MIXORNET: ENCODING SCENT MIXTURES VIA COMBINATORIAL PROTEIN ACTIVATIONS FOR OLFACTORY PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Perceptual modeling of complex scent mixtures remains largely underexplored in computational olfaction. Existing approaches share two critical limitations: (1) they rely solely on the chemical properties of constituent odourants, neglecting selective olfactory receptor (OR) binding, and (2) they model perception via pairwise molecular attention, ignoring the fact that humans perceive mixtures holistically. We propose MixORNet, the first framework to encode scent mixture perception at the receptor level. MixORNet represents the mixture as a fully connected graph and uses a graph attention network (GAT) to extract holistic mixture-level features. These features further interact with a panel of N OR proteins via cross-attention, yielding N-dimensional OR-activation fingerprints, with each dimension denoting the activation probability of each receptor. MixORNet achieves state-of-the-art performance on perceptual distance prediction and multilabel olfactory annotation of scent mixtures. Moreover, it can reproduce the olfactory white phenomenon, aligning with established psychophysics. Notably, ablation of the OR panel size reveals that the optimal N is substantially smaller than the approximately 400 human ORs in specific scenarios, providing the first AI-derived estimate for the intrinsic dimension of the human olfactory perceptual subspace. The main code is available in the anonymous repository `https://anonymous.4open.science/r/MixORNet/`

1 INTRODUCTION

The sense of smell is directly linked to our everyday experience, from food and beverages to consumer products such as perfumes, where pleasant olfactory stimuli carry substantial commercial value (Feng et al., 2026; Taleb et al., 2024; Zhang et al., 2026). Smell perception has two indispensable foundations: the binding of volatile molecules to olfactory receptors (ORs) and the combinatorial coding of OR-activation patterns (Malnic et al., 1999). At the molecular level, olfactory perception begins in the nasal epithelium, where OR proteins selectively bind with volatile molecules (Buck & Axel, 1991). Malnic et al. (2004) reported that there are approximately 400 functional ORs. At the receptor level, each odourant activates a characteristic combination of multiple ORs, which constitutes the neural representation of odour identity Malnic et al. (1999). This mechanism enables us to distinguish an enormous variety of smells from a finite receptor repertoire.

Motivated by these findings, the computational olfaction community has made substantial progress in quantitative structure-odour relationship (QSOR) modeling, which predicts single-odourant perception from molecular structure. Early approaches relied on hand-crafted physicochemical descriptors and shallow machine learning models (Keller et al., 2017). More recently, GNNs have enabled end-to-end structure-to-perception modeling, with the principal odour map (Lee et al., 2023) achieving human-level performance in odour-quality prediction. However, these advances address only a single odourant, whereas natural scent stimuli are always mixtures. The few existing mixture-level models (Tom et al., 2025; Anonymous, 2026; Sisson et al., 2025) highlight two critical challenges. First, the perception of scent mixtures involves complex interactions among components. When multiple odourants co-occur at the olfactory epithelium, they induce competitive antagonism, synergistic enhancement, and even mixed-emergent receptor activation (Xu et al., 2020; Rospars et al., 2008), rendering the mixed OR-activation pattern an emergent property irreducible to the linear

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

combination of its component activations. We must shift the research paradigm to treat the scent mixture as a unique entity rather than a superposition of its components. Second, existing methods model scent mixture perception solely in molecular chemical space, overlooking receptor-level combinatorial coding. This fundamental mismatch leads to the well-documented structure–odour discontinuity: structurally similar molecules can smell entirely differently, whereas structurally diverse molecules can smell alike (Lee et al., 2023). We suppose that this discontinuity is interpretable through OR combinatorial coding, where similarity in OR-activation patterns reflects the perceptual similarity (Malnic et al., 1999). Hence, predicting the OR-activation patterns for scent mixtures is indispensable.

We address both challenges with MixORNet, the first computational framework to encode scent-mixture perception at the level of OR-activation. MixORNet has two key designs. First, each mixture is represented as a fully connected graph over its constituent odourants, with node features extracted by a pretrained directed message-passing neural network (D-MPNN) (Anonymous, 2026). A graph attention network (GAT) with residual connections further processes this graph to encode the mixture embedding. Second, the mixed embedding interacts with a panel of N OR proteins encoded by ESM-2 (Lin et al., 2023) via cross-attention, yielding an N-dimensional OR-activation fingerprint where each dimension denotes the predicted activation probability of the corresponding receptor. This fingerprint is applied directly to downstream tasks, including predicting perceptual distance and multilabel annotations of scent mixtures. Both designs enjoy rigorous biological grounding in the modulation model of Xu et al. (2020). First, the modulation model indicates that receptor responses to individual molecules are shaped by the presence of all co-occurring molecules. In an extreme case, mixture-emergent activation occurs when no single constituent alone activates a given OR. MixORNet models each mixture as a complete graph with single-molecule features as nodes, sharing the same initial single-molecule feature space. The subsequent GAT produces a mixture-level embedding that spans regions of this space unreachable by any single-component feature, which is consistent with mixture-emergent activation phenomena. Second, given that identical molecular pairs produce antagonism at some receptors and enhancement at others, we apply cross-attention between the mixed embedding and ESM-2 OR features, allowing each receptor to selectively query its own relevant activation signal. MixORNet achieves state-of-the-art performance on DREAM2024 (Satarifard et al., 2025) and DREAM2025 (Sage Bionetworks, 2025) and reproduces the *"olfactory white"* phenomenon, demonstrating that this model is aligned with established psychophysics. Moreover, the optimal OR panel size is substantially smaller than the canonical $\sim$400 human ORs in practical scenarios, providing a novel AI-driven estimate for the intrinsic dimension of the human olfactory perceptual subspace. In conclusion, our contributions are threefold.

- MixORNet is the first model to introduce receptor-level combinatorial coding into scent mixture perception, which is a paradigm shift from existing QSOR models. We demonstrate that explicitly predicting the OR-activation fingerprints for scent mixtures can yield state-of-the-art performance on two scent-mixture perception tasks.
- MixORNet enjoys clear biological groundings. The mixture-as-graph design yields mixture-level embeddings that extend the single-molecule feature space and align with the modulation model of Xu et al. (2020). The cross-attention mechanism further enables each receptor to selectively query its own relevant activation signal, thereby realizing selective binding (Malnic et al., 1999).
- Experimental results on the two-scent mixture perception challenges demonstrate the effectiveness of MixORNet. Moreover, MixORNet can reproduce the *"olfactory white"* phenomenon and estimate the intrinsic dimensions of the olfactory perception subspace, validating its scientific significance.

