# DOES THIS SMELL THE SAME? LEARNING REPRESENTATIONS OF OLFACTORY MIXTURES USING INDUCTIVE BIASES

**Gary Tom**[1,2*], **Cher Tian Ser**[1,2*], **Ella M. Rajaonson**[1,2*], **Stanley Lo**[1,2], **Hyun Suk Park**[1], **Brian K. Lee**[3], **Benjamin Sanchez-Lengeling**[1,2,†]
[1]University of Toronto
[2]Vector Institute for Artificial Intelligence
[3]Independent
{gtom, ctser, rajao}@cs.utoronto.ca
[†]ben.sanchez@utoronto.ca

## ABSTRACT

Olfaction—how molecules are perceived as odors to humans—remains poorly understood. Recently, the principal odor map (POM) was introduced to digitize the olfactory properties of single compounds. However, smells in real life are not pure single molecules, but complex mixtures of molecules, whose representations remain relatively under-explored due to limited data in olfactory mixtures. We introduce POMMIX, a mixture model extension of POM which leverages mono-molecular olfactory data to build meaningful mixture representations of smells. Our model builds upon the symmetries of the problem space in a hierarchical manner: (1) graph neural networks for building mono-molecular embeddings, (2) attention mechanisms for aggregating molecular representations into mixture representations, and (3) cosine prediction heads to encode olfactory perceptual distance in the mixture embedding space. POMMIX achieves state-of-the-art performance across multiple datasets. We evaluate the generalizability of the model, explore olfactory phenomena with the representations, and analyze the interpretability of the representations. Our work advances the effort to digitize olfaction, highlighting the synergy of domain expertise and deep learning in crafting mixture representations in low-data regimes.

## 1 INTRODUCTION

A central challenge in neuroscience is deciphering the link between the physical properties of a stimulus and its perceptual characteristics. While this relationship is well-defined for senses like vision (wavelength to color) and audition (frequency to pitch), it remains elusive for olfaction, a chemical sense, where the mapping from chemical structure to odor perception is complex and not fully understood (Sell, 2006; Barwich & Lloyd, 2022; Barwich, 2022). A recent advance towards digitizing olfaction came with the introduction of the principal odor map (POM) by Lee et al. (2023), a high-dimensional, data-driven representation of odor perceptual space learned from molecular structures. This model demonstrated human-level performance in predicting odor qualities of single molecules and generalized well to other olfactory tasks. However, naturally occurring olfactory stimuli are not comprised of single molecules, but rather complex mixtures of molecules, whose representations remain relatively unexplored within the existing literature. This work introduces POMMIX—a mixture and distance-aware extension of the POM representation.

A searchable, rankable, and optimizable digital representation of olfactory space has potential applications in diverse areas (Spence et al., 2017). Such a representation could be used to develop mosquito repellents (Wei et al., 2024), inform agricultural practices by enabling targeted manipulation of insect behavior (Conchou et al., 2019), improve food quality and reduce waste through enhanced spoilage detection (Jung et al., 2023), and accelerate the design of novel fragrance and

*Equal contribution; random order.

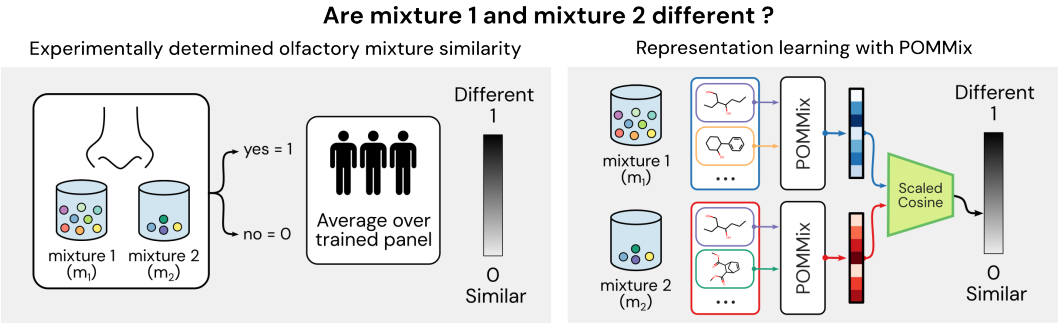


Figure 1: **Task schematic**. Experimental evaluation of olfactory mixture similarities (*top*), and our approach to predicting olfactory mixture similarities (*bottom*).

flavor compounds, which is particularly valuable given increasing regulatory constraints on existing ingredients (Demyttenaere, 2012; IFRA, 2024).

Deep learning models enable the construction of task-optimized data representations, learning complex relationships directly from data (Bengio et al., 2012). However, in low-data regimes, the success of deep learning hinges on incorporating appropriate inductive biases, effectively injecting domain-specific knowledge to guide the learning process and improve generalization (Tom et al., 2023). Olfactory data is currently in this regime—gathering olfactory data is expensive and labor-intensive as it requires training human panelists, filtering potentially toxic molecules, and navigating ethical review boards. Existing work on olfactory mixture modeling remains limited, employing a small pool of compounds with low coverage of chemical space (Ravia et al., 2020; Snitz et al., 2019; Sisson et al., 2023; Snitz et al., 2013). While the perfume industry reportedly utilizes 10,000–20,000 compounds routinely, the largest publicly available dataset GoodScents–Leffingwell (GS-LF) contains only around 5,000 molecules (Sanchez-Lengeling et al., 2019; Lee et al., 2023). Significantly larger repositories of mixture characterizations (blends and perfumes) exist within the industry, but remain inaccessible behind private doors.

POMMIX is built by training a neural network to tackle the mixture similarity problem by jointly training a POM with an attention-based mixture model to predict the perceptual similarity of mixtures (Figure 1). This approach also allows us to combine mono-molecular datasets (up to 5,000 data points) and more limited mixture data sources (up to 1,000 data points). At each stage of our work, we take care to respect the natural symmetries of the problem space—namely, the permutation invariance of molecular descriptions (introduced by the graph representation of a molecule), the permutation invariance of mixture compositions (the model should not care in what order the mixture ingredients are presented), and the symmetry of mixture similarities (the model should predict that the similarity of mixture 1 and mixture 2 is the same as the similarity for mixture 2 and mixture 1). The end result is an extension of the POM to mixtures, and a new model building block, dubbed CHEMIX, for encoding mixtures of molecules.

We summarize our contributions below:

- We introduce POMMIX, the first extension of the POM to predict the olfactory similarities between mixtures of molecules.
- We compile and comprehensively analyze the limited publicly available olfactory mixture perception data, and provide a useful benchmark dataset for future study.
- Our model takes into account the inductive biases of the problem and achieves state-of-the-art predictions on olfactory mixture similarity. We find that POMMix outperforms pre-trained molecular transformer-based representations.
- We significantly broaden the scope and analysis of mixture modelling efforts compared to previous work Zhang et al. (2023).
- We test our representation in several olfactory settings: the olfactory white-noise hypothesis, generalization to unseen molecules and mixture sizes, and a qualitative study of the interpretability of components within mixtures.
- Our data and code are available for reproducibility and future extensions: https://anonymous.4open.science/r/anon_pommix-E213.

### 1.1 Related Works

The modeling of molecular structure-property relationships has a rich history. Within the olfactory domain, previous contributions have utilized hand-picked expert descriptors with classical machine learning algorithms (e.g. tree-based models, support vector machine, and linear models), and/or similarity measures (e.g., cosine, angle) (Snitz et al., 2013; Keller et al., 2017; Kowalewski & Ray, 2020; Vigneau et al., 2018). More recently, deep learning based models have been actively explored to create a more expressive molecular representation of olfactory space (Lee et al., 2023; Tran et al., 2018; Zhang et al., 2024; Sisson, 2022; Maziarka et al., 2020), notably by integrating olfactory receptor information using representations from protein language models (Chithrananda et al., 2024).

Deep learning techniques have been explored in modeling molecular structure-property relationships, including variational autoencoders (Gómez-Bombarelli et al., 2018; Oliveira et al., 2022), large language models (Chithrananda et al., 2020; Ross et al., 2022), and graph neural networks (GNNs) (Yang et al., 2019; Wang et al., 2021). Graph neural networks and graph attention networks (GATs) (Heid et al., 2024; Wu et al., 2023; Buterez et al., 2024) in particular have shown state-of-the-art performance in many molecular property prediction tasks including modeling olfactory space.

Olfactory mixture property prediction is a much more difficult task with fewer effective attempts (Lapid et al., 2008; Khan et al., 2007; Olsson, 1998; Dhurandhar et al., 2023; Ravia et al., 2020). Molecular mixtures have been studied before for battery electrolytes by Zhang et al. (2023). The work, however, uses a large dataset (10,000 mixtures), and focuses on property prediction, rather than mixture representation learning. To work in the low-data regime of our olfactory mixture dataset, POMMix uses pre-training techniques (Honda et al., 2019; Shoghi et al., 2023; Goh et al., 2018) and designed inductive biases to improve the expressivity of the molecular representation and attention mechanisms (Wang et al., 2019; Xiong et al., 2020; Maziarka et al., 2020).

