevant to the olfactory mixture [1,2]. We utilized these chemical descriptors to train a model for predicting 19 semantic labels as well as the intensity and pleasantness of single molecules [3,4]. Additionally, we employed an aroma-chemical pair model to predict binary mixture olfactory semantic descriptors [6]. Moreover, we found that the quantile transformer, along with data augmentation, improves the performance of the predictive model on the leaderboard set results.

Methods

The schematic workflow used in our study is summarized and shown in Figure 1. In the following we will discuss the details of different steps leading to the final predictions.

1- Dataset Curation.

The original dataset provided by the DREAM challenge organizers [7] included four different datasets from three separate studies: Snitz 1, Snitz 2, Ravia 4, and Bushdid. In the early stages of model training, we discovered several parsing errors in the Snitz 1 and Snitz 2 datasets, as well as numerous mislabeling issues and missing CIDs in the Bushdid dataset. Moreover, we realized the possibility of adding three more experiments to the original dataset that could potentially improve the model's performance. In light of these points, we curated a new dataset from original publications [1,2,8] and added three more datasets: Ravia 1, Ravia 2, and Ravia 5, to the full training dataset. We ultimately compiled a dataset consisting of 850 unique mixtures and 780 similarity measurements. To be compatible with the challenge original design we excluded the leaderboard dataset from the training dataset.

2- Feature Selection.

Single molecule semantic descriptors: Inspired by previous studies [3,4], we utilized 21 Dragon chemical descriptors [1,2], including 'nCIR', 'ZM1', 'GNar', 'S1K', 'piPC08', 'MATS1v', 'MATS7v', 'GATS1v', 'Eig05_AEA(bo)', 'SM02_AEA(bo)', 'SM03_AEA(dm)', 'SM10_AEA(dm)', 'SM13_AEA(dm)', 'SpMin3_Bh(v)', 'RDF035v', 'G1m', 'G1v', 'G1e', 'G3s', 'R8u+', and 'nRCOSR', in conjunction with an XGBoost regressor to predict pleasantness, intensity, and 19 semantic descriptors: 'bakery', 'sweet', 'fruit', 'fish', 'garlic', 'spices', 'cold', 'sour', 'burnt', 'acid', 'warm', 'musky', 'sweaty', 'ammonia', 'decayed', 'wood', 'grass', 'flower', 'chemical' for each molecule.

Binary mixture semantic descriptors: We trained a model with 33 aroma-chemical pair semantic descriptors [6], which included: 'alliaceous', 'coffee', 'floral', 'fruity', 'green', 'herbal', 'minty', 'sulfurous', 'waxy', 'balsamic', 'aldehydic', 'buttery', 'caramellic', 'creamy', 'earthy', 'ethereal', 'fatty', 'fermented', 'musk', 'soapy', 'spicy', 'tropical', 'woody', 'amber', 'cooling', 'citrus', 'animal', 'berry', 'honey', 'vanilla', 'nutty', 'musty', 'camphoreous'. We dropped 8 labels 'aldehydic', 'fermented', 'spicy', 'tropical', 'animal', 'berry', 'Nutty' and 'musty' due to their low performance with an AUROC below 0.5, while all other labels showed good performance with average AUROC of 0.80. For each mixture, we averaged over the predictions of semantic descriptors for all possible binary combinations of compounds in the mixture.

Finally, each mixture was represented by a binary vector of molecular IDs (235 unique molecules), combined with intensity and the highest single molecule semantic descriptor values, as well as 25-dimensional binary mixture semantic descriptors. This set of features leverages both single molecule and binary mixture semantic descriptors, along with molecular overlap of mixtures, to enhance the prediction accuracy and interpretability of the model.

3- Data Augmentation.

To address the challenge of model overfitting due to a limited dataset, we implemented a data augmentation strategy by manipulating molecular counts in mixtures. This involved either increasing or decreasing the experimental similarity by a fixed amount. Specifically, if a molecule present in mixture A was added to mixture B, which lacked it, the experimental distance between the two mixtures decreased by a predetermined amount, a parameter fine-tuned through hyperparameter tuning. This adjustment resulted in significant data expansion: from an initial count of 780 similarity measurements (treating A-B and B-A as distinct pairs), the dataset grew to 4,078 synthetic similarity entries with the addition of one molecule per mixture, and to 7,331 entries with two molecular adjustments. We further augmented the data by excluding common molecules; specifically, if mixture A and mixture B shared a molecule, removing it from both increased their experimental distance by a predetermined amount. This step increased the total dataset size to 10,692, of which approximately 93% was synthetic data. Though not fully optimized, preliminary tests indicated that these augmentation steps, adding up to two molecules and removing one, improved the model performance on the leaderboard set.

4- Model Training and Prediction.