2 Related works

2.1 The QSOR model.

QSOR modeling has been studied for decades. Early work (Keller et al., 2017) demonstrated that physicochemical descriptors with shallow regression models can predict the intensity and quality of 49 odour attributes. Sharma et al. (2021) encoded SMILES strings through deep neural networks to improve generalization over descriptor-based baselines. The Principal Odour Map (Lee et al., 2023)

trained a GNN end-to-end on the curated GoodScents/Leffingwell dataset (Lee et al., 2023), producing a molecular odour space that matched the descriptions of odour quality provided by trained human panelists. Despite these advances, extending QSOR to scent mixtures remains underexplored. OdourPair (Sisson et al., 2025) applied MPNN-GNN architectures to aroma-chemical pairs, capturing nonlinear perceptual relationships in binary blends. POMMix Tom et al. (2025) introduced a hierarchical extension of POM that aggregates single-molecular embeddings via pairwise self-attention. In contrast, MixScentNet (Anonymous, 2026) represents scent mixtures as unified graph structures and extracts mixture-level features holistically, yielding improved perceptual prediction over component-wise aggregation baselines. MixORNet shares this mixture-as-graph design but addresses a fundamentally distinct problem: deriving OR-activation fingerprints for scent mixtures and validating their utility in downstream perceptual prediction tasks. Such an objective has not been addressed in prior work.

2.2 OR-MOLECULE ACTIVATION PREDICTION MODEL

The OR-activation prediction has long been restricted to single molecules. Early work by Bushdid et al. (2014) trained a support vector machine on molecular chemical descriptors to screen agonists for human ORs. Kowalewski & Ray (2020) subsequently showed that ML-derived OR-activation profiles across 34 human ORs could predict 146 distinct perceptual qualities and demonstrated that each odour percept could be recovered from a small OR subset. Hladiš et al. (2023) presented the first deep learning model to jointly encode the receptor sequence and the molecular graph for OR-activation prediction, combining ProtBERT embeddings with a tailored GNN and outperforming baselines on a dataset of 46,700 OR-molecule pairs. More recently, Chithrananda et al. (2024) demonstrated that ESM-2-derived OR-activation profiles substantially improve percept label assignment. Despite this progress, all these prediction models operate in the single-molecule setting, without any mechanism to model how scent mixtures interact at the receptor level. In contrast, MixORNet bridges both QSOR modeling and OR-molecule activation prediction by introducing mixture-as-a-graph and OR-mixture cross-attention. The final receptor activation fingerprint directly reflects the joint engagement of the olfactory repertoire, and its effectiveness is demonstrated in two olfactory perception tasks for scent mixtures.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND NETWORK OVERVIEW

Given a scent mixture $A = \{a_1, a_2, \ldots, a_m\}$ and a panel of N OR proteins, MixORNet produces an N-dimensional activation fingerprint $\mathbf{P}_A = \{p_1, p_2, \ldots, p_n\}$, where each p_i denotes the predicted activation probability of the i-th receptor. This fingerprint is applied to two downstream tasks: (1) predicting the perceptual label for A and (2) estimating the perceptual distance between two mixtures A and B. The mathematical definition of the two problems can be formalized as:

$$\begin{cases} \hat{\mathbf{y}}_A = \text{MLP}(\mathbf{P}_A) \in \mathbb{R}^K, & \text{Task 1} \\ \hat{d}_{\{A,B\}} = D(\mathbf{P}_A, \mathbf{P}_B) \in [0,1], & \text{Task 2} \end{cases} \tag{1}$$

where $\hat{\mathbf{y}}_A \in \mathbb{R}^K$ is a K-dimensional vector of predicted label scores, and we directly apply a simple multilayer perceptron (MLP) to regress it. $D(\cdot,\cdot)$ is a distance function operating on the activation fingerprints. We used the continuous Tanimoto distance, which has been shown to outperform other distance metrics for comparing OR-activation patterns (Rácz et al., 2018).

The network overview is shown in Figure 1, which consists of three components: encoders for single-molecule components and OR proteins, a mixture encoder to capture unified mixture-level features, and a cross-attention module for OR-mixture selective binding. During evaluation, users select the OR panel (a set of amino acid sequences) and provide the scent mixture (a set of SMILES strings for its components) as inputs. The protein and single-molecule encoders then encode them into deep features. The single-molecule features are organized as a complete graph and fed into a 3-layer residual GAT to obtain mixture-level features. The mixture-level features and N OR protein features are incorporated via OR-mixture cross-attention, resulting in N-dimensional OR-activation fingerprints.

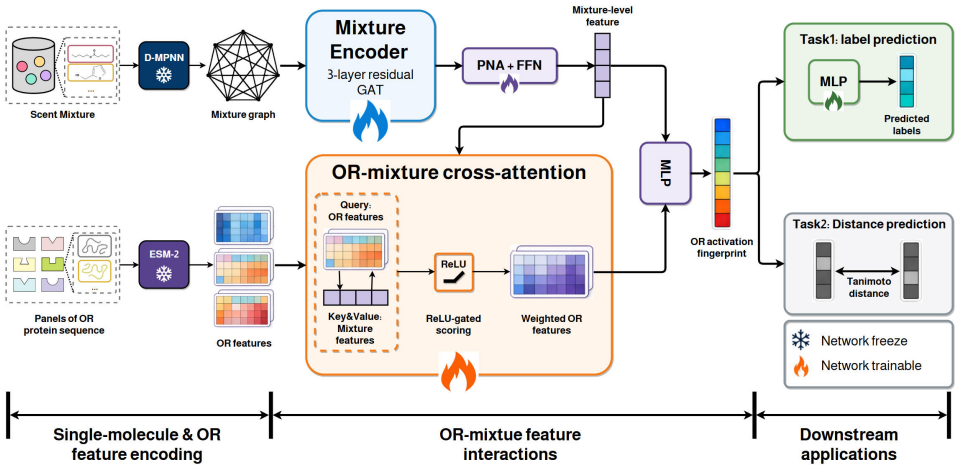


Figure 1: Overview of the network.