## 2 Methods

### 2.1 Data

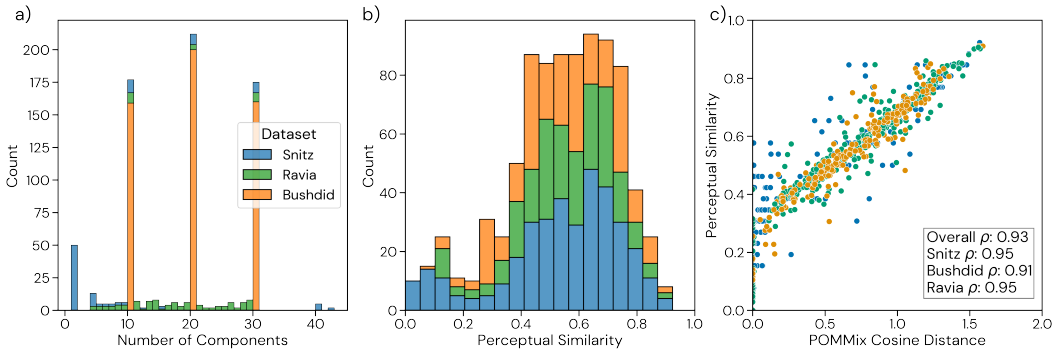


Figure 2: **Snitz, Ravia, and Bushdid mixture datasets at a glance**. **a**) Most mixtures contain 4-30 molecules, with a handful of mono-molecule mixtures from **Snitz**. **b**) Histogram of perceptual scores; most mixtures are somewhat different, scored 0.4-0.8. **c**) POMMix embeddings are carefully tuned for the task of discriminating mixture percepts. Pearson $\rho$ correlation constants are annotated in inset. Across all three subplots, color labels indicate the dataset source.

We combine mono-molecular datasets and multi-molecular (mixture) datasets. Mono-molecular datasets list a set of odor labels ("grassy", "fishy", etc.) for a single molecule, and the most exhaustive compilation is found in the GoodScents/Leffingwell (GS-LF) dataset (Barsainyan et al., 2024). We further clean the GS-LF dataset by canonicalizing SMILES (Weininger, 1988) strings with RDKit (Landrum et al., 2022), removing duplicate entries, inorganic, charged or multi-molecular (e.g. salts) entries, molecules with molecular weight < 20 and > 600, and small inorganic molecules. We further removed infrequently applied odor labels that appeared for fewer than 20 molecules and

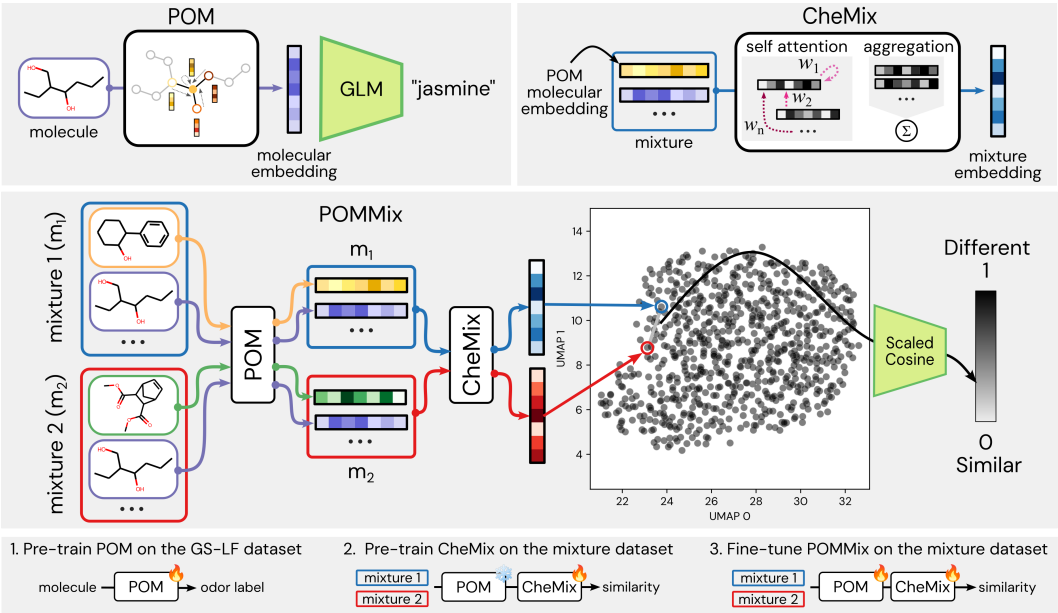


Figure 3: **POMMIX model combines POM with mixture modeling**. (*Top*) The POM model with a generalized linear model (GLM) is pre-trained with mono-molecular olfactory data, and mixture modeling is performed through the CHEMIX attention model. (*Middle*) The two modules are joined to produce mixture embeddings which are trained to encode the olfactory perceptual distance of two mixtures using a scaled cosine distance predictor head. (*Bottom*) A multi-step model fitting procedure is used, where certain model weights are updated (flame) and other pre-trained model weights are frozen (snowflake).

subsequently removed molecules with no remaining labels (see Appendix A.1 for details on dataset cleaning).

Multi-molecular datasets were compiled from previous publications, hereby referred to as **Snitz** (Snitz et al., 2013) (containing data from Weiss et al. (2012)), **Ravia** (Ravia et al., 2020), and **Bushdid** (Bushdid et al., 2014). Data for each of these publications was obtained from *pyrfume* (Castro et al., 2022). In aggregate, we have 743 unique mixtures, containing between 1 to 43 unique molecular components (Figure 2a).

These mixtures are described by 865 pairwise mixture comparisons (Figure 2b) corresponding to labels from two types of experiments:

- *Explicit similarity* (**Snitz, Ravia**): participants are asked to explicitly rate the perceptual similarity of two mixtures from 0 (completely similar) to 1 (completely different), and the final similarity for a mixture pair is averaged across all participants;
- *Triangle discrimination* (**Bushdid**): participants are provided three mixtures, of which two are identical, and asked to identify which mixture was different. These results are aggregated for each mixture triplet, and the percentage of correct identifications is treated as the label for the two unique mixtures in the triplet.

We note that the interpretation of the triangle discrimination task is congruent with the explicit similarity task, as a score of "1.0" in a triangle discrimination task shows that all tested participants could identify the mixture that was different, which meant that the two unique mixtures in the triplet were perceptually distinct. However, calibration of intermediate scores may differ; we did not attempt to correct for this effect. Additionally, **Snitz** and **Bushdid** are intensity balanced, meaning that their components are first diluted to equal odor intensities using an odorless solvent and then proportionally mixed. POMMIX does not explicitly account for intensity or concentration of odorant mixtures, and would likely underperform in predicting mixture similarity if presented with mixtures that are not intensity-balanced.

### 2.2 MODELING

A schematic of the POMMIX model is provided in Figure 3. The POMMIX model can be divided into three hierarchical components: (1) a mono-molecular GNN POM embedding model, (2) a multi-molecular CHEMIX mixture attention model, and (3) a similarity scoring function.

The POM is a GNN which takes in molecular graphs derived from the SMILES representations of molecules. Each graph, written as $G = (U, V, E)$, consists of a special global vertex $U$ connected to all other vertices $V$, and a set of edges $E$. The global vertex $U$ encodes overall properties of the molecule and is initialized with 200 normalized RDKIT cheminformatics molecular descriptors (Landrum et al., 2022) obtained from DESCRIPTASTORUS (Kelley et al., 2024). The atoms of the molecules are the vertices (nodes), with node vectors $V = \{v_i\}_{i=1}^{N_v}$ for a molecule with $N_v$ atoms, where $v_i$ are feature vectors encoding atomic properties. Covalent bonds between atoms are represented as edges $E = \{(e_k, r_k, s_k)\}_{k=1}^{N_e}$ for a molecule with $N_e$ bonds, where $e_k$ are feature vectors of edge properties, and $r_k, s_k \in [1, \ldots, N_v]$ are indices of the two atoms that the bond joins together. Note $r_k \neq s_k$, since bonds must be between two different atoms (see Appendix A.2 for details of node and edge properties).

The POM GNN is based on the GRAPHNETS architecture (Battaglia et al., 2018), with message-passing blocks for the edge, node and global properties of the molecular graphs. The architecture is designed to be lightweight in order to avoid overfitting on the limited amounts of olfactory mixture data. Edge updates use feature-wise linear modulation (FiLM) layers (Perez et al., 2017; Brockschmidt, 2019), while node updates use graph attention layers (Veličković et al., 2017; Brody et al., 2021) with self-attention. The global embeddings are updated through principal neighborhood aggregation (PNA) (Corso et al., 2020; Zaheer et al., 2017). The GNN is composed of four of these GRAPHNET layers, and the final global embedding serves as the POM embedding.

The CHEMIX model processes a set of molecular POM embeddings, and generates an embedding representing the entire mixture. Mixtures are first represented by concatenating POM embeddings of constituent molecules, and mixtures with fewer molecules are padded to the length of the largest mixture. CHEMIX uses molecule-wise self-attention, where each molecule attends to all other molecules, followed by PNA. This ensures invariance of the mixture embeddings in the permutation of molecules within a given mixture. This model can be viewed as isomorphic to a GAT on a fully-connected graph (Joshi, 2020), with each molecule as a node, and the mixture embedding is the global embedding.

Finally, the distances between the mixture embeddings are obtained through a similarity score. For this, we use a predictive head based on cosine distance (Koch et al., 2015), commonly used for distance-aware high-dimensional learned representations. A final two-parameter linear layer is used to encode for human bias and experimental noise present in the dataset (see section 3), followed by a HardTanh activation to enforce output in the [0,1] range, while maintaining linearity. We note that this scaled cosine prediction head is invariant to the order of the mixtures due to the symmetry of the cosine distance operation.

