Advancing Odor Mixture Discriminability with Pretrained Embeddings and Boosting

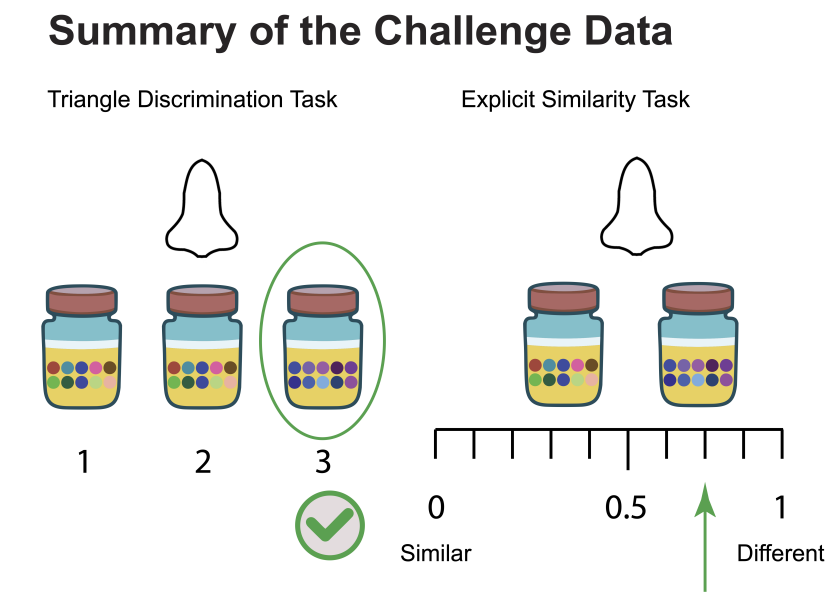
Yikun Han Zehua Wang Stephen Yang Ambuj Tewari

Department of Statistics, University of Michigan

October 1, 2024

Table of Contents

Dataset



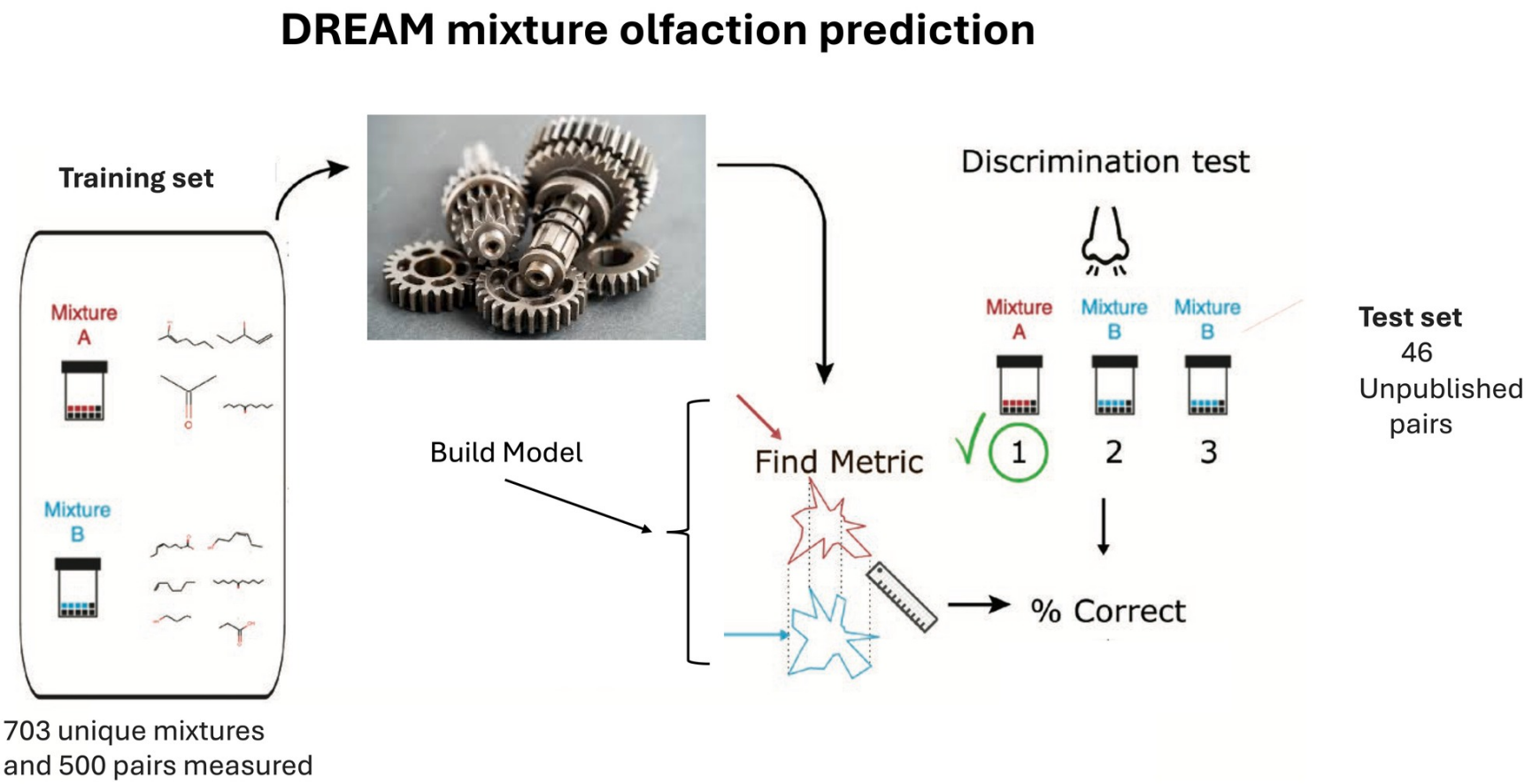
- In both tasks, the outcome measure varies from 0 to 1
- We consider Similarity score to be equivalent to the Discrimination score
- In the triangle discrimination task, a score of score of 0.7 means 70% of panelists could identify the odor that was different

