DREAM Olfactory Mixtures Prediction Challenge 2025- Write Up

Alicia Wang, Sophia Shiu
University of Maryland, Boston University

Summary Sentence

We used random forest regression for both task 1 and task 2 as we wanted to predict numerical values.

Background/Introduction

Predicting human olfactory perception from chemical structure remains a significant challenge in computational neuroscience. While past work has shown that molecular descriptors can capture useful information about a compound's structure, turning that into accurate predictions of perceptions is still challenging.

In this study, we aimed to predict human perceptual ratings for both individual odor molecules and mixtures. We chose Random Forest Regression due to its ability to handle high-dimensional data. For Task 1, we focused on monomolecular stimuli, leveraging Mordred descriptors aligned with the stimulus definitions. For Task 2, which involved complex mixtures, we aggregated component-level features (including Mordred, Morgan fingerprints, and OpenPOM perceptual vectors) to derive averaged representations of each mixture.

The motivation for our approach was to balance interpretability and predictive power using ensemble models, while integrating diverse molecular features across multiple datasets to improve generalization to unseen compounds and mixtures.

Methods

Two independent pipelines were developed for Task 1 with a single-molecule stimuli and Task 2 with multiple components in mixtures. Each pipeline consisted of integrating datasets, cleaning and preprocessing, training a multi-output regression model, and applying the model to the leaderboard for outcomes. All of the processing was performed in Python 3.11 using pandas, numpy, and scikit-learn.

For Task 1, we began with merging certain features together to create a more coherent,organized feature table. We combined TASK1_Stimulus_definition.csv with Mordred_Descriptors.csv on the “molecule” field. Additionally, merged the molecular feature table with TASK1_training.csv on the “stimulus field”. We then moved onto feature selection to remove any features that were not used for inputs or are not numeric-columns as it would complicate the prediction process. Columns such as “stimulus”, “molecule”, “dilution”, “solvent”, “Intensity_label”, “Intensity”, and “Pleasantness” were all removed as they are non-predictive metadata. We ended feature processing with cleaning and scaling using StandardScaler to remove zero-variance columns. Infinite values were replaced with NAN and missing values were replaced with 0.

Moving onto Task 2, the beginning methods were similar as we started with merging descriptors from Mordred_Descriptors.csv and Morgan_fingerprints.csv on “molecule”. Additionally, we added SMILES (Simplified Molecular Input Line Entry System) and external data from OpenPOM_Dream_RATA.csv to be able to bring perceptual ratings. Afterward, the merged molecular feature table was joined on component identifiers to create a full feature set for each component. This allowed us to aggregate the features and split the components into lists based on their IDs. For each stimulus, numeric features for every component were taken out. If there were multiple ones, the features would be averaged out for a single feature vector per mixture. If the mixture had no features or they were missing, then NaN vector values would take place and turn into zeros for the prediction process.

Both tasks had us scale the features using the training set and align the columns with the training feature set. As the process goes onto the algorithm and training portion, we developed the idea to use a Random Forest Regressor through MultiOutRegressor from scikit-learn for both tasks. This algorithm allowed for independent regression for every sensory descriptor. We set the hyperparameters to be the default setting of n_estimators =100, and random_state = 42. The train-validation split was 80-20 for Task 1 and 90-10 for Task 2 as we found them to be the best divide for each task’s outcome. For the last step of the prediction, we developed model evaluation. We used the Root Mean Squared Error (RMSE) as the main validation metric for both tasks. For Task 2, we added the Mean Cosine Distance between the predicted and true descriptor vectors and the Mean Pearson Correlation across descriptors.

While we did not use outside data, we did incorporate outside knowledge to better prepare and understand the tasks' missions and desired outcomes. We learned about how molecular structures are represented which enabled us to learn about SMILES strings, and molecular fingerprints. In addition, we learned about the overall pathway from smells to distinction. We absorb readings on the olfactory system and the development through odor molecules to become odorant patterns. Furthermore, we dove into the idea of valence and intensity to a human's overall sensory experience. As opinions have heavily influenced how one perceives certain scents, it is important to get an idea of a specific scale our algorithms can depend on. These allowed us to develop full confidence in our analysis and predictions, to ensure that the outcomes we provide reflect highly on the challenges and desired output needed to understand more about the olfactory system.

Conclusion/Discussion

Overall, using Random Forest Regression allowed us to create a reliable and interpretable model for both single-molecule and mixture olfactory prediction tasks. Task 1 was simpler to model, as each stimulus corresponded to a single molecule with directly associated descriptors. In contrast, Task 2 involved predicting perceptual ratings for mixtures. However, aggregating component-level features provided a reasonable baseline for mixture representation.

We found that combining multiple types of molecular data helped the model generalize better, especially unfamiliar mixtures. One of the main challenges was handling missing or incomplete feature sets, particularly in Task 2, which we addressed by imputing missing values with zeros to maintain model stability.

Future improvements could involve models that capture interactions between mixture components, moving beyond simple feature averaging. These findings highlight the effectiveness of combining diverse molecular features and suggest that even simple aggregation methods can capture meaningful patterns in human odor perception.

References (limit 10 references)

1. Boldini, D., Ballabio, D., Consonni, V., Todeschini, R., Grisoni, F., & Sieber, S. A. (2024, March 25). *Effectiveness of molecular fingerprints for exploring the chemical space of natural products*. Journal of cheminformatics. https://pmc.ncbi.nlm.nih.gov/articles/PMC10964529/
2. Megalakaki, O., Ballenghein, U., & Baccino, T. (2019, February 1). *Effects of valence and emotional intensity on the comprehension and memorization of texts*. Frontiers in psychology. https://pmc.ncbi.nlm.nih.gov/articles/PMC6367271/
3. U.S. Department of Health and Human Services. (2025a, May 8). *How the nose decodes complex odors*. National Institutes of Health. https://www.nih.gov/news-events/nih-research-matters/how-nose-decodes-complex-odors
4. DREAM Olfactory Mixtures Prediction Challenge 2025 (syn64743570)

Authors Statement

- Data Acquisition and Curation of Task 1 and 2: Alicia Wang and Sophia Shiu
- Methodology of Task 1 and 2: Alicia Wang and Sophia Shiu
- Writing- Statement Section: Sophia Shiu
- Writing- Background Section: Alicia Wang
- Writing- Method Section: Sophia Shiu
- Writing- Conclusion Section: Alicia Wang