### 2.3 TRAINING AND OPTIMIZATION

In order to effectively train a deep learning model in a low-data regime, we adopt a transfer learning strategy (Figure 3). The POM GNN is first pre-trained to predict the olfactory binary multi-labels of molecules with binary cross-entropy loss on the GS-LF dataset, using a 80/20 training/validation random split. All training is performed using the Adam optimizer (Kingma & Ba, 2014). To determine the architecture, we perform a Bayesian optimization hyperparameter search to maximize the area under receiver operator curve (AUROC) metric. Early stopping is used to prevent overfitting. The best model achieves a validation AUROC 0.884, in line with previous work (Sanchez-Lengeling et al., 2019). We explore other state-of-the-art models such as graph transformers GRAPHORMER (Shi et al., 2022; Ying et al., 2021) and GPS (Rampášek et al., 2022), but we find our POM architecture remains optimal for the GS-LF olfactory dataset (Appendix Table 3).

The frozen POM embeddings from the pre-trained GNN form the vector representation of mixtures for the CHEMIX model. Again, the architecture is determined through hyperparameter tuning. The training is performed with mean absolute error (MAE) loss on a 80/20 training/validation split of the combined mixture dataset, stratified across **Snitz**, **Ravia**, and **Bushdid**. The stratification process

fixes the proportion of each dataset across the splits, ensuring equal representation of any experimental differences. Additionally, we ablate the CHEMIX prediction head, and find that the scaled cosine prediction head is optimal for learning mixture embedding similarities (Appendix Table 4). Early stopping terminates on maximal validation Pearson correlation coefficient ($\rho$) between the ground truth labels and the prediction. The optimal model found in the search achieved a maximal $\rho = 0.794$ on the validation set. Further details about the hyperparameter tuning for both models are provided in Appendix A.3.

In the final stage of training POMMIX, the POM GNN is directly joined to the CHEMIX model, and all model weights fine-tuned. A lower learning rate is used for the POM GNN model weights, as they are already well-conditioned from pre-training on the larger mono-molecular odor dataset. The results following this section are based on the final POMMIX model. Models were built with PYTORCH (Paszke et al., 2019) and PYTORCH GEOMETRIC (Fey & Lenssen, 2019).

## 3 RESULTS

We evaluate our approach on the mixture dataset by training and testing on 5-fold cross-validation (CV) splits, stratified across the **Snitz**, **Ravia**, and **Bushdid** datasets. For early stopping, a validation split is randomly split from the training set, producing a final split of 70/10/20 training/validation/test sets. The performances of the models are then evaluated on the test sets.

We evaluate POMMIX on a progressive ladder of modeling components. For the simplest baseline, we follow the methods of Snitz et al. (2013), who performed extensive feature selection on molecular descriptors, which are then averaged together for the mixtures (see Appendix A.4). The angle distance between the vector descriptors are then correlated with the experimental results. We perform the same analysis using normalized RDKIT molecular features on our aggregated mixture dataset. We ensure that the feature selection is only performed on the training set.

We also provide comparisons with the gradient-boosted random forest XGBOOST model (Chen & Guestrin, 2016) (details in section A.5), and use features with varying levels of inductive biases. In addition to RDKIT features, and POM embeddings, we also use molecule embeddings from the pre-trained chemical language transformer model MOLT5 (Edwards et al., 2022), which was trained with self-supervised learning on unlabeled natural language and molecular strings and subsequently fine-tuned on annotated chemical data. Mixture representations are created with PNA-style aggregation of molecular descriptors, including RDKIT features, MOLT5 embeddings or the frozen POM embeddings. Additionally, we augment the training data by permuting the mixture pairs, as the symmetry of the mixture similarity is not encoded in XGBOOST.

### 3.1 PREDICTIVE PERFORMANCE

We report results across three metrics: Pearson correlation coefficient $\rho$, root-mean-squared error (RMSE), and the Kendall ranking coefficient $\tau$, each reflecting different strengths of the model. The test results compiled from the CV splits for all models evaluated are tabulated in Table 1. We show that incorporating more inductive biases into the model leads to a dramatic increase in model performance. The Snitz baseline of angle distances calculated from empirically selected features produces a weak positive correlation with the ground truth distances, but has high RMSE when treated as a regression problem. The XGBOOST model improves upon these predictions, and explicitly models the experimental distances, achieving significantly lower RMSE than the Snitz baseline. However, the correlation and ranking of the mixture similarity is only slightly increased through the use of the boosted RF model.

When applying the XGBOOST model to the POM embeddings, we find that the performance is slightly improved over the RDKIT descriptors, signaling that deep learning architectures are needed to further extract useful information out of the POM embeddings. Interestingly, we find that the pretrained MOLT5 embeddings are outperformed by both the RDKIT and POM features. Across all models, we find the best performance with the POM embeddings. Recent work by Shin et al. (2018) studying the use of transformer-based language models and in combination with graph models show that GNN methods are still optimal for this modeling problem.

For our approaches, CHEMIX shows excellent test performance for predicting olfactory mixture similarities, even when trained with frozen POM embeddings, demonstrating the efficacy of incorpo-

Table 1: **Model performances on mixture data with ablation of features and model components**. 5-fold cross validation metrics for all baseline models, CHEMIX and POMMIX. The mean and standard deviation are reported.

| Model | Pearson $\rho$ (↑) | RMSE (↓) | Kendall $\tau$ (↑) |
|---|---|---|---|
| Snitz Baseline | 0.399 ± 0.050 | 0.334 ± 0.010 | 0.292 ± 0.042 |
| XGBOOST + MOLT5 | 0.432 ± 0.030 | 0.171 ± 0.012 | 0.347 ± 0.036 |
| XGBOOST + RDKIT | 0.485 ± 0.048 | 0.166 ± 0.012 | 0.373 ± 0.040 |
| XGBOOST + POM | 0.497 ± 0.041 | 0.165 ± 0.012 | 0.388 ± 0.033 |
| CHEMIX + MOLT5 | 0.672 ± 0.021 | 0.144 ± 0.006 | 0.498 ± 0.023 |
| CHEMIX + RDKIT | 0.732 ± 0.030 | 0.132 ± 0.007 | 0.552 ± 0.040 |
| CHEMIX + POM | 0.746 ± 0.030 | 0.130 ± 0.007 | 0.545 ± 0.032 |
| POMMIX | **0.779 ± 0.028** | **0.118 ± 0.004** | **0.596 ± 0.022** |

rating domain knowledge and inductive biases into model architectures. For POMMIX, we observe further increases in model performance when the POM and CHEMIX are trained end-to-end, as the POM embeddings are fine-tuned for use in mixture representations. We note that the end-to-end training results in larger improvements in Kendall $\tau$ than in $\rho$. We hypothesize that inherent human noise in the experimental results create a performance ceiling for the model's real-valued predictive capabilities. However, the ranking correlation still improves as it is more robust to experimental noise and outliers (Tom et al., 2024). We further explore this human bias in Section 3.3.

### 3.2 GENERALIZATION TO NEW MIXTURE SIZES AND MOLECULES

We further study the effects of the inductive biases of the model, and the capabilities of POMMIX in explaining physical olfactory phenomena. In particular, we study the generalization of POMMIX to different splits based on the number of mixture components, and the molecular identities within the mixtures.

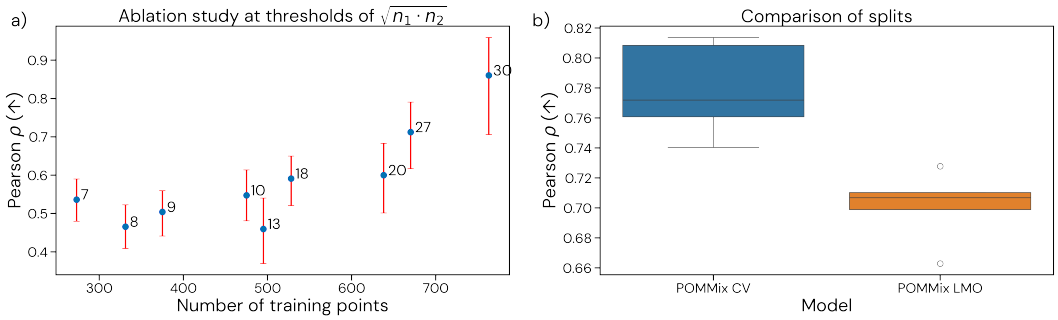


Figure 4: **Generalization to new mixture sizes and molecules**. **a)** Ablation study with training data only containing mixtures with geometric average number of molecules less than a threshold. The thresholds are indicated for each split. **b)** Boxplot of POMMIX test Pearson correlation on random CV splits, and the LMO splits.

In Figure 4a, we show the test results of an ablation study, in which the training data is ablated based on thresholds on the geometric average number of components found in a mixture. In other words, for a given threshold, the training set only contains mixtures with components that have a geometric mean number of components less than the threshold, and the test set contains only mixtures that are above the threshold. We observe sufficient generalization capabilities of the model to larger mixtures, achieving performances similar to the RDKIT baselines, even when the training sets are thresholded at ten mixture components, and only about two-thirds of the available training data. We also observe a general increase in test performance, measured by $\rho$, as the training set grows, indicating that more high quality experimental olfactory mixture data can greatly improve the modeling performance.

We observe a significant decrease in performance when considering new chemistries. For Figure 4b, we study POMMIX performance on leave-molecules-out (LMO) splits, in which the test sets are split from the dataset such that certain molecules will not appear in the training set. Note that, unlike the random CV splits, the training sets are not mutually exclusive, since there is significant overlap in the molecular identities across different mixtures. This additional challenge in studying new molecules is an important consideration when validating models, and also planning future mixture similarity experiments. More olfactory mixture data with diverse arrays of molecules can help build better and more generalizable POMMIX representations.