3.2 Single-molecule and protein encoders

Single-molecule encoder. Each constituent molecule $a_i \in A$ is encoded by a D-MPNN with pretrained weights from Anonymous (2026). They pretrained the model on ∼1 million PubChem molecules via a masked regression objective: given the molecular graph with RDKit-based features as input, the model predicts the set of Mordred descriptors (Moriwaki et al., 2018) with 85% of the targets randomly masked. The RDKit-based features and the Mordred descriptors are computable from SMILES strings and free of olfactory annotations. The D-MPNN weights are **frozen** during training, encoding each a_i into a fixed embedding $\mathbf{h}_{a_i} \in \mathbb{R}^{d_{mol}}$ with a molecular dimension d_{mol}. These features serve as the nodes in the mixture graph. Note that there is no prior knowledge from any olfactory annotation, aiming to show that molecular-level physicochemical properties are essential to olfactory perceptions.

OR protein encoder. The N OR proteins are encoded by ESM-2 (Lin et al., 2023), a transformer-based protein language model pretrained on 250 million sequences from UniRef (Suzek et al., 2007) via masked language modeling, which encourages the model to encode evolutionary conservation and implicit protein structure information without annotated olfactory data. We employ the 650 M-parameter variant and obtain a fixed-dimensional embedding for each OR, where the ESM-2 weights are also **frozen** during training, resulting in N protein embeddings $\{\mathbf{R}_j\}_{j=1}^{n}$, where $\mathbf{R}_j \in \mathbb{R}^{L \times d_{or}}$ and (L, d_{or}) are the lengths of the amino acid residues and OR feature dimensions, respectively. We freeze ESM-2 (Lin et al., 2023) to show that MixORNet's effectiveness stems from its architecture, which combines molecular- and receptor-level features to model selective OR-mixture binding.

3.3 Mixture encoder and OR-mixture cross-attention

Mixture encoder. Let $\mathbf{H}^{(0)} = \{\mathbf{h}_{a_1}, ..., \mathbf{h}_{a_m}\} \in \mathbb{R}^{m \times d_{\text{mol}}}$ denote the D-MPNN embeddings of the m constituent molecules in mixture A. The number of components m depends on the scent mixture. If the mixture has fewer than m components, we use zero-padding to expand it. We apply the 3-layer GAT network with residual connections to extract the mixture-level representations. Specifically, the first and second layers apply post-norm residual attention, which can be formulated as :

$$\mathbf{H}^{(i)} = \text{Dropout}\left(\text{LN}\left(\text{MHA}_i\left(\mathbf{H}^{(i-1)}, \mathbf{H}^{(i-1)}, \mathbf{H}^{(i-1)}; \mathbf{\Phi}_i\right) + \mathbf{W}_{\text{res}}^{(i-1)}\mathbf{H}^{(i-1)}\right)\right), i = 1, 2 \quad (2)$$

where $\mathbf{\Phi}_i$ represents the attention parameters for the multihead attention module $\text{MHA}_i(*)$. $\text{LN}(*)$ denotes layer normalization. $\mathbf{W}_{\text{res}}^{(i-1)}$ are linear transformations for $\mathbf{H}^{(i-1)}$ to allow richer residual information. Layer 3 uses a single-head attention followed by ELU activation applied to the residual sum:

$$\mathbf{H}^{(3)} = \text{LN}\left(\text{ELU}\left(\text{MHA}_3\left(\mathbf{H}^{(2)}, \mathbf{H}^{(2)}, \mathbf{H}^{(2)}; \mathbf{\Phi}_3\right) + \mathbf{W}_{\text{res}}^{(2)}\mathbf{H}^{(2)}\right)\right). \quad (3)$$

Padding positions are excluded at every layer via a Boolean key-padding mask (set to $-\infty$). Finally, we adopt a principal neighborhood aggregation (PNA) readout (Corso et al., 2020) to derive the

mixture-level embeddings, which concatenate four complementary statistics computed over valid (nonpadded) molecules:

$$\mathbf{r}_A = \left[\, \mu(\mathbf{H}^{(3)}) \parallel \sigma(\mathbf{H}^{(3)}) \parallel \min(\mathbf{H}^{(3)}) \parallel \max(\mathbf{H}^{(3)}) \,\right] \in \mathbb{R}^{4d_{mol}}. \tag{4}$$

A feedforward network FFN then compresses the concatenated statistics to the working dimension: $\mathbf{z}_A = \mathrm{FFN}(\mathbf{r}_A) \in \mathbb{R}^{d_{mol}}$. We use the PNA because it treats the scent mixture as a complete entity. Specifically, the mean and standard deviation (μ, σ) capture the central tendency and compositional diversity of the constituent distribution, whereas the extremal statistics $(\min, \max)$ preserve information about the most salient molecular contributors that could be removed by mean pooling. The four statistics enable the encoder to represent nonlinear, nonadditive functions of the component set, which is indispensable in modeling the holistic nature of human perception (Coureaud et al., 2022).

OR-mixture cross-attention. Given the mixture embedding $\mathbf{z}_A$ and the per-residue ESM-2 embeddings $\{\mathbf{R}_j\}_{j=1}^{n}$, we first unify their feature dimensions into d_h via linear projection. Afterward, a cross-attention module introduces a molecule-specific interaction for each OR. The module projects protein residues into the query space and the mixture into the key and the value space:

$$\mathbf{q}_j = \mathbf{R}_j \mathbf{W}_Q^\top \in \mathbb{R}^{L \times d_h}, \quad \mathbf{k} = \mathbf{z}_A \mathbf{W}_K^\top \in \mathbb{R}^{d_h}, \quad \mathbf{v} = \mathbf{z}_A \mathbf{W}_V^\top \in \mathbb{R}^{d_h}, \tag{5}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learned linear projections. A per-residue compatibility score is then computed between each amino acid residue of receptor j and the mixture:

$$e_{j,s} = \frac{\mathbf{q}_{j,s} \cdot \mathbf{k}}{\sqrt{d_h}}, \quad s = 1, ..., L. \tag{6}$$