- In the explicit similarity task, a score of score of 0.7 is the panel average rating of dissimilarity

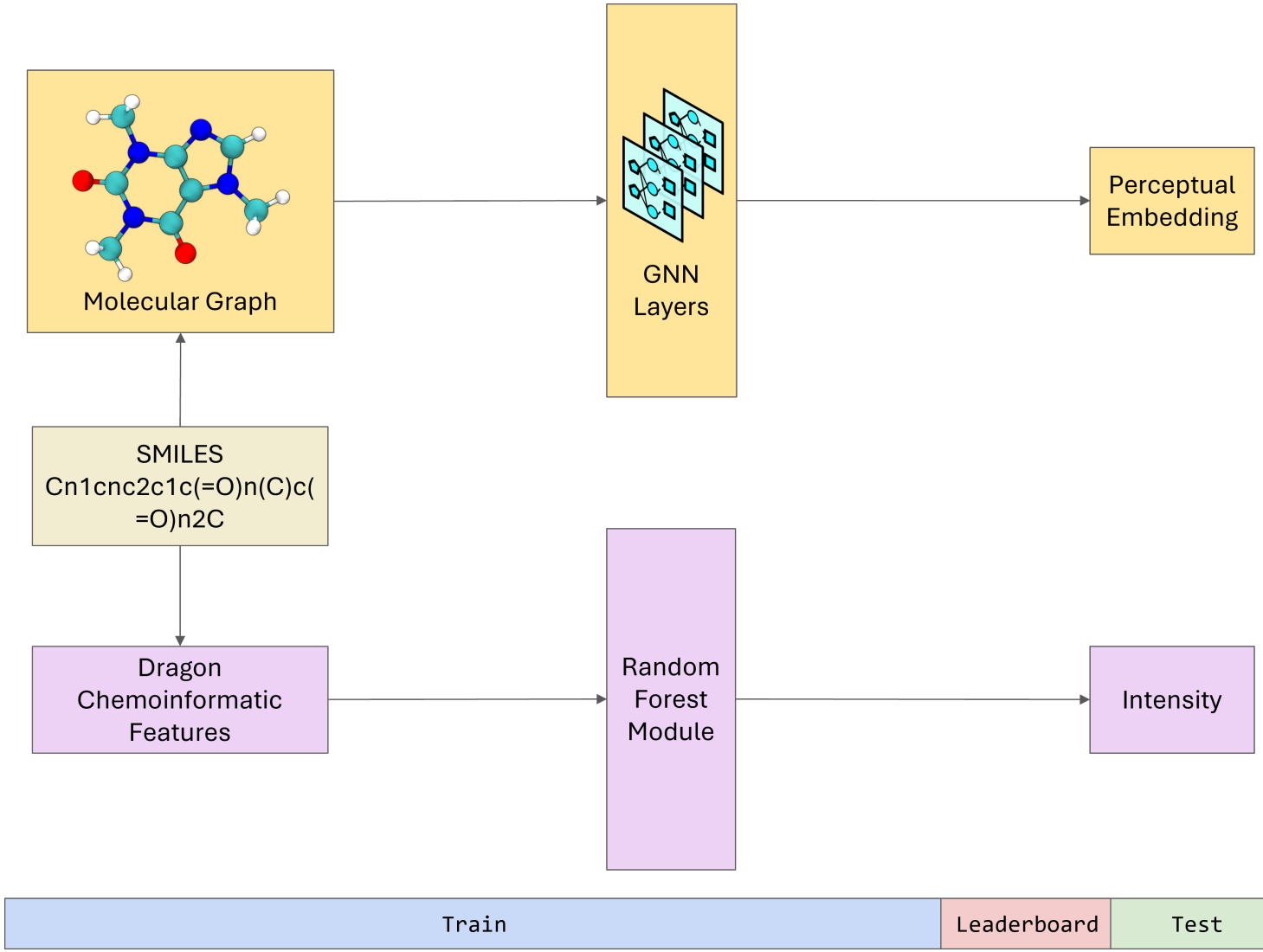
Dataset	Measurement
Snitz et al., 2013	Explicit similarity
Bushdid et al., 2014	Triangle discrimination
Ravia et al., 2020	Explicit similarity