### 3.3 EXPLORING OLFACTORY PHENOMENA WITH POMMIX EMBEDDINGS

The *white noise hypothesis* states that intensity-balanced mixtures with an increasingly large number of components become increasingly indistinguishable, even if they share no common molecular components, and approach a scent characterized as an "olfactory white" (Weiss et al., 2012). Using the POMMIX embedding, we reproduce the "olfactory white" phenomena (Figure 5a). In our investigation, we observe this decrease in POMMIX embedding distances as a function of the geometric mean of components in mixture pairs for our larger dataset, which includes **Bushdid** and **Ravia**. This demonstrates the ability of POMMIX in capturing and explaining physiological olfaction phenomena, allowing it to build toward an expressive odor perceptual space.

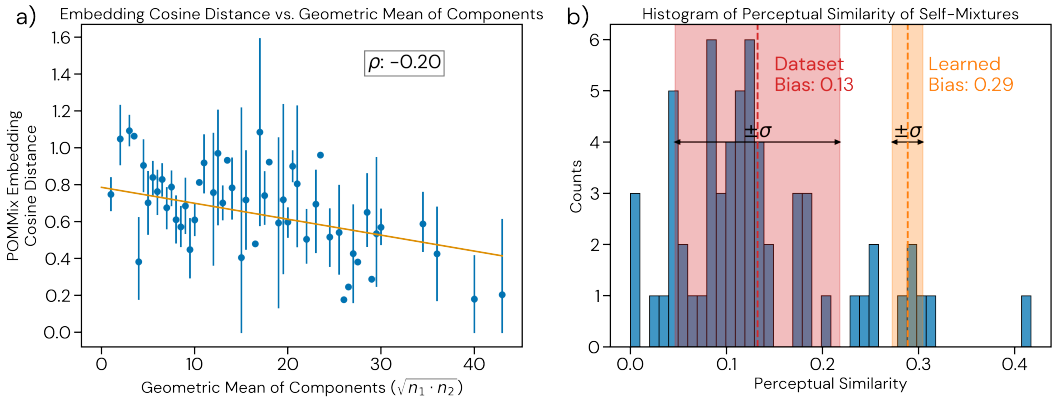


Figure 5: **Investigating inductive biases in perceptual olfactory data**. **a**) The white noise hypothesis (Weiss et al., 2012), where larger mixtures become less distinguishable from one another. **b**) Inherent human inaccuracies in the identification of two identical mixtures (data from **Snitz**). The mean (dashed line) and standard deviation (shaded region) of perceptual similarity for identical mixtures within the dataset is in red, while the learned bias of the similarity prediction head is in orange. The learned biases are averaged across the five CV splits.

It is important to note that the perceptual similarity metrics obtained across the datasets are inherently subjective and biased as they are gathered from humans. We show a subset of the data where the panelists are asked to rate the similarities of two *identical* mixtures, and show that a significant portion of identical mixture pairs (58 out of 61) are labeled as having non-zero similarities (Figure 5b). While the observed bias could be descriptive of average human olfactory inaccuracies, because the number of panelists sampled was low ($\sim$300), the bias could be local to the panelists. This human bias is modeled by the learned bias term of the scaled cosine similarity prediction head. In general, we observe that the learned bias is slightly higher than the dataset bias. When we physically ground the model by enforcing that the learned similarity of identical mixtures should be zero (i.e., $b = 0$), we observe poorer model performance (see Appendix A.6). This could be a result of having a small dataset (< 1,000 points); enforcing this inductive bias may be more relevant when considering larger data regimes and experiments with more human panelists, where the perceptual distance of identical mixtures approach zero, making the model more generalizable and not subject to the bias of a specific dataset.

### 3.4 BUILDING INTERPRETATIONS OF MIXTURES

An unanswered question relevant to mixture modeling is how the mixture components interact with each other and contribute to the prediction of mixture similarities. To probe at this question and generate hypotheses for future investigation, we modified CHEMIX to be more interpretable as an additive model (Agarwal et al., 2021). Specifically we express the self-attention-based mixing component as a one-layer additive model by using sigmoid normalization (Ramapuram et al., 2024) rather than softmax, allowing the model to attend to all components. In a simplified way, this is a pairwise interaction model. Although this modified model is simpler and more constrained, it achieves performance comparable to that of our best model.

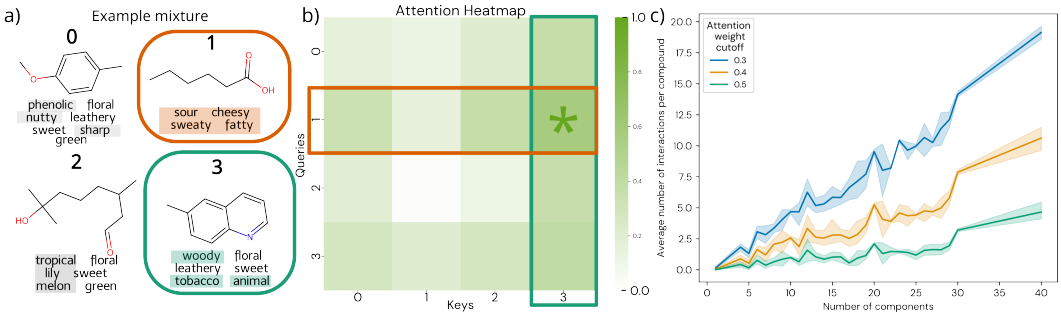


Figure 6: **Mixture attention maps**. **a)** Example mixture with molecules and their odor labels. **b)** Sigmoidal self-attention heatmap, with compound 1 and 3 highlighted. Strongest interaction is indicated with an asterisk. **c)** Number of average interactions per compound as function of mixture size across all datasets. Each color represents a different threshold for a meaningful interaction.

In Figure 6, we showcase how such sigmoidal self-attention maps can be used to analyze the information passing between representations of molecules within a mixture. More complex examples can be found in Appendix A.9. In this simple example, when comparing the GS-LF labels associated to each molecule (Figure 6a) to the attention weights attributed to each query (Figure 6b), we notice that the strongest "interaction"—namely, the highest attributed attention weight—is found between query molecule 1 and key molecule 3. In general, we observe that molecules that are most different from the rest, either by chemical structure (e.g., presence of N or S atom) or by olfactory perception (e.g., presence of rare or numerous labels), tend to have stronger interactions. To further our analysis, we derive label-guided structural heuristics about molecules across the set of unique mixtures in Appendix A.10. We find that higher attention is attributed to chemical structures with strong, pungent, and unique smells. These include compounds with sulfur, nitrogen, and aromatic structures. However, it is important to keep in mind that the attention map showcased here is intrinsically linked to the task of differentiating between mixtures and is therefore likely biased towards attributing higher attention weights to molecular embeddings that carry discriminative power only relative to this task. Careful experimentation with synthetic mixture tasks and dataset, where the number of data points is not as limited, might provide guidance on the strengths and failure modes of these approaches to interpretability (Sanchez-Lengeling et al., 2020). Prospective validation from new experimentation would also strengthen these hypotheses.

A multi-headed ($k$) softmax attention mechanism can be interpreted as attending to $k$ tokens (Joshi, 2020; Sanchez-Lengeling et al., 2021). In understanding interpretability, a natural question for mixtures might be: how many compounds do we need to attend to on average? Figure 6c attempts to answer this by looking at the average number of interactions per compounds across the dataset. We consider three attention weight cutoffs (0.3, 0.4 and 0.5) to define a significant interaction and observe that the number of interactions grows approximately linearly with the number of components in the mixtures. For the 0.5 cutoff, this is approximately two interactions per compound for mixtures of less than 30 components. We observe an increase in average number of interactions after 30 components; however, we also note that data is quite sparse here, precluding the formation of firm conclusions about the data (Figure 2a).

## 4 Conclusion and Discussion

We introduce POMMix, an extension of the POM for predicting olfactory similarities between mixtures of molecules. Our approach combines graph neural networks for molecular representation, with attention mechanisms for mixture modeling, and incorporates inductive biases by considering cosine similarities between mixture embeddings to predict olfactory similarity. POMMix demonstrates state-of-the-art performance, creating meaningful representations of olfactory mixtures, and we show how each component of inductive bias contributes to this performance. Our work highlights the value of incorporating domain knowledge and inductive biases, particularly in low-data regimes. By respecting problem-space symmetries, we create a flexible and expressive representation for olfactory mixtures, offering a potential solution for modeling other multi-component systems. Furthermore, we provide a method towards interpretable modeling of mixture components interactions studying the attention weights of mixture components in CheMix and studying how molecular information attends to itself within a mixture.

While POMMix shows promising results, we acknowledge several limitations. The small size of the available mixture dataset (< 1,000 samples) raises concerns about overfitting, despite our regularization efforts. Additionally, the limited coverage of chemical odorant space in current public datasets (only $\sim$200 unique odorant compounds) constrains the model's ability to generalize to a wider range of chemical compounds. We also observed challenges in generalizing to new datasets due to potential distribution shifts from varying experimental setups and human biases. Despite making conscious design choices on the modeling side, the interpretability of mixture components interactions in CheMix remains qualitative; further experimental investigations are required to validate our conjectures.

We believe that the primary bottleneck in advancing olfactory modeling is the generation of high-quality, diverse, and representative datasets. Future work should focus on expanding the coverage of chemical space, incorporating various experimental conditions (e.g., dilution, intensity), and collecting rich textual descriptions of odors. Such comprehensive datasets will be crucial for developing more robust, interpretable and generalizable models of olfactory perception.

## References

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711, 2021.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Aryan Amit Barsainyan, Ritesh Kumar, Pinaki Saha, and Michael Schmuker. GitHub - BioMachineLearning/openpom: Replication of the Principal Odor Map paper by Lee et al (2022). The model is implemented such that it integrates with DeepChem, 2024. URL `https://github.com/BioMachineLearning/openpom`.