A scalar gate α is derived from the value projection to modulate the overall attention magnitude:

$$\alpha = \mathbf{v}^\top \mathbf{w}_\alpha \in \mathbb{R}, \tag{7}$$

where $\mathbf{w}_\alpha \in \mathbb{R}^{d_h}$ is a learnable vector. The final per-residue attention weight of receptor j is formed by combining the score and the gate via the ReLU activations:

$$\tilde{w}_{j,s} = \mathrm{ReLU}(e_{j,s}) \cdot \alpha + b_\alpha, s = 1, ..., L, \tag{8}$$

where b_α is a learnable scalar bias. We do not use softmax because the module should be able to assign zero attention to all residues when the mixture is entirely incompatible with a receptor. The interaction representation for receptor j is then computed as the weighted sum across all residues:

$$\tilde{\mathbf{r}}_j = \sum_{s=1}^{L} \tilde{w}_{j,s}\, \mathbf{R}_{j,s} \in \mathbb{R}^{d_h}. \tag{9}$$

The final receptor-mixture feature is obtained by concatenating the weighted protein representation with the mixture embedding:

$$\mathbf{g}_j = \left[\, \tilde{\mathbf{r}}_j \parallel \mathbf{z}_A \,\right] \in \mathbb{R}^{2d_h}, j = 1...n. \tag{10}$$

An MLP with a sigmoid head $\sigma(*)$ is applied to pool the concatenated features, mapping each $\mathbf{g}_j$ to the predicted activation probability:

$$p_j = \sigma(\mathrm{MLP}(\mathbf{g}_j)) \in [0, 1], j = 1...n. \tag{11}$$

and the full set $\mathbf{P}_A = \{p_1, \ldots, p_n\}$ constitutes the OR-activation fingerprint of mixture A.

3.4 Training details

The OR panel selection. We select the OR panels from the M2OR database (Lalis et al., 2024), which contains 46,564 OR-molecule activation pairs with 1,237 unique ORs. For the default MixORNet, we select the $N = 100$ panels via stratified sampling. Specifically, we rank the ORs via the number of molecules that can activate them and then select ORs by uniformly sampling 100 indices over the rank-ordered list, ensuring that the panels span both broadly tuned and narrowly tuned ORs. This occurs because we find that narrowly tuned ORs help discriminate differences in scent mixtures, and $N = 100$ achieves sufficiently strong performance on both downstream tasks while maintaining favorable computational efficiency. We also conduct a detailed exploration of how to pick the OR panels in 4.4.

Trainable modules and loss functions. During downstream fine-tuning, we freeze the D-MPNN and ESM-2 and train only the mixture encoder and the cross-attention module. For scent mixture label prediction, we train an additional multilabel MLP for final-score regression, as indicated in Eq. (1). For loss functions, **Task 1** minimizes the mean squared error (MSE) loss between the predicted 51-dimensional label vector $\hat{\mathbf{y}}_A$ and the ground-truth annotation $\mathbf{y}_A$. **Task 2** minimizes the simulated annealing loss between the predicted perceptual distance $\mathbf{d}_{(A,B)}$ and the ground truth perceptual distance $\hat{\mathbf{d}}_{(A,B)}$. This loss, called the $\mathrm{MSEtoPLCC}$ loss, linearly interpolates the MSE and Pearson linear correlation coefficient (PLCC) loss to include a warmup period of T_{ramp} epochs to stabilize early training. The definition is as follows:

$$\mathcal{L}_{\text{MSEtoPLCC}}(t) = (1 - \alpha_t)\,\mathcal{L}_{\text{MSE}} + \alpha_t\,\mathcal{L}_{\text{PLCC}}, \qquad \alpha_t = \min\left(\frac{t}{T_{\text{ramp}}}, 1\right), \tag{12}$$

where t is the current epoch index, and the PLCC loss is defined as follows:

$$\mathcal{L}_{\text{PLCC}} = 1 - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \frac{(\hat{\mathbf{d}}_i - \text{avg}(\hat{\mathbf{d}}_i))^\top(\mathbf{d}_i - \text{avg}(\mathbf{d}_i))}{\|\hat{\mathbf{d}}_i - \text{avg}(\hat{\mathbf{d}}_i)\|_2\,\|\mathbf{d}_i - \text{avg}(\mathbf{d}_i)\|_2 + \varepsilon}, \tag{13}$$

where $\mathcal{B}$ is the batch size and $\text{avg}(*)$ is the mean of the vector. In all the experiments, we use the full batch on each database. Both tasks use the Adam optimizer (Kingma & Ba, 2014) with a step learning rate schedule that halves the rate every 1,000 epochs.

4 Experiments

4.1 Experimental setup

Datasets and evaluation metrics. We evaluate MixORNet on two olfactory mixture benchmarks. For multilabel prediction, DREAM2025 (Sage Bionetworks, 2025) contains more than 650 mixtures, with perceptual labels collected from trained panelists via the rate-all-that-apply (RATA) method across 51 odour descriptors. For perceptual distance estimation, DREAM2024 (Satarifard et al., 2025) contains 865 mixture pairs with perceptual distance measurements onto a continuous 0-1 scale, where 0 denotes indistinguishable mixtures and 1 denotes the maximal perceptual distance. For DREAM2025, we report the per-sample average Pearson linear correlation coefficient (Avg PLCC) and the cosine distance between the predicted and ground-truth 51-dimensional score vectors. For DREAM2024, we report the PLCC and the KRCC between the predicted and ground-truth distances. All the experiments are conducted on Ubuntu 16.04 with an Intel Xeon Silver 4214R CPU and an NVIDIA RTX 3090 GPU.