[1] *DREAM Olfactory Mixtures Prediction Challenge* (n.d.). Synapse (syn53470621)

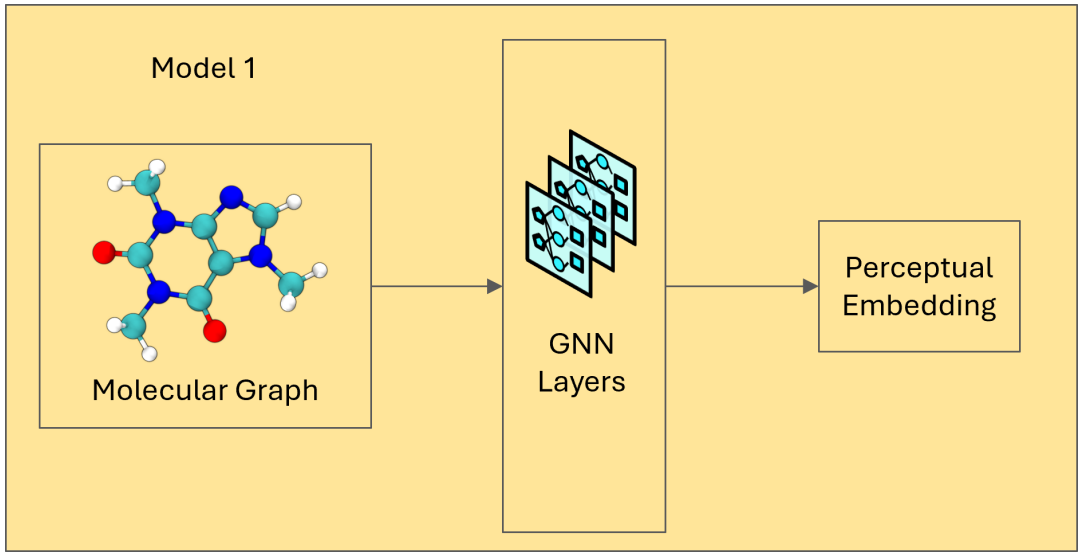
Task


DREAM mixture olfaction prediction
Training set
Mixture A
Mixture B
703 unique mixtures and 500 pairs measured
Build Model
Find Metric
Discrimination test
Mixture A
Mixture B
Mixture B
1
2
3
% Correct
Test set
46 Unpublished pairs

Embeddings for Single Molecules



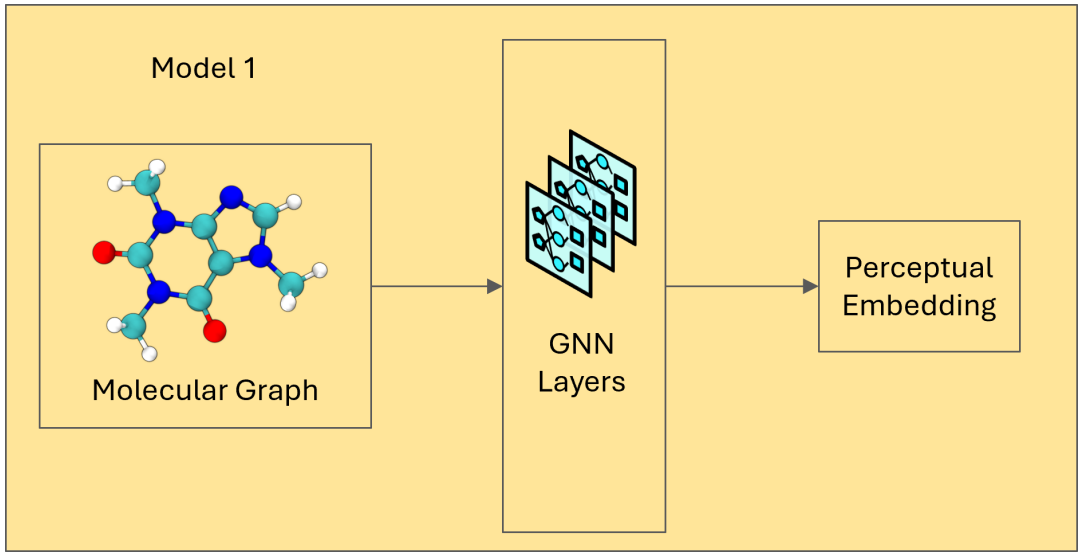
Embeddings for Single Molecules



[2] B. K. Lee et al. (2023). "A principal odor map unifies diverse tasks in olfactory perception". In: *Science*

