DREAM Olfactory Mixtures Prediction Challenge 2025

SystemsCBLab_OMP: Submission Write-up

Sheerin Irfaanaa[1]*, Ashok Palaniappan*[1]

[1]Systems Computational Biology Lab, School of Chemical and Biotechnology, SASTRA Deemed University, Thanjavur India.

- Will you be able to make your submission public as part of the challenge archive? **Yes**

Github repo for code & data:
https://github.com/apalania/DREAM-Olfactory-Mixtures-Prediction-Challenge-2025

Summary Sentence: We used an ensemble of hyperbolic neural networks and random forests trained on selected (embeddings, engineered features) from the DREAM challenge datasets for the tasks.

Introduction: Our principal innovations in this contribution include:

- Design of feature space spanned by **MAP4 fingerprints (1024 count features), Mordred descriptors** (1826 descriptors), and **MACCS keys (166 keys)**
- Use of **ChEMBERTa pretrained model to generate embeddings** (767-dimensional) from SMILES strings
- Feature selection on the designed feature spaces and embeddings via SelectKBest, PCA and MultiTask-Lasso.
- Consolidation of three different datasets for training **Olfaction percept** models
- Experiments with two different model classes for training:
  - **Hyperbolic neural networks** (HNNs) with multi-task output layers and custom loss functions that combine pearson and cosine loss
  - Random Forest MultiOutput Regressor with 100 estimators and default hyperparameters – predicting all 52 sensory attributes simultaneously, but independent RF models for each target dimension. Custom metrics evaluated included Pearson correlation score and Cosine similarity score; models optimized on 'mse'.
- Ensembling the best models for inference on the test set

We chose to work with HNNs since the geometry of olfaction space is known to be hyperbolic. We used HNNs with three fully connected layers (Input → 128 units → 64 units → Output (52 perceptual attributes)), and maintaining 'Tanh' Activation between hidden layers in hyperbolic space.

Below we provide more task-specific details.

Methods:: Task-1: Olfactory Percept - single molecules

I. Feature engineering:

As noted above, for each molecule we generated MAP4 fingerprints, MACCS keys, and Mordred descriptors. For task-1, the following metadata were used:

1. Dilution (log-transformed or other),
2. solvent (with three classes; Label_Encoder) and
3. intensity_label (with two classes; Label_Encoder).

We also generated the ChEMBERTa embeddings for each molecule

II. Dataset consolidation

We used the following feature selection techniques:

(i) PCA-99 & PCA-95 on the Embeddings

(ii) SelectKBest on the tabular feature set using F-regression scores to identify top 50 features.

From these results, we engineered the following datasets for predicting olfactory percepts:

(i) (Embeddings + metadata) followed by PCA-99, yielding 22 components

(ii) SelectKbest on (tabular feature set including metadata), yielding exactly 50 features

(iii) (Embeddings-PCA95 + SelectKBest) combined dataset yielding a fusion of 50 features and 16 components

III. Models trained:

1. HNNs with custom PearsonCorr + CosineSim Loss with a tunable alpha-weight parameter. To mitigate overfitting, we used shared parameters in a multi-task multi-output model setting, with hyperparameters: Adam optimizer,mini-batch size of 32, 100 epochs, and learning rate 1e-3
2. Random Forest MultiOutput Regressor to predict all 52 target attributes

So we trained 2*3 = 6 models

IV. Ensemble predictions:

We used the following ensembles for the three test set submissions:

1. mean (SelectKbest RF, (SelectKbest + Embeddings-PCA95) RF, Embeddings-PCA99 HNN)
2. SelectKBest RF

3. mean (Embeddings–PCA99 RF, Selectkbest RF)

Task-2: Olfactory mixtures percept

I. Feature engineering:

For each molecule in the mixture we generated MAP4 fingerprints, MACCS keys, and Mordred descriptors, which were aggregated using a **dilution-weighted average** to get a single feature vector per stimulus. We also generated the ChEMBERTa embeddings for each molecule

II. Dataset consolidation

We used the following feature selection techniques:

(i) PCA-99 & PCA-95 on the Embeddings, yielding 54 and 29 components respectively

(ii) SelectKBest on the tabular feature set using F-regression scores to identify top 50 features.

From these results, we engineered the following datasets for predicting olfactory mixture percepts:

(i) (Embeddings–PCA-99), with 54 components

(ii) SelectKbest on (tabular feature set), yielding exactly 50 features

(iii) (Embeddings–PCA95 + SelectKBest) combined dataset yielding a fusion of 29 components and 50 features

III. Models trained:

1. HNN: with custom PearsonCorr + CosineSim Loss with a tunable alpha-weight parameter, and same hyperparameter values as in Task-1.
2. RF: Random Forest MultiOutput Regressor to predict all 52 target attributes in aggregate [RF]

Given three datasets, we trained 2*3 = 6 models totally.

IV. Ensemble predictions:

We used the following ensembles for the three test set submissions:

1. mean (SelectKbest RF, (SelectKbest + Embeddings-PCA95) RF, Embeddings-PCA99 HNN)
2. (SelectKbest + Embeddings-PCA95) RF
3. mean (Embeddings–PCA99 RF, (SelectKbest + Embeddings-PCA95) RF)

REFERENCES:

1. Capecchi, A., Probst, D. & Reymond, JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminform* 12, 43 (2020).
2. Moriwaki, H., Tian, YS., Kawashita, N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* 10, 4 (2018).
3. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chemical Information Computer Sci 42:1273–1280
4. Chithrananda S, Grand G, Ramsundar B (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv:2010.09885
5. Ganea OE, Bécigneul G, Hofmann T. (2018) Hyperbolic Neural Networks. arXiv:1805.09112
6. Yuansheng Zhou et al. Hyperbolic geometry of the olfactory space. Sci. Adv. 4, eaaq1458 (2018).DOI:10.1126/sciadv.aaq1458
7. DREAM Olfactory Mixtures Prediction Challenge 2025. https://www.synapse.org/Synapse:syn64743570/wiki/