Baseline methods and setups. We compare MixORNet against both classical and deep learning baselines. The classical baselines include the Snitz angle metric (Snitz et al., 2013) and XGBoost (Tyagi et al., 2024), which are equipped with three different features: MolT5 (Edwards et al., 2022), RDKit descriptors (RDKit Contributors, 2006), and POM embeddings (Lee et al., 2023). Deep learning baseline comparisons include CheMeleon (Green et al., 2026), MolSets (Zhang et al., 2024), POMMix (Tom et al., 2025), and MixScentNet Anonymous (2026). CheMeleon is a single-molecule encoder, and we compare it to show that our mixed encoder and OR-mixture cross-attention are indispensable for achieving significant performance gains. Specifically, we directly sum the component embeddings of CheMeleon to obtain the mixture-level embeddings, and use the same MLP head as MixORNet for label prediction and cosine distance for perceptual distance prediction. CheMeleon is unfrozen during cross-fold validation. MolSets was originally proposed for battery electrolyte conductivity prediction and is adapted here by replacing its readout layer with our multilabel regression head for DREAM2025 and by using the cosine distance between mixture-level features for DREAM2024 (details in the Appendix). The POMMix combines the POM encoder with the CheMix aggregation module. We evaluate two variants of it: **POM+CheMix**, where only CheMix is fine-tuned, and **POMMix**, where both POM and CheMix are jointly fine-tuned end-to-end. MixScentNet is a state-of-the-art method that models a scent mixture as a complete graph and extracts mixture-level features for downstream perception tasks. We compare it to show that our MixORNet, with the OR-mixture cross-attention mechanism, can match its superior performance while improving robustness on unseen molecules.

Table 1: Standard 5-fold cross-validation on DREAM2025 (multilabel regression task) and DREAM2024 (perceptual distance prediction task).

Method (Standard Split)	DREAM2025 (label)		DREAM2024 (distance)	
	Avg PLCC(↑)	cos-dist (↓)	PLCC (↑)	KRCC (↑)
Snitz baseline	/	/	0.4154±0.0031	0.3154±0.0532
XGBOOST + MOLT5	0.6726±0.0113	0.2058±0.0060	0.4691±0.0596	0.3669±0.0214
XGBOOST + RDKIT	0.7210±0.0163	0.1745±0.0116	0.4891±0.0644	0.3872±0.0381
XGBOOST + POM	0.6854±0.0145	0.1981±0.0093	0.5243±0.0625	0.4159±0.0125
MolSets	0.5742±0.0212	0.2615±0.0082	0.6175±0.0598	0.4330±0.0490
Chemeleon	0.7466±0.0113	0.1588±0.0056	0.7140±0.0236	0.5163±0.0211
POM+Chemix	0.7277±0.0189	0.1712±0.0133	0.7357±0.0432	0.5569±0.0353
POMMix	0.7496±0.0163	0.1575±0.0104	0.7558±0.0276	**0.5700±0.0258**
MixScentNet	<u>0.7611±0.0140</u>	<u>0.1523±0.0088</u>	**0.7853±0.0078**	0.5615±0.0126
MixORNet	**0.7651±0.0141**	**0.1489±0.0083**	<u>0.7842±0.0143</u>	<u>0.5648±0.0316</u>

Table 2: Unseen molecule cross-validation on DREAM2025 (multilabel regression task) and DREAM2024 (perceptual distance prediction task).

Method (Unseen Split)	DREAM2025 (label)		DREAM2024 (distance)	
	Avg PLCC(↑)	cos-dist (↓)	PLCC (↑)	KRCC (↑)
Snitz baseline	/	/	0.4262±0.0578	0.3105±0.0610
XGBOOST + MOLT5	0.5756±0.0514	0.2529±0.0310	0.3882±0.1351	0.2322±0.0936
XGBOOST + RDKIT	0.6087±0.0625	0.2338±0.0361	0.4222±0.1074	0.2787±0.0736
XGBOOST + POM	0.6306±0.0260	<u>0.2237±0.0178</u>	0.4703±0.0937	0.3280±0.0733
MolSets	0.5360±0.0639	0.2769±0.0332	0.5353±0.0625	0.2769±0.0324
Chemeleon	0.6368±0.0348	0.2239±0.0239	0.5943±0.0719	0.4088±0.0429
POM+Chemix	0.6236±0.0374	0.2393±0.0272	0.6124±0.0906	0.4202±0.0571
POMMix	0.6354±0.0458	0.2336±0.0314	0.6015±0.1039	0.4233±0.0550
MixScentNet	<u>0.6495±0.0364</u>	0.2262±0.0265	<u>0.6724±0.0842</u>	<u>0.4318±0.0543</u>
MixORNet	**0.6829±0.0194**	**0.1995±0.0139**	**0.7045±0.0556**	**0.4643±0.0433**

4.2 Performance Comparison

5-fold cross-validation. In this experiment, all the models are evaluated under 5-fold cross-validation with an 80%/20% split. We report the mean and standard deviation of the evaluation metrics over five folds. The results are presented in Table 1. The bold and underlined numbers are the best and second-best mean scores in each column. "↑" indicates that the larger the score is, the better the performance, and vice versa. Note that the "Snitz baseline" is only for the perceptual distance task and is not applicable to the label prediction task. We find that MixORNet achieves state-of-the-art performance on DREAM2025, with an Avg PLCC of $\mathbf{0.7651 \pm 0.0141}$ and a Cos-dist of $\mathbf{0.1489 \pm 0.0083}$, outperforming all the baselines. On DREAM2024, MixORNet records a PLCC of $\underline{0.7842 \pm 0.0143}$ and a KRCC of $\underline{0.5648 \pm 0.0316}$, matching the top-performing MixScentNet (gap of 0.0011, within overlapping standard deviations) and outperforming MolSets, CheMeleon, and POM+CheMix by $+26.98\%$, $+9.83\%$, and $+6.59\%$ in PLCC, respectively. Across both tasks, MixORNet is the only method producing a biologically interpretable OR-activation fingerprint, demonstrating that receptor-level encoding is effective for scent mixture perception.