[3] *openpom [GitHub repository]* (n.d.). BioMachineLearning

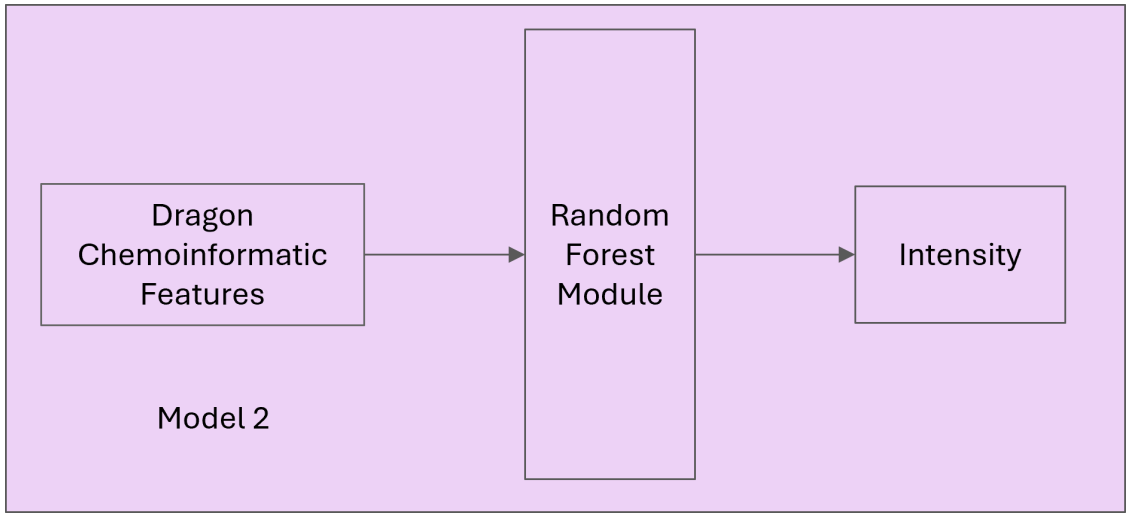
Embeddings for Single Molecules



GoodScents-Leffingwell Dataset

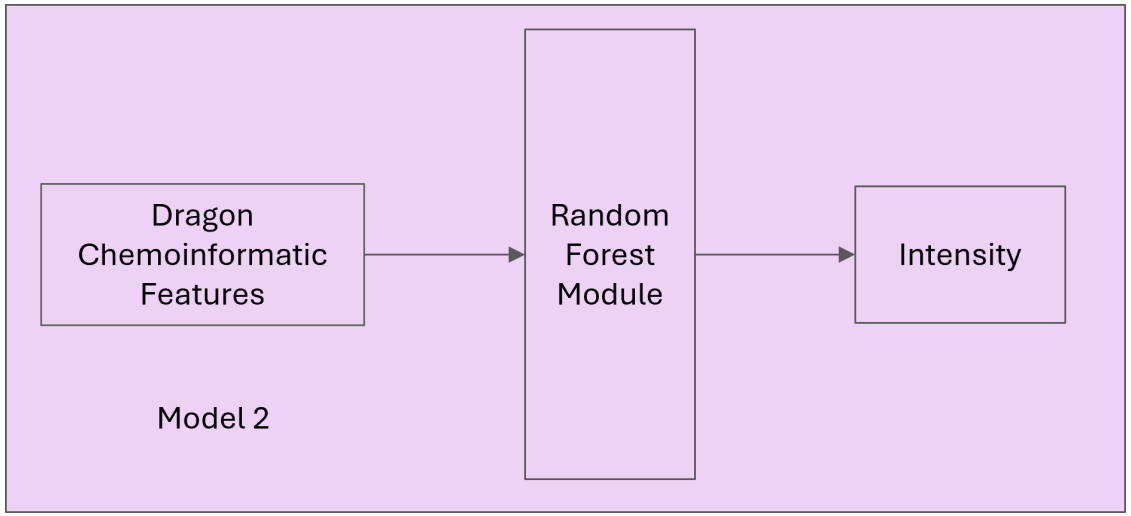
- 4983 samples
- 138 labels
- Multi-label classification

Embeddings for Single Molecules



[4] A. Keller et al. (2017). "Predicting human olfactory perception from chemical features of odor molecules". In: *Science*