A S Barwich. *Smellosophy*. Harvard University Press, 2022. ISBN 9780674278721.

Ann-Sophie Barwich and Elisabeth A Lloyd. More than meets the AI: The possibilities and limits of machine learning in olfaction. *Frontiers in neuroscience*, 16:981294, 2022. ISSN 1662-4548,1662-453X. doi: 10.3389/fnins.2022.981294. URL `https://doi.org/10.3389/fnins.2022.981294`.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL `https://arxiv.org/abs/1806.01261`.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2012. URL http://arxiv.org/abs/1206.5538.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Marc Brockschmidt. GNN-FiLM: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning*, 2019. URL http://arxiv.org/abs/1906.12192.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021. URL http://arxiv.org/abs/2105.14491.

C Bushdid, M O Magnasco, L B Vosshall, and A Keller. Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343:1370–1372, 2014. ISSN 0036-8075,1095-9203. doi: 10.1126/science.1249168. URL http://dx.doi.org/10.1126/science.1249168.

David Buterez, Jon Paul Janet, Steven J Kiddle, Dino Oglic, and Pietro Lió. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature communications*, 15:1517, 2024. ISSN 2041-1723,2041-1723. doi: 10.1038/s41467-024-45566-8. URL https://www.nature.com/articles/s41467-024-45566-8.

Jason B Castro, Travis J Gould, Robert Pellegrino, Zhiwei Liang, Liyah A Coleman, Famesh Patel, Derek S Wallace, Tanushri Bhatnagar, Joel D Mainland, and Richard C Gerkin. Pyrfume: A window to the world's olfactory data. *bioRxiv*, pp. 2022–09, 2022.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *arXiv [cs.LG]*, 2016. URL http://arxiv.org/abs/1603.02754.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Seyone Chithrananda, Judith Amores, and Kevin K Yang. Mapping the combinatorial coding between olfactory receptors and perception with deep learning. *bioRxiv*, pp. 2024–09, 2024.

Lucie Conchou, Philippe Lucas, Camille Meslin, Magali Proffit, Michael Staudt, and Michel Renou. Insect odorscapes: From plant volatiles to natural olfactory scenes. *Frontiers in physiology*, 10: 972, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2019.00972. URL http://dx.doi.org/10.3389/fphys.2019.00972.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 2020. URL http://arxiv.org/abs/2004.05718.

Jan C R Demyttenaere. The new european union flavouring regulation and its impact on essential oils: production of natural flavouring ingredients and maximum levels of restricted substances. *Flavour and fragrance journal*, 27:3–12, 2012. ISSN 0882-5734,1099-1026. doi: 10.1002/ffj.2093. URL http://dx.doi.org/10.1002/ffj.2093.

Amit Dhurandhar, Hongyang Li, Guillermo A Cecchi, and Pablo Meyer. Expansive linguistic representations to predict interpretable odor mixture discriminability. *Chemical senses*, 48:bjad018, 2023. ISSN 0379-864X,1464-3553. doi: 10.1093/chemse/bjad018. URL https://academic.oup.com/chemse/article/doi/10.1093/chemse/bjad018/7188234.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*, pp. 1437–1446. PMLR, 2018.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*, may 2019. URL https://github.com/pyg-team/pytorch_geometric.

Garrett B Goh, Charles Siegel, Abhinav Vishnu, and Nathan Hodas. Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018. ISBN 9781450355520. doi: 10.1145/3219819.3219838. URL https://dl.acm.org/doi/10.1145/3219819.3219838.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64:9–17, 2024. ISSN 1549-9596,1549-960X. doi: 10.1021/acs.jcim.3c01250. URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c01250.

Shion Honda, Shoi Shi, and Hiroki R Ueda. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv [cs.LG]*, 2019. URL http://arxiv.org/abs/1911.04738.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

IFRA. IFRA transparency list, 2024. URL https://ifrafragrance.org/priorities/ingredients/ifra-transparency-list.

Chaitanya K Joshi. Transformers are graph neural networks. *The Gradient*, 2020. URL https://thegradient.pub/transformers-are-graph-neural-networks/.

Gyuweon Jung, Jaehyeon Kim, Seongbin Hong, Hunhee Shin, Yujeong Jeong, Wonjun Shin, Dongseok Kwon, Woo Young Choi, and Jong-Ho Lee. Energy efficient artificial olfactory system with integrated sensing and computing capabilities for food spoilage detection. *Advanced science (Weinheim, Baden-Wurttemberg, Germany)*, 10:e2302506, 2023. ISSN 2198-3844. doi: 10.1002/advs.202302506. URL http://dx.doi.org/10.1002/advs.202302506.

Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, Celine Vens, Leander Schietgat, Kurt De Grave, Raquel Norel, DREAM Olfaction Prediction Consortium, Gustavo Stolovitzky, Guillermo A Cecchi, Leslie B Vosshall, and Pablo Meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355:820–826, 2017. ISSN 0036-8075,1095-9203. doi: 10.1126/science.aal2014. URL https://www.science.org/doi/full/10.1126/science.aal2014.

Brian Kelley et al. GitHub - bp-kelley/descriptastorus: Descriptor computation (chemistry) and (optional) storage for machine learning, 2024. URL https://github.com/bp-kelley/descriptastorus.

Rehan M Khan, Chung-Hay Luk, Adeen Flinker, Amit Aggarwal, Hadas Lapid, Rafi Haddad, and Noam Sobel. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 27:10015–10023, 2007. ISSN 0270-6474,1529-2401. doi: 10.1523/JNEUROSCI.1158-07.2007. URL https://www.jneurosci.org/content/jneuro/27/37/10015.full.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv [cs.LG]*, 2014. URL http://arxiv.org/abs/1412.6980.

Gregory R Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, 2015. URL https://www.cs.utoronto.ca/~rsalakhu/papers/oneshot1.pdf.

Joel Kowalewski and Anandasankar Ray. Predicting human olfactory perception from activities of odorant receptors. *iScience*, 23:101361, 2020. ISSN 2589-0042. doi: 10.1016/j.isci.2020.101361. URL https://www.cell.com/iscience/fulltext/S2589-0042(20)30548-4.

Greg Landrum et al. RDKit: Open-source cheminformatics software, 2022. URL https://github.com/rdkit/rdkit/releases/tag/Release_2022_03_4.

Hadas Lapid, David Harel, and Noam Sobel. Prediction models for the pleasantness of binary mixtures in olfaction. *Chemical senses*, 33:599–609, 2008. ISSN 0379-864X,1464-3553. doi: 10.1093/chemse/bjn026. URL https://academic.oup.com/chemse/article/33/7/599/330603?login=true.

Brian K Lee, Emily J Mayhew, Benjamin Sanchez-Lengeling, Jennifer N Wei, Wesley W Qian, Kelsie A Little, Matthew Andres, Britney B Nguyen, Theresa Moloy, Jacob Yasonik, Jane K Parker, Richard C Gerkin, Joel D Mainland, and Alexander B Wiltschko. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381:999–1006, 2023. ISSN 0036-8075,1095-9203. doi: 10.1126/science.ade4401. URL http://dx.doi.org/10.1126/science.ade4401.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv [cs.LG]*, 2020. URL http://arxiv.org/abs/2002.08264.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

André F Oliveira, Juarez LF Da Silva, and Marcos G Quiles. Molecular property prediction and molecular design using a supervised grammar variational autoencoder. *Journal of Chemical Information and Modeling*, 62(4):817–828, 2022.

M J Olsson. An integrated model of intensity and quality of odor mixtures. *Annals of the New York Academy of Sciences*, 855:837–840, 1998. ISSN 0077-8923,1749-6632. doi: 10.1111/j.1749-6632.1998.tb10672.x. URL https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1998.tb10672.x.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *arXiv [cs.LG]*, 2019. URL http://arxiv.org/abs/1912.01703.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. URL http://arxiv.org/abs/1709.07871.

Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russ Webb. Theory, analysis, and best practices for sigmoid self-attention, 2024. URL https://arxiv.org/abs/2409.04431.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Aharon Ravia, Kobi Snitz, Danielle Honigstein, Maya Finkel, Rotem Zirler, Ofer Perl, Lavi Secundo, Christophe Laudamiel, David Harel, and Noam Sobel. A measure of smell enables the creation of olfactory metamers. *Nature*, 588:118–123, 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2891-7. URL https://www.nature.com/articles/s41586-020-2891-7.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5898–5910. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/417fbbf2e9d5a28a855a11894b2e795a-Paper.pdf.

Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alex Wiltschko. A gentle introduction to graph neural networks. *Distill*, 6, 2021. ISSN 2476-0757. doi: 10.23915/distill.00033. URL http://dx.doi.org/10.23915/distill.00033.

C S Sell. On the unpredictability of odor. *Angewandte Chemie*, 45:6254–6261, 2006. ISSN 1433-7851. doi: 10.1002/anie.200600782. URL http://dx.doi.org/10.1002/anie.200600782.

Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. Benchmarking graphormer on large-scale molecular modeling datasets. *arXiv preprint arXiv:2203.04810*, 2022. URL https://arxiv.org/abs/2203.04810.

Daniel Shin, Gao Pei, Priyadarshini Kumari, and Tarek R Besold. Optimizing learning across multimodal transfer features for modeling olfactory perception. 2018.

Nima Shoghi, Adeesh Kolluru, John R Kitchin, Zachary W Ulissi, C Lawrence Zitnick, and Brandon M Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv [cs.LG]*, 2023. URL http://arxiv.org/abs/2310.16802.

Laura Sisson. Odor descriptor understanding through prompting. *arXiv [cs.LG]*, 2022. URL http://arxiv.org/abs/2205.03719.