Model robustness to unseen molecules. Table 2 reports the 5-fold cross-validation performance under the unseen-molecule protocol. Specifically, for each fold, we randomly select a molecule from the database as the '*unseen molecule*', and all the samples that contain it are assigned to the test set. This greedy accumulation continues until the test set reaches 20% samples, with all remaining samples assigned to training. Consequently, no SMILES string in any test sample appears in any training sample, providing a strictly harder evaluation than the standard split does. We find that the model performances in Table 2 are generally inferior to those in Table 1. Nevertheless, MixORNet achieves the best performance across both benchmarks. On DREAM2025, MixORNet

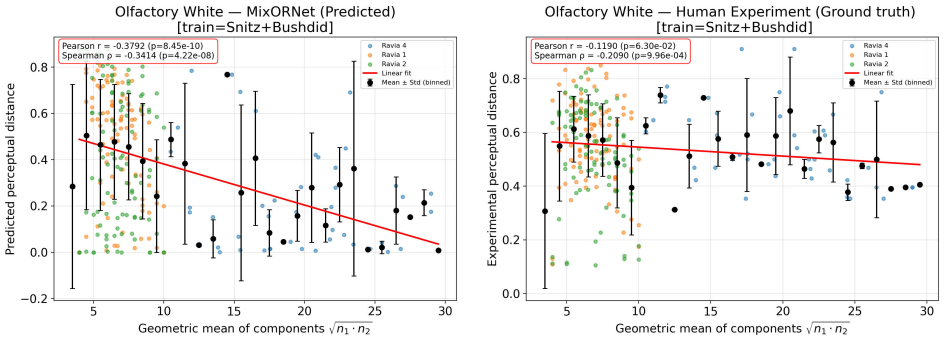


Figure 2: Reproducing the '*olfactory white*' phenomena on the Ravia subset in DREAM2024.

records an Avg PLCC of $\mathbf{0.6829 \pm 0.0194}$ and a cos-dist of $\mathbf{0.1995 \pm 0.0139}$, outperforming the second-best method MixScentNet by $+5.14\%$ in Avg PLCC, indicating that MixORNet is more robust to unseen molecules. For DREAM2024, MixORNet achieves a PLCC of $\mathbf{0.7045 \pm 0.0556}$ and a KRCC of $\mathbf{0.4643 \pm 0.0433}$, surpassing MixScentNet by $+4.77\%$ in the PLCC and $+7.53\%$ in the KRCC. MixORNet also has smaller standard deviations across both metrics, indicating a more stable fold-level performance. We attribute the generalization capability to the OR-activation fingerprint. The frozen ESM-2 provides a fixed biological fingerprint grounded in receptor biology, which is more robust to shifts in molecular identity. In contrast, MixScentNet relies entirely on molecular physicochemical features without receptor-level regularization, making it more susceptible to distributional shifts when novel molecules are present.

4.3 Reproducing the Olfactory White Phenomena

The "*olfactory white*" phenomenon (Weiss et al., 2012) refers to the perceptual convergence of complex odour mixtures: as the number of diverse components increases, the mixture loses its distinctive character and converges toward a common, undifferentiated percept. To examine whether MixORNet reproduces this phenomenon, we conduct a leave-one-subset-out evaluation on DREAM2024, which comprises three subsets: Snitz (Snitz et al., 2013), Ravia (Ravia et al., 2020), and Bushdid (Bushdid et al., 2014). For DREAM2024, the olfactory white hypothesis implies that pairs of mixtures with more components should be predicted to be closer. The representative result (trained on Snitz+Bushdid, tested on Ravia) is shown in Figure 2, where each point represents a mixture pair with horizontal coordinates $\sqrt{n_1 \cdot n_2}$ and vertical coordinates of the predicted perceptual distance. The geometric mean is used to denote the complexity of the mixture pair and is aligned with Weiss et al. (2012). The black error bars denote the within-bin mean $\pm$ standard deviation. The red ordinary least squares (OLS) regression line of the MixORNet predictions confirms a significant negative slope, with Pearson $r = -0.3792$ ($p = 8.45 \times 10^{-10}$) and Spearman $\rho = -0.3414$ ($p = 4.22 \times 10^{-8}$). These results demonstrate that the predicted distances decrease as mixture complexity increases, consistent with the olfactory white phenomenon. Notably, MixORNet captures this convergence trend even more strongly than the human ground-truth ratings do. This behavior of the MixORNet emerges without any olfactory-white supervision, providing strong evidence that the OR-activation fingerprint encodes a biologically meaningful representation of mixture identity. More results can be found in A.3.

4.4 Discussions on Selecting OR Panels

OR panel selection involves two key considerations: database choice and panel size. First, we explore OR panels from the HORDE database (Olender et al., 2013), which provides 853 human OR gene entries (386 functional OR genes and 467 pseudogenes). Unlike M2OR (Lalis et al., 2024), HORDE does not annotate OR-molecule activations but rather pseudogenes that are generally believed not to participate in functional olfactory signaling. Selection is based on the pseudogene label to pick functional ORs. Second, for the panel size, we choose $N = \{10, 20, 40, 80, 100, 200, 300, 400\}$ on the M2OR and HORDE databases. The results are shown in Figure 3. For each point, we conduct 5-fold standard cross-validation and report the mean perfor-

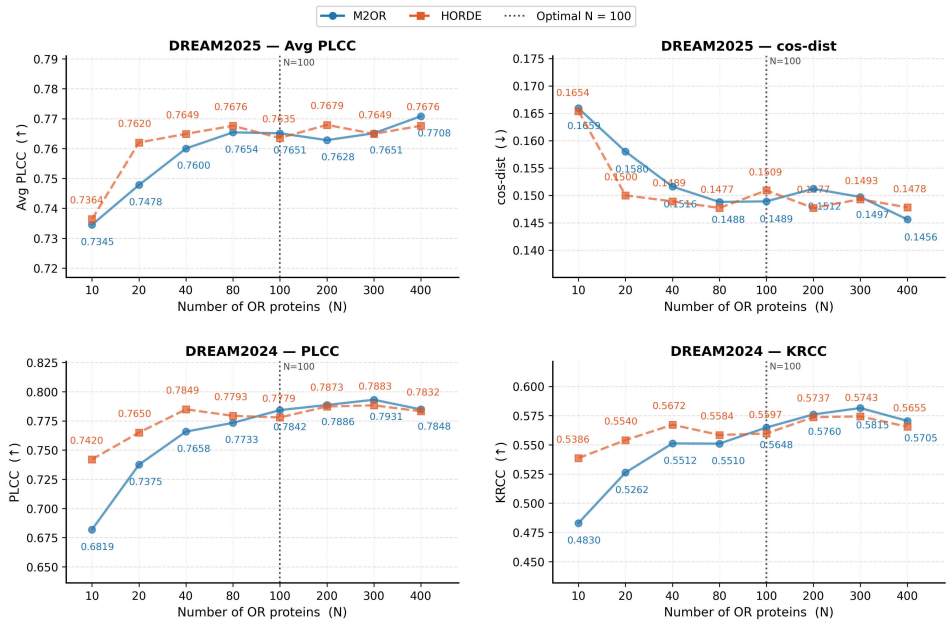


Figure 3: Performance on selecting different numbers of ORs from the M2OR (Lalis et al., 2024) and the HORDE (Olender et al., 2013).