For model training, we utilized XGBoost as the primary machine learning (ML) technique to train predictive models for olfactory mixture similarity. The fast training speed of this technique is well-suited for small datasets and requires fewer computational resources. Built-in regularization helps prevent overfitting, which is crucial for handling small data sets. We employed root mean squared error (RMSE) as the metric for calculating the loss function and used Optuna and wandb for hyperparameter optimization. Furthermore, we implemented 10-fold cross-validation to evaluate the model's performance. Finally, we used the trained model to predict the outcomes on the test set.

Conclusion/Discussion

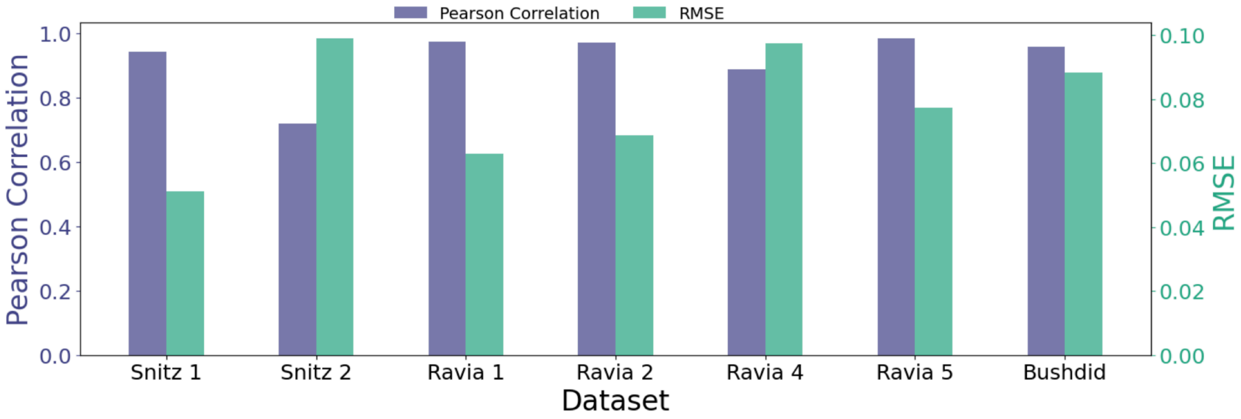

Figure 2: Model Performance on Each Dataset. The purple bars display the Pearson correlation, and teal bars show the root mean squared error (RMSE).

In summary, we have utilized XGBoost to train a predictive ML model for olfactory mixture similarity. To address the limited size of our experimental data, we implemented data augmentation and various regularization techniques, along with 10-fold cross-validation to prevent overfitting. Our model demonstrates excellent performance on the full dataset (RMSE=0.08 and Pearson Correlation=0.89) on each separate dataset, as illustrated in Figure 2, as well as on the leaderboard dataset. In addition to our discussed efforts, we experimented with using single molecular semantic descriptors from Open-POM [10] as input features, in combination with symbolic regression, random forest, elastic net, and lasso as ML techniques. However, none of these efforts resulted in improved performance particularly for correlations. It is worth mentioning that since the 46 hidden test sets included only implicit measurements, one might expect that a training model based on 360 implicit measurements (Figure 1) could lead to a better predictive model on the test set. We look forward to exploring this once the results of the test set are revealed at the conclusion of the DREAM challenge. Looking ahead, we plan to refine our augmentation techniques and parameters to enhance the dataset for future olfactory mixture studies. Finally, we believe that as the number of similarity measurements of olfactory mixtures increases in the future, the potential for training more robust and predictive models will drastically improve, as currently the small size of these measurements limits the ML algorithms from learning more general aspects of the olfactory mixtures.

References

[1] Snitz, Kobi, et al. "Predicting odor perceptual similarity from odor structure." PLoS computational biology 9.9 (2013): e1003184.

[2] Ravia, Aharon, et al. "A measure of smell enables the creation of olfactory metamers." Nature 588.7836 (2020): 118-123.

[3] Dhurandhar, Amit, et al. "Expansive linguistic representations to predict interpretable odor mixture discriminability." Chemical Senses 48 (2023): bjad018.

[4] Keller, Andreas, et al. "Predicting human olfactory perception from chemical features of odor molecules." Science 355.6327 (2017): 820-826.

[5] Pfister, Patrick, et al. "Odorant receptor inhibition is fundamental to odor encoding." Current Biology 30.13 (2020): 2574-2587.

[6] Sisson, Laura. "Olfactory Label Prediction on aroma-chemical Pairs." arXiv preprint arXiv:2312.16124 (2023).

[7] DREAM Olfactory Mixtures Prediction Challenge (Project SynID: **syn53470621**)

[8] Bushdid, Caroline, et al. "Humans can discriminate more than 1 trillion olfactory stimuli." Science 343.6177 (2014): 1370-1372.

[9] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[10] Lee, Brian K., et al. "A principal odor map unifies diverse tasks in olfactory perception." Science 381.6661 (2023): 999-1006.

Authors Statement

VS, WY, and AR performed data preprocessing, algorithm implementation, model training, and drafting the write-up. **MB, KS, DK, NC, and NS** provided supervisory guidance. All authors discussed the results and contributed to the final write-up.