Laura Sisson, Aryan Amit Barsainyan, Mrityunjay Sharma, and Ritesh Kumar. Olfactory label prediction on aroma-chemical pairs. *arXiv preprint arXiv:2312.16124*, 2023. URL http://arxiv.org/abs/2312.16124.

Kobi Snitz, Adi Yablonka, Tali Weiss, Idan Frumin, Rehan M Khan, and Noam Sobel. Predicting odor perceptual similarity from odor structure. *PLoS computational biology*, 9:e1003184, 2013. ISSN 1553-734X,1553-7358. doi: 10.1371/journal.pcbi.1003184. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003184.

Kobi Snitz, Ofer Perl, Danielle Honigstein, Lavi Secundo, Aharon Ravia, Adi Yablonka, Yaara Endevelt-Shapira, and Noam Sobel. SmellSpace: An odor-based social network as a platform for collecting olfactory perceptual data. *Chemical senses*, 44:267–278, 2019. ISSN 0379-864X,1464-3553. doi: 10.1093/chemse/bjz014. URL http://dx.doi.org/10.1093/chemse/bjz014.

Charles Spence, Marianna Obrist, Carlos Velasco, and Nimesha Ranasinghe. Digitizing the chemical senses: Possibilities & pitfalls. *International journal of human-computer studies*, 107:62–74, 2017. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2017.06.003. URL http://dx.doi.org/10.1016/j.ijhcs.2017.06.003.

Gary Tom, Riley J Hickman, Aniket Zinzuwadia, Afshan Mohajeri, Benjamin Sanchez-Lengeling, and Alán Aspuru-Guzik. Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. *Digital Discovery*, 2:759–774, 2023. doi: 10.1039/D2DD00146B. URL https://pubs.rsc.org/en/content/articlelanding/2023/dd/d2dd00146b.

Gary Tom, Stanley Lo, Samantha Corapi, Alan Aspuru-Guzik, and Benjamin Sanchez-Lengeling. Ranking over regression for bayesian optimization and molecule selection. *arXiv preprint arXiv:2410.09290*, 2024.

Ngoc B Tran, Daniel R Kepple, Sergey A Shuvaev, and A Koulakov. DeepNose: Using artificial neural networks to represent the space of odorants. *bioRxiv*, 97:6305–6314, 2018. doi: 10.1101/464735. URL https://proceedings.mlr.press/v97/tran19b/tran19b.pdf.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. URL http://arxiv.org/abs/1710.10903.

E Vigneau, P Courcoux, R Symoneaux, L Guérin, and A Villière. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. *Food quality and preference*, 68:135–145, 2018. ISSN 0950-3293,1873-6343. doi: 10.1016/j.foodqual.2018.02.008. URL https://www.sciencedirect.com/science/article/abs/pii/S0950329318301599?via%3Dihub.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888*, 2021.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, 2019. ISBN 9781450366663. doi: 10.1145/3307339.3342186. URL https://dl.acm.org/doi/pdf/10.1145/3307339.3342186.

Jennifer N Wei, Carlos Ruiz, Marnix Vlot, Benjamin Sanchez-Lengeling, Brian K Lee, Luuk Berning, Martijn W Vos, Rob W M Henderson, Wesley W Qian, D Michael Ando, Kurt M Groetsch, Richard C Gerkin, Alexander B Wiltschko, Jeffrey Riffel, and Koen J Dechering. A deep learning and digital archaeology approach for mosquito repellent discovery. *bioRxiv*, pp. 2022.09.01.504601, 2024. doi: 10.1101/2022.09.01.504601. URL https://www.biorxiv.org/content/10.1101/2022.09.01.504601v5.abstract.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28:31–36, 1988. ISSN 0095-2338,1520-5142. doi: 10.1021/ci00057a005. URL https://pubs.acs.org/doi/10.1021/ci00057a005.

Tali Weiss, Kobi Snitz, Adi Yablonka, Rehan M Khan, Danyel Gafsou, Elad Schneidman, and Noam Sobel. Perceptual convergence of multi-component mixtures in olfaction implies an olfactory white. *Proceedings of the National Academy of Sciences*, 109:19959–19964, 2012. doi: 10.1073/pnas.1208110109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1208110109.

Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature communications*, 14:2585, 2023. ISSN 2041-1723,2041-1723. doi: 10.1038/s41467-023-38192-3. URL https://www.nature.com/articles/s41467-023-38192-3.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63:8749–8760, 2020. ISSN 0022-2623,1520-4804. doi: 10.1021/acs.jmedchem.9b00959. URL https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.9b00959.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=OeWooOxFwDa.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 2017. URL http://arxiv.org/abs/1703.06114.

Hengrui Zhang, Jie Chen, James M Rondinelli, and Wei Chen. Molsets: Molecular graph deep sets learning for mixture property modeling. *arXiv preprint arXiv:2312.16473*, 2023.

Mengji Zhang, Yusuke Hiki, Akira Funahashi, and Tetsuya J Kobayashi. A deep position-encoding model for predicting olfactory perception from molecular structures and electrostatics. *npj systems biology and applications*, 10:76, 2024. ISSN 2056-7189,2056-7189. doi: 10.1038/s41540-024-00401-0. URL https://www.nature.com/articles/s41540-024-00401-0#Sec9.

# A APPENDIX

## A.1 GS-LF DATASET FILTERING

The open-source version of the GoodScents/Leffingwell (GS-LF) dataset by Barsainyan et al. (2024) initially contains 4983 molecules with 138 odor descriptors. These filters were applied in the following order:

- **Inorganic atom filter.** 110 molecules containing these atoms were removed: ["He", "Na", "Mg", "Al", "Si", "K", "Ca", "Ti", "V", "Cr", "Fe", "Co", "Cu", "Zn", "Bi"].
- **Duplicate SMILES filter.** 0 duplicate molecules were removed.
- **Salts and charged molecule filter**. 10 molecules containing charges (including salts) were removed.
- **Multimolecular filter.** 36 SMILES strings containing multiple molecules were removed (characterized by SMILES strings containing the "." character).
- **Molecular weight filter**. 1 molecule with MW < 20 was removed. 11 molecules with MW > 600 were removed.
- **Non-carbon molecule filter.** 1 molecule containing only non-carbon atoms was removed.

This filtering process results in a dataset of 4814 molecules, which was then used to train the POM.

## A.2 DETAILS OF MOLECULAR GRAPH REPRESENTATION

The node features used in the molecular graph representation as input to the POM GNN are 85-dimensional one-hot encoding vectors, encoding categorical information about the atoms. The edge features encode the categorical information about the bonds as 14-dimensional one-hot encoding vectors. The molecular information for the features are shown in Table 2.

Table 2: **Features for node and edge features of molecular graphs**. All categories are one-hot encoded and stacked to give a singular bit vector. `UNK` stands for "unknown", and is a catch-all category.

| Node features | Categories |
|---|---|
| Atomic number | 1 (hydrogen) to 54 (iodine), `UNK` |
| Atom degree | 0, 1, 2, 3, 4, 5, `UNK` |
| Formal charge | -2, -1, 0, 1, 2, `UNK` |
| Chirality | unspecified, CW, CCW, other, `UNK` |
| Number of hydrogens | 0, 1, 2, 3, 4, 5, 6, 7, 8, `UNK` |
| Hybridization | sp, sp2, sp3, sp3d, sp3d2, `UNK` |
| Aromatic | True/False |
| **Edge features** | **Categories** |
| Bond type | single, double, triple, aromatic, `UNK` |
| Is conjugated | True/False |
| In ring | True/False |
| Stereo-configuration | none, $Z$, $E$, *cis*, *trans*, any, `UNK` |

## A.3 HYPERPARAMETER SEARCH

We perform hyperparameter searches for the pre-training of both the POM GNN and the CHEMIX models. For the POM GNN, we use Optuna (Akiba et al., 2019), with the Tree-structured Parzen Estimator algorithm (Bergstra et al., 2011), with a budget of 200 runs. The final embedding space is fixed to 196 dimensions. The node GAT model and edge FiLM model is fixed to a single layer, while the global PNA model has 2 layers. The search space is defined as follows (bolded values are the optimal):

- Number of GRAPHNETS layers: [2, 3, **4**]
- Hidden dimensions for all models: [64, 128, 192, 256, **320**]
- Dropout rate: [0, 0.05, **0.1**, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5]
- Learning rate: [1e-2, 5e-3, 1e-3, 5e-4, **1e-4**, 5e-5]

For CHEMIX, the search was performed using Weights & Biases (Biewald, 2020) with BOHB (Falkner et al., 2018) algorithm and a budget of 200 runs. Early stopping was implemented with patience set to 100 epochs The search space was defined as follows (bolded values are the optimal):

- Embedding dimension: [32, 64, **96**, 128]
- Number of MolecularAttention (self attention) layers: values: [0, **1**, 2, 3]
- Number of attention heads: values: [1, 4, **8**, 16]
- Addition of an MLP head on top of MolecularAttention: ["True", **"False"**]
- Type of molecular aggregation: ["mean", **"pna"**, "attention"]
- Scaled cosine activation function: ["sigmoid", **"hardtanh"**]
- Attention type: ["standard", **"sigmoidal"**]
- Dropout rate: [0, 0.05, **0.1**, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5]
- Learning rate: [**8e-5**, 1e-4, 5e-4, 8e-4, 1e-3]
- Loss type: [**"mae"**, "mse", "huber"]

### A.4 PROCEDURE FOR SNITZ BASELINE

The Snitz baselines are reproduced following the procedure outlined in Snitz et al. (2013). There are three steps involved in optimizing the angle similarity model for the best descriptors. Prior to the optimization campaign, we normalize the 200 RDKIT features obtained from DESCRIPTASTORUS (Kelley et al., 2024) and average across all the molecules in the mixture.