mance scores. Two key observations emerge. First, the model performance improves with increasing OR panel size, but then saturates. For M2OR-selected ORs, the most substantial gains occur between $N = 10$ and $N = 100$: DREAM2025 increases the PLCC from 0.7345 to 0.7654 ($+4.1\%$), and DREAM2024 increases the PLCC from 0.6819 to 0.7842 ($+14.9\%$). Further expansion to 200-400 yields only marginal, non-monotonic improvements. For HORDE-selected ORs, saturation occurs earlier, with $N \approx 40$ sufficient for strong performance. This implies that the perceptual subspace spanned by both benchmarks can be effectively captured by a compact OR panel smaller than the canonical $\sim$400 human ORs. Such findings are highly valuable. When designing the electronic nose in practical scenarios, 40-100 bio-inspired sensing elements are sufficient, avoiding the cost and integration overhead of deploying the full OR complement. Second, HORDE-selected ORs consistently outperform M2OR-selected receptors at small panel sizes ($N \leq 40$), with the advantage being most pronounced for DREAM2024 (ΔPLCC $= +0.0601$ at $N = 10$). We believe this is because we only select functional ORs in HORDE. As N increases, the performance gap narrows because M2OR sampling progressively covers the functional OR subspace. This observation also has practical applications: in fragrance and flavour design, customers can select functional ORs with high fragrance relevance to develop a more efficient odour retrieval system.

5 Conclusions

We propose MixORNet, the first computational olfaction model to address scent-mixture perception, by explicitly modeling selective OR binding. MixORNet applies a GAT-based mixture encoder to extract the mixture-level embedding that is further coupled with ESM-2 representations of OR panels via cross-attention, yielding an N-dimensional OR-activation fingerprint. MixORNet achieves state-of-the-art performance on downstream perceptual tasks, with particularly strong generalization capability to unseen molecules. Moreover, MixORNet reproduces the olfactory white phenomenon without explicit supervision, confirming alignment with human olfactory systems. We further explore the OR panel size on DREAM2024 and DREAM2025 and find that the perceptually optimal N is substantially smaller than the canonical $\sim$400 human OR, which is the first AI-derived estimate for the intrinsic dimensions of the olfactory perceptual subspace. We hope that MixORNet can open a novel direction at the intersection of computational olfaction and receptor biology, and the OR-activation fingerprint paradigm will serve as a useful perceptual representation for future work in fragrance design and olfactory-inspired sensing systems.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) to aid and polish writing. We have not used generative AI tools for any other task, such as generating synthetic data, formulating mathematical claims, or implementing the core model architecture. We have reviewed all AI-polished parts in the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have released our source code and logs of the main results in the manuscript in the anonymous repository `https://anonymous.4open.science/r/MixORNet/`.

REFERENCES

Anonymous. Mixscentnet: A multiscale graph-based framework for predicting scent mixture perception. *Under review*, 2026.

Linda Buck and Richard Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–187, 1991.

Caroline Bushdid, Marcelo O Magnasco, Leslie B Vosshall, and Andreas Keller. Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343(6177):1370–1372, 2014.

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

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in neural information processing systems*, 33:13260–13271, 2020.

Gérard Coureaud, Thierry Thomas-Danguin, Jean-Christophe Sandoz, and Donald A Wilson. Biological constraints on configural odour mixture perception. *Journal of Experimental Biology*, 225 (6):jeb242274, 2022.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Dewei Feng, Wei Dai, Carol Li, Alistair Pernigo, and Paul Liang. Smellnet: A dataset for sensor-based smell recognition and mixture prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103061–103092, 2026.

William Green, Jackson Burns, Akshat Shirish Zalte, Charlles Abreu, Jochen Sieg, Christian Feldmann, and Miriam Mathea. Deep learning foundation models from classical molecular descriptors. 2026.

Matej Hladiš, Maxence Lalis, Sébastien Fiorucci, and Jérémie Topin. Matching receptor to odorant with protein language and graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Joel Kowalewski and Anandasankar Ray. Predicting human olfactory perception from activities of odorant receptors. *IScience*, 23(8), 2020.

Maxence Lalis, Matej Hladiš, Samar Abi Khalil, Loïc Briand, Sébastien Fiorucci, and Jérémie Topin. M2or: a database of olfactory receptor–odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Research*, 52(D1):D1370–D1379, 2024.

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, et al. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Bettina Malnic, Junzo Hirono, Takaaki Sato, and Linda B Buck. Combinatorial receptor codes for odors. *Cell*, 96(5):713–723, 1999.

Bettina Malnic, Paul A Godfrey, and Linda B Buck. The human olfactory receptor gene family. *Proceedings of the National Academy of Sciences*, 101(8):2584–2589, 2004.

Idan Menashe, Ronny Aloni, and Doron Lancet. A probabilistic classifier for olfactory receptor pseudogenes. *BMC bioinformatics*, 7(1):393, 2006.

Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1):4, 2018.

Tsviya Olender, Noam Nativ, and Doron Lancet. Horde: comprehensive resource for olfactory receptor genomics. In *Olfactory Receptors: Methods and Protocols*, pp. 23–38. Springer, 2013.

Anita Rácz, Dávid Bajusz, and Károly Héberger. Life beyond the tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of cheminformatics*, 10(1):48, 2018.

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(7836):118–123, 2020.

RDKit Contributors. Rdkit: Open-source cheminformatics. `https://www.rdkit.org`, 2006. Accessed: 2026-04-08.

Jean-Pierre Rospars, Petr Lansky, Michel Chaput, and Patricia Duchamp-Viret. Competitive and noncompetitive odorant interactions in the early neural coding of odorant mixtures. *The Journal of Neuroscience*, 28(10):2659–2666, 2008.