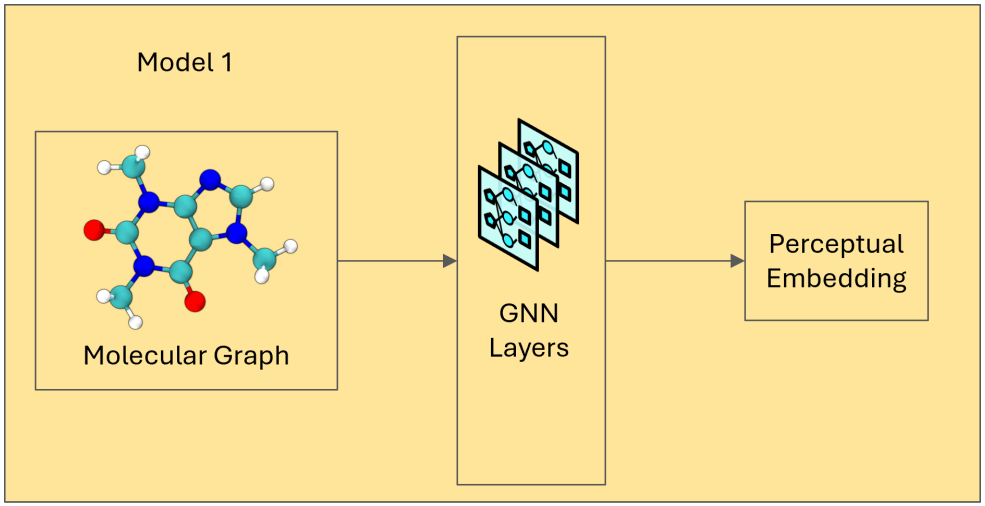
Embeddings for Single Molecules



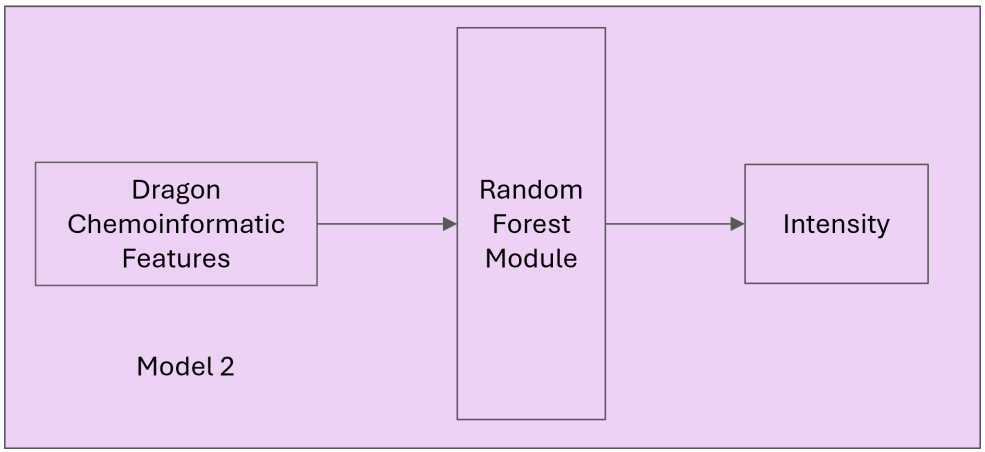
Keller-Vosshall Dataset

- 338 samples
- 19 odor descriptors, pleasantness, **intensity**
- Multi-label classification, regression

Embeddings for Single Molecules

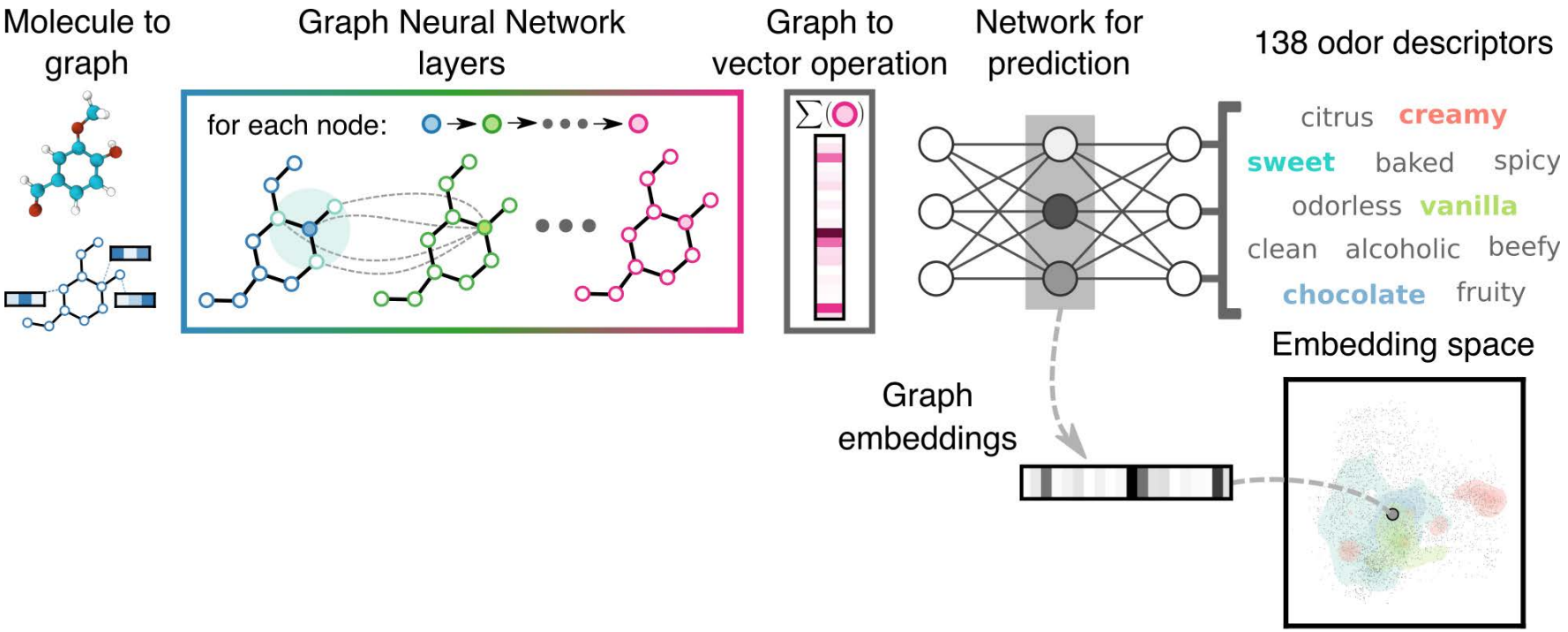


Model 1:
Weighted the molecules in the intersection of pre-trained data and competition data more heavily.