In step 1, we determine the appropriate number of descriptors by randomly sampling 20,000 times, without replacement, $n \in [2, 200]$ descriptors, resulting in 199 sets of 20,000 samples each of predictions. We then evaluate the RMSE from the predictions of the similarity model for each value of $n$. The optimal number of descriptors was scored by minimizing $\mu_{RMSE} - \sigma_{RMSE}$ across all 20,000 samples for a given $n$. The appropriate number of descriptors was between $n = 5$ and $n = 7$, depending on the CV split.

In step 2, we evaluate the efficacy of each descriptor. We set the number of descriptors $n$ for each CV split based on step 1, and randomly sampled $n - 1$ descriptors 2,000 times. Then, cycling through each individual descriptor, we appended it to each set of sampled $n - 1$ descriptors, producing a vector of $n$ features, and again evaluated RMSE from the predictions of the similarity model. We take the mean RMSE from the 2,000 samples and the most relevant descriptors are determined by minimizing the mean RMSE.

In step 3, we first calculate the score of each descriptor from step 2. The 2,000 samples of $n - 1$ features with a specially appended feature $i \in [1, 200]$ provides a score for the $i$-th descriptor in the representation. This score for the $i$-th descriptor is given by

$$score(i) = \max\left(0, -\frac{\text{RMSE}_i - \mu_{\text{RMSE}}}{\sigma_{\text{RMSE}}}\right), \tag{1}$$

which only provides a positive score if the feature achieves lower RMSE than the average RMSE achieved over all features. Only positive scored features are kept. We then randomly sample, 4,000 times, $n = 5$ to $n = 7$ descriptors depending on the appropriate CV split out of the set of descriptors that performed better than the average RMSE value (i.e. positive score). Out of the 4,000 samples, we pick the best-performing set of descriptors (lowest RMSE) on the training set and perform a final evaluation on the test set. This procedure produces the values found in Section 3.

### A.5 XGBOOST MODELING

The XGBOOST model was given a maximum of 1,000 estimators and tree depth of 1,000. To ensure the model does not overfit, we use the validation set for early stopping, with a patience of 25 epochs. The model is trained with mean squared error, with a learning rate of 0.01.

### A.6 PARITY PLOTS FOR ALL BASELINE MODELS, CHEMIX, POMMIX, AND ZERO-BIAS POMMIX

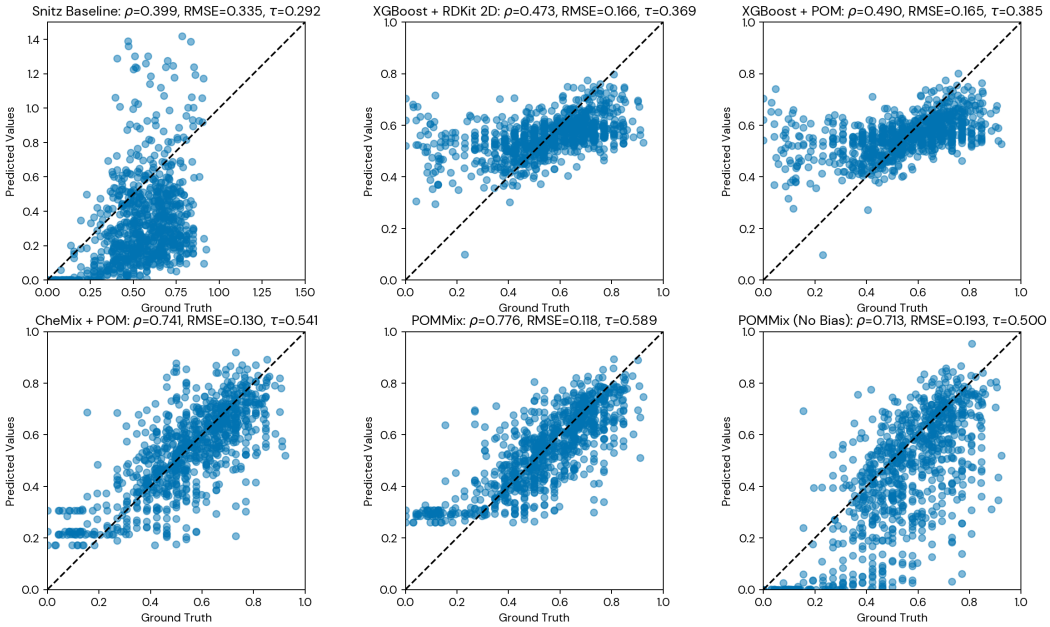


Figure 7: **Parity plots for all models evaluated**. Ground truth labels versus predicted values across all five cross-validation splits, with Pearson $\rho$, RMSE and Kendall $\tau$ reported.

### A.7 ADDITIONAL ABLATION STUDIES

In addition to the ablation studies in Section 3, we perform additional ablations of the POM graph model, and the CHEMIX prediction head.

In Table 3, we compare the chosen GRAPHNETS GNN with graph transformer models GRAPHORMER (Shi et al., 2022; Ying et al., 2021) and GPS (Rampášek et al., 2022). These models have shown state-of-the-art performance on large molecular datasets, such as Open Graph Benchmark (OGB) (Hu et al., 2020; 2021), the Open Catalyst Challenge (Chanussot et al., 2021), and the ZINC 250k dataset (Gómez-Bombarelli et al., 2018).

Table 3: **Other GNN models**. Validation results on GS-LF model for Graphormer and GPS models. While both graph transformers achieve state-of-the-art performances on larger molecular datasets, the lightweight and tuned POM performs as well as or better than the models.

| Model | Validation AUROC (↑) |
|---|---|
| GRAPHORMER (slim) | 0.8564 |
| GPS | 0.8647 |
| Our POM | **0.8843** |

In Table 4, we provide the cross validation test performance results for CHEMIX with frozen POM embeddings and different prediction heads. We train four additional models of CHEMIX with differ-

ent prediction heads: Mean + Linear, Concatenate + Linear, PNA-like + Linear, and unscaled cosine distance. The unscaled cosine distance prediction head achieves poor RMSE, but better correlation metrics ($\rho$ and $\tau$) when compared to the regressive prediction heads. Of the aggregation methods, we find that the PNA-like aggregation produces the best results with the linear regression head. The scaled cosine prediction head combines the strengths of both, achieving the best test performance across all three metrics. Additionally, we want to imbue the mixture embedding space with a notion of distance and similarity. The scaled cosine similarity was finally chosen, based on our experiments, as the best POMMIX prediction head.

Table 4: **Ablation of CHEMIX prediction head**. 5-fold cross validation metrics for CHEMIX with various prediction heads. The mean and standard deviation are reported. The scaled cosine distance is what was finally chosen for POMMIX, and is reproduced here for comparison.

| | Test predictive performance | | |
|---|---|---|---|
| CHEMIX prediction head | Pearson $\rho$ (↑) | RMSE (↓) | Kendall $\tau$ (↑) |
| Mean + Linear | 0.085 ± 0.123 | 0.204 ± 0.018 | 0.097 ± 0.057 |
| Concatenate + Linear | 0.266 ± 0.089 | 0.201 ± 0.009 | 0.190 ± 0.063 |
| PNA-like + Linear | 0.477 ± 0.066 | 0.174 ± 0.005 | 0.295 ± 0.048 |
| Cosine distance | 0.676 ± 0.049 | 0.208 ± 0.009 | 0.436 ± 0.028 |
| Scaled cosine distance | **0.746 ± 0.030** | **0.130 ± 0.007** | **0.545 ± 0.032** |

### A.8 PRE-TRAINING WITH AUGMENTED DATA

Due to the scarcity of mixture data, especially those with perceptual similarities between single molecules, we sought to investigate if we could augment data using available larger mono-molecular datasets. We investigated if the Jaccard distance between the odor descriptors of two molecules (obtained from GS-LF) was a good proxy to pairwise single-molecule perceptual similarities. Based on 75 single-molecular pairwise perceptual similarity measurements already in our dataset, we discovered a modest correlation ($\sim$0.49) between the Jaccard distance and the perceptual similarity (Figure 8a). Thus, we pre-trained CHEMIX with this augmentation strategy with a total of 15571 augmented datapoints, followed by fine-tune training of POMMIX , but found that it did not provide improved structure for the embedding space for single-molecular mixtures (Figure 8b), where the pairwise distance of the POMMIX mono-molecular mixture embeddings remained the same. Additionally, the pre-training causes reduced model performance in all tracked metrics for CHEMIX (Table 5).

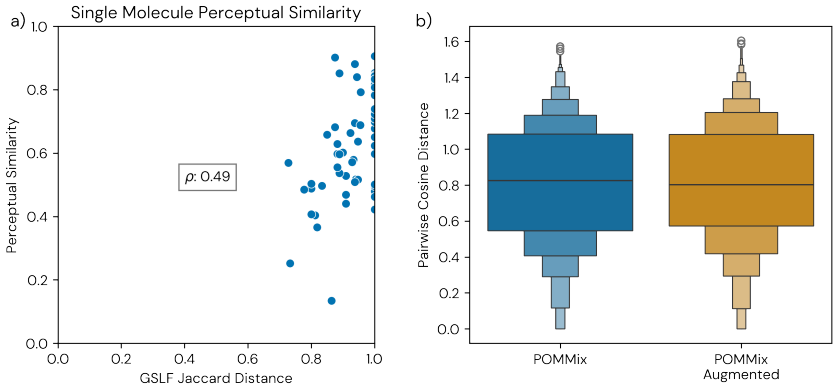


Figure 8: **Augmentation with GS-LF odor label Jaccard similarities**. **a**) Correlation between the Jaccard distance of the GS-LF odor labels of two single molecules, versus their perceptual similarity. **b**) Boxen plot of all pairwise cosine distances between the embeddings of single molecules for POMMIX, with and without augmentation.

Table 5: **Model performances on pre-training with augmented data**. 5-fold cross validation metrics for CHEMIX pre-trained with and without augmented data. The mean and standard deviation are reported.

| | Test predictive performance | | |
|---|---|---|---|
| Model | Pearson $\rho$ (↑) | RMSE (↓) | Kendall $\tau$ (↑) |
| CHEMIX + POM (augmented) | 0.628 ± 0.048 | 0.163 ± 0.007 | 0.479 ± 0.039 |
| CHEMIX + POM | **0.746 ± 0.030** | **0.130 ± 0.007** | **0.545 ± 0.032** |

### A.9 ATTENTION HEATMAP EXAMPLES

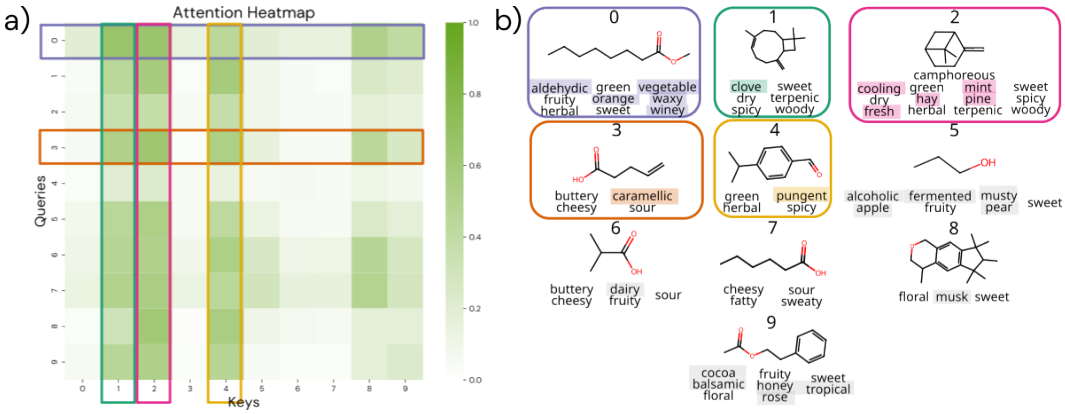


Figure 9: **Mixture attention map example with 10 components**. **a)** Sigmoid attention heatmap, compounds with 3 or more significant interactions (cutoff=0.5) are highlighted. **b)** Example mixture with molecules and their odor labels. Most interacting molecules are highlighted, and unique labels have a shaded rectangle.

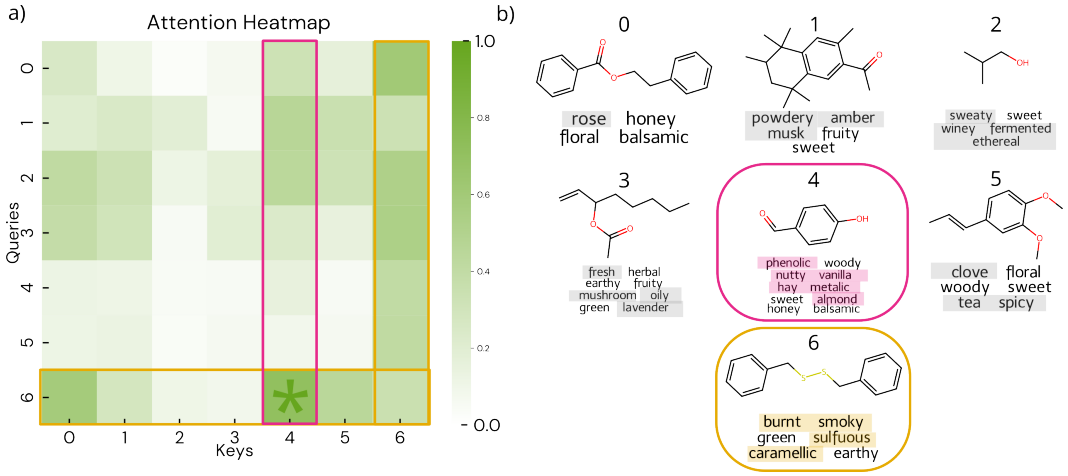


Figure 10: **Mixture attention map example with 7 components**. **a)** Sigmoid attention heatmap, compounds with 3 or more significant interactions (cutoff=0.5) are highlighted. Strongest interaction is indicated with an asterisk. **b)** Example mixture with molecules and their odor labels. Most interacting molecules are highlighted and unique labels have a shaded rectangle.

### A.10 MIXTURE SET LABEL-GUIDED STRUCTURAL INSIGHTS ON KEY MOLECULES

To derive structural heuristics across the entire set of unique mixtures, we analyze the key-ed molecules associated with extrema attention weight values for each query, focusing on queries that interact "strongly" with a key (an interaction is considered strong if the attention weight is above 0.5). We visualize the UMAP (McInnes et al., 2018) of the POM embeddings projected by CHEMIX through one linear layer of key-ed molecules exclusively found as maximizing/minimizing attention weights (Figure 11, left). We observe a clear separation between the two classes, suggesting that certain types of molecules are prioritized (high interactions) or de-prioritized (low interactions) when it comes to updating the molecular embeddings within a mixture.

We then performed hierarchical clustering (Müllner, 2011) of these key-ed molecules with SCIPY (Virtanen et al., 2020) based on the pairwise Jaccard similarity of the binary GS-LF odor descriptor labels. We selected a few representative molecules for each of the clusters and observed strong structural differences between them (Figure 11, right). This is implicitly expected from the structure-property relationship between scent and molecular structure. More importantly, we note that molecules within clusters are generally either "strongly" or "weakly" interacting, suggesting our model established a relationship between specific molecular structures and attention weight values. Through this analysis, we observe that ester/aldehydes with long alkane chains tend to have low interaction keys (cluster 2, 3 and 7; Figure 11, right), while sulfur-containing molecules and molecules containing aromatic rings tend to be highly interacting ones (cluster 1, 4 and 5; Figure 11, right). These structural insights derived from label-driven clustering confirm the idea that certain molecules receive more attention than others.

One possible explanation for why ester/aldehydes with long alkane chains are "non-interacting" keys would be that such molecules generally have a pleasant, sweet, or fruity smell. These odor descriptions are highly prevalent in the mixture dataset (119 ($58.62\%$) "sweet" molecules and 92 ($45.32\%$) "fruity" molecules in 203 unique molecules across the 743 unique mixtures), and could therefore not be informative in distinguishing mixtures. On the other hand, sulfur-containing molecules generally have a pungent, garlicky smell and occur less in the mixture dataset dataset (10 ($4.92\%$) "garlic" molecules).

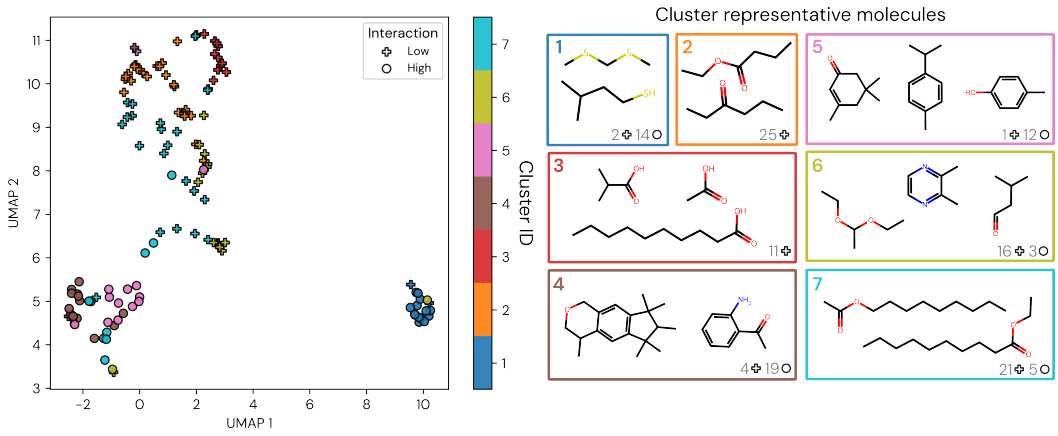


Figure 11: **Visualizing the embeddings of maximally/minimally interacting key-ed molecules across unique mixtures**. *(Left)* UMAP visualization of the embeddings of key-ed molecules exclusively found as maximizing/minimizing attention weights, for each query exhibiting significant interaction (attention weight > 0.5) across all unique mixtures. The interaction strength, determined by the attention weight, of the molecules are indicated by the markers. The molecules are colored by cluster identity. Molecules without GS-LF labels are excluded from the visualization. *(Right)* Representative molecules for each of the label-based Jaccard distance clusters. The number of strong/weak interaction molecules for each cluster is indicated in the bottom right corner of each box.

Table 6: **Model performances of MolSets architectures**. 5-fold cross validation metrics on mixture data with MolSets, adapted to be compatible with molecular mixture data. The mean and standard deviation are reported.

| Model | Pearson $\rho$ (↑) | RMSE (↓) | Kendall $\tau$ (↑) |
|---|---|---|---|
| MolSets SAGEConv | 0.418 ± 0.063 | 0.171 ± 0.014 | 0.294 ± 0.051 |
| MolSets DMPNN | 0.329 ± 0.092 | 0.178 ± 0.015 | 0.198 ± 0.071 |
| POMMIX | **0.779 ± 0.028** | **0.118 ± 0.004** | **0.596 ± 0.022** |