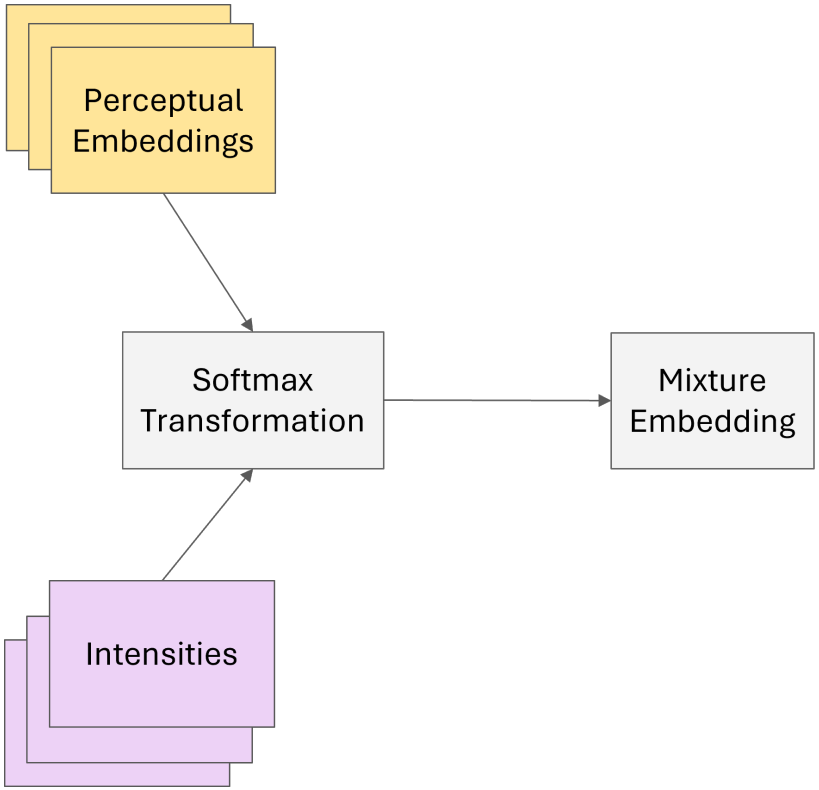
Model 2:
Merged the leaderboard and test set.

Graph Neural Network



[5] B. Sanchez-Lengeling et al. (2019). "Machine learning for scent: Learning generalizable perceptual representations of small molecules". In: *arXiv*. arXiv: 1910.10685

Embeddings of Molecules to Embedding of Mixtures



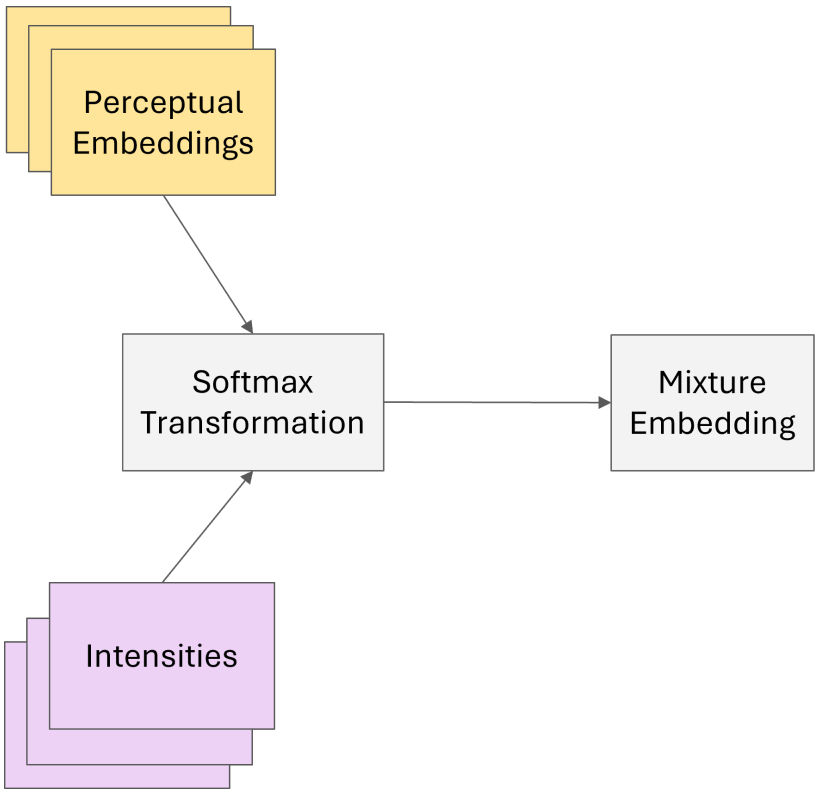
We apply a **softmax transformation** to normalize the intensities:

$$w_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

where w_i are the normalized weights. The final **mixture embedding** is the weighted sum of perceptual embeddings:

$$\mathbf{E}_{\text{mixture}} = \sum_i w_i \mathbf{E}_i$$

Embeddings of Molecules to Embedding of Mixtures

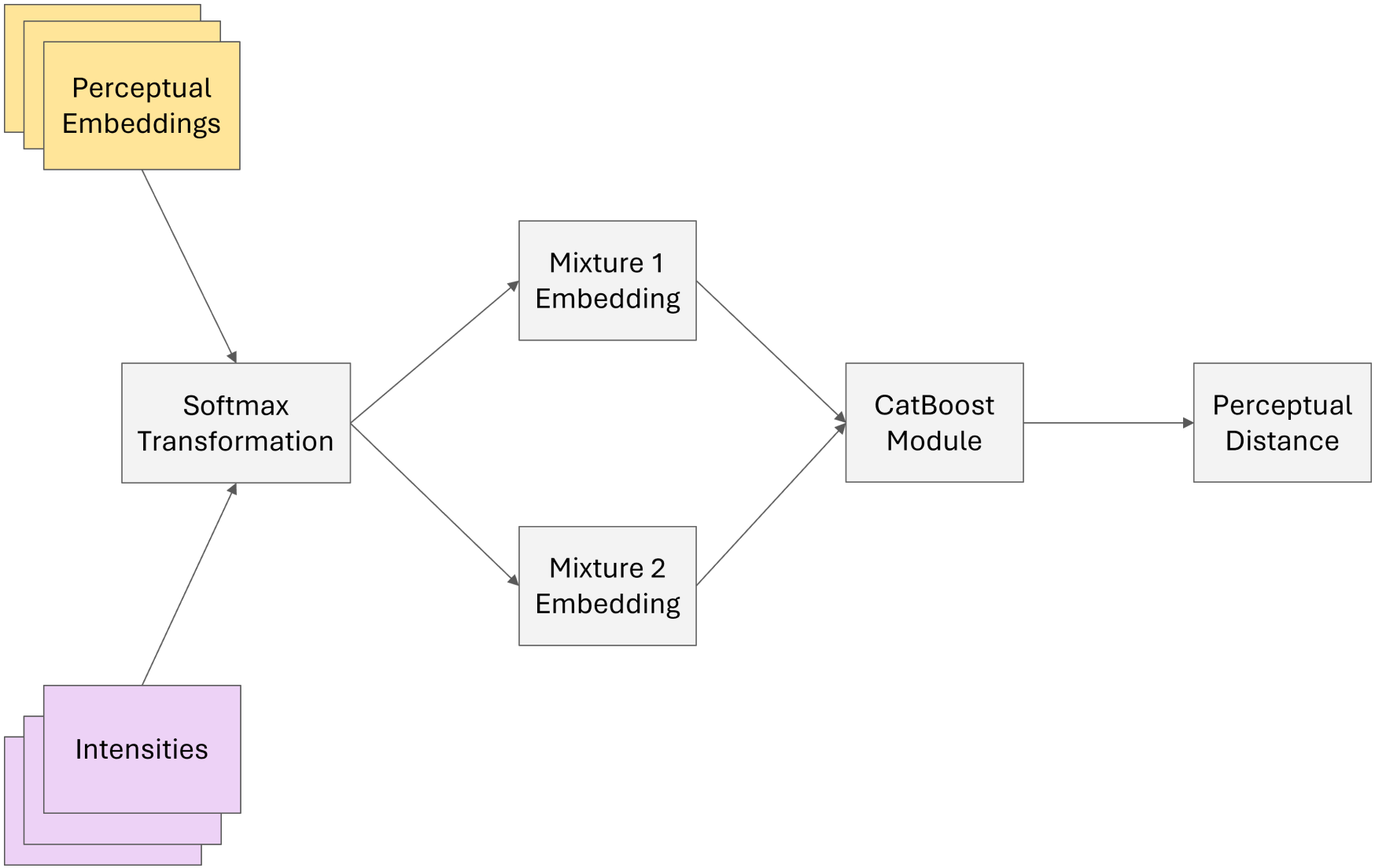


We introduce a **temperature parameter** τ to control the smoothness of the softmax:

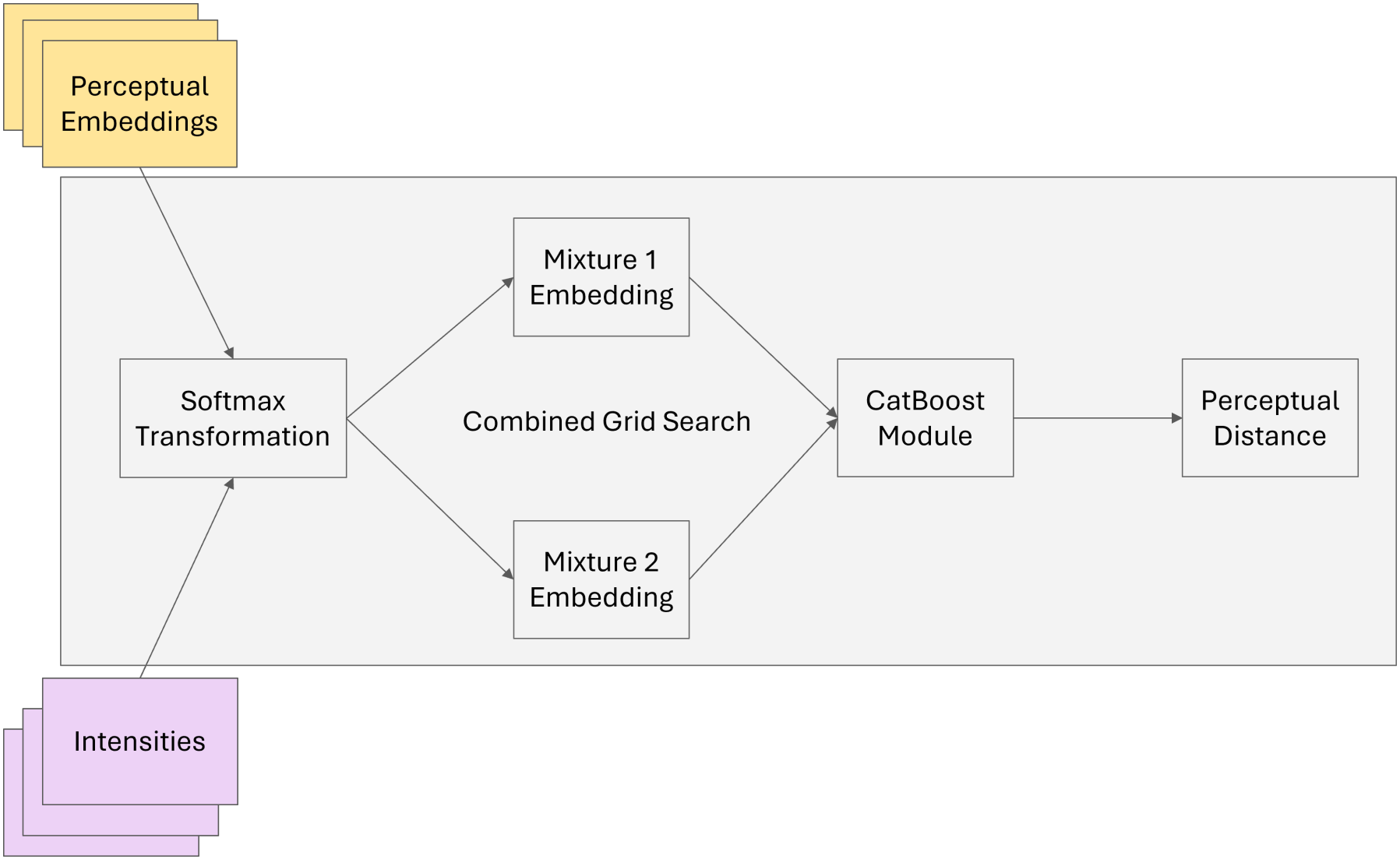
$$w_i = \frac{e^{x_i/\tau}}{\sum_j e^{x_j/\tau}}$$

A lower τ makes the weights sharper, focusing on larger values, while a higher τ distributes the weights more evenly. This helps adjust the influence of each molecule in the mixture.

Embedding to Distance



Embedding to Distance



CatBoost

Symmetric Trees

- Trees where the same splitting criterion is applied across all nodes at the same depth
- Efficient computation
- Fast predictions
- Regularization effect

Results

	LightGBM	XGBoost	RF	CatBoost
Valid Pearson	0.601	0.697	0.636	0.633
Valid RMSE	0.128	0.117	0.124	0.122
Leaderboard Pearson	0.549	0.502	0.596	0.625
Leaderboard RMSE	0.131	0.137	0.123	0.123
Test Pearson	-	-	-	0.501
Test RMSE	-	-	-	0.089

Key Insights from Small Data

- **Pretrained Models:**
  We relied heavily on pretrained models.
- **Data Augmentation:**
  Tried substituting a single molecule in a mixture with its nearest neighbor based on GNN embeddings, but this approach did not yield improvements.
- **Simple Models:**
  Small data led us to keep models and featurization straightforward.
- **Regularization:**
  CatBoost integrates strong **regularization** techniques, which help prevent overfitting more effectively than other boosting and bagging methods.

Looking to the Future

- **More Pretrained Models:**
  As larger datasets become available, we will explore more pretrained models, particularly focusing on olfactory receptor (OR) modeling.
- **Deep Learning and Attention:**
  Deep learning, especially attention mechanisms, will be leveraged to better capture the relationships between sets and create more sophisticated embeddings.
- **Advanced Data Augmentation:**
  With more data, advanced data augmentation techniques will be explored to further enhance model performance and generalizability.

Thanks

Thanks for listening!