Sage Bionetworks. DREAM olfactory mixtures prediction challenge 2025. `https://www.synapse.org/Synapse:syn64743570/wiki/632379`, 2025. Synapse ID: syn64743570. Accessed: 2026-04-16.

Vahid Satarifard, Laura Sisson, Yikun Han, Pedro Ilídio, Matej Hladiš, Maxence Lalis, Xuebo Song, Wenjie Yin, Aharon Ravia, CiCi Xingyu Zheng, et al. High-fidelity tuning of olfactory mixture distances in the perceptual space of smell through a community effort. *biorxiv*, pp. 2025–12, 2025.

Anju Sharma, Rajnish Kumar, Shabnam Ranjta, and Pritish Kumar Varadwaj. Smiles to smell: decoding the structure–odor relationship of chemical compounds using the deep neural network approach. *Journal of Chemical Information and Modeling*, 61(2):676–688, 2021.

Laura Sisson, Aryan Amit Barsainyan, Mrityunjay Sharma, and Ritesh Kumar. Deep learning for odor prediction on aroma-chemical blends. *ACS omega*, 10(9):8980–8992, 2025.

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(9):e1003184, 2013.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Farzaneh Taleb, Miguel Vasco, Nona Rajabi, Mårten Björkman, and Danica Kragic. Can foundation models smell like humans? In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Gary Tom, Cher Tian Ser, Ella M Rajaonson, Stanley Lo, Hyun Suk Park, Brian K Lee, and Benjamin Sanchez-Lengeling. Does this smell the same? learning representations of olfactory mixtures using inductive biases. *Machine Learning: Science and Technology*, 6(3):035063, 2025.

Pankaj Tyagi, Anju Sharma, Rahul Semwal, Uma Shanker Tiwary, and Pritish Kumar Varadwaj. Xgboost odor prediction model: finding the structure-odor relationship of odorant molecules using the extreme gradient boosting algorithm. *Journal of Biomolecular Structure and Dynamics*, 42(20):10727–10738, 2024.

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(49):19959–19964, 2012.

Lu Xu, Wenze Li, Venkatakaushik Voleti, Dong-Jing Zou, Elizabeth MC Hillman, and Stuart Firestein. Widespread receptor-driven modulation in peripheral olfactory coding. *Science*, 368 (6487):eaaz5390, 2020.

Hengrui Zhang, Tianxing Lai, Jie Chen, Arumugam Manthiram, James M. Rondinelli, and Wei Chen. Learning molecular mixture property using chemistry-aware graph neural network. *PRX Energy*, 3:023006, Jun 2024. doi: 10.1103/PRXEnergy.3.023006. URL `https://link.aps.org/doi/10.1103/PRXEnergy.3.023006`.

Ziqi Zhang, Eunyeong Jin, Miguel Vasco, Farzaneh Taleb, Nona Rajabi, Alexandra Gutmann, Jonathan Williams, Antônio H Ribeiro, and Danica Kragic. Aligning mass spectra with molecular structure for fast computational olfaction. In *ICLR 2026 Workshop on Representational Alignment*, 2026.

Table A1: Ablation studies with standard 5-fold cross-validation on DREAM2025 and DREAM2024. Bold and underlined numbers are the best and second best results in each column.

Method (Standard Split)	DREAM2025 (label)		DREAM2024 (distance)	
	Avg PLCC(↑)	cos-dist (↓)	PLCC (↑)	KRCC (↑)
w/o GAT	0.7212±0.0208	0.1756±0.0124	0.7164±0.0267	0.4944±0.0322
w/o CA	0.7477±0.0179	0.1543±0.0113	0.7479±0.0301	0.5451±0.0258
w ESM-2(3B)	**0.7681±0.0119**	**0.1472±0.0074**	0.7770±0.0199	0.5554±0.0319
w MAE loss	0.7565±0.0056	0.1548±0.0030	0.7057±0.0564	0.4762±0.0440
w MSE loss	<u>0.7651±0.0141</u>	<u>0.1489±0.0083</u>	0.7193±0.0530	0.4817±0.0459
w PLCC loss	0.7647±0.0166	0.4866±0.0272	<u>0.7833±0.0209</u>	**0.5687±0.0262**
w MAE	/	/	0.7552±0.0455	0.5339±0.0534
w MSE	/	/	0.7587±0.0390	0.5392±0.0299
w cos-dist	/	/	0.7718±0.0133	0.5504±0.0269
MixORNet	<u>0.7651±0.0141</u>	<u>0.1489±0.0083</u>	**0.7842±0.0143**	<u>0.5648±0.0316</u>

A Appendix

A.1 Ablation studies

GAT and cross-attention module. We ablate the mixture-encoder GAT and the OR-mixture cross-attention (CA) module. For MixORNet without the GAT, we sum the mixed component features to aggregate the features. For MixORNet without CA, the raw mixture and protein features are concatenated directly and passed to the MLP to obtain the OR-activation fingerprint. The results are in Table A1. We find that removing the GAT causes the largest performance drop: DREAM2025 Avg PLCC decreases from 0.7651 to 0.7212 (-5.74%), and Cos-dist increases from 0.1489 to 0.1756 ($+17.9\%$); DREAM2024 PLCC decreases from 0.7842 to 0.7164 (-8.64%), and KRCC decreases from 0.5648 to 0.4944 (-12.46%). These degradations confirm that the summation of the naive component fails to yield an informative mixture-level embedding. Similarly, removing CA yields consistent but less severe degradation. DREAM2025 Avg PLCC decreases to 0.7477 (-2.27%), and DREAM2024 PLCC decreases to 0.7479 (-4.63%), confirming that the selective query mechanism is indispensable for OR-specific interaction modeling; without it, the OR fingerprint reduces to a receptor-blind projection that loses the combinatorial binding selectivity on which biological validity depends.

Different ESM-2 backbone. We further tested ESM-2 with 3B weights, which scaled up from the default 650 M variant to approximately 5 times more parameters, increasing the transformer layer count from 33 to 36 and expanding the per-residue embedding dimension from 1,280 to 2,560. Nevertheless, the gains are marginal. With respect to DREAM2025, ESM-2(3B) improves Avg PLCC from 0.7651 to 0.7681 ($+0.39\%$) and cos-dists from 0.1489 to 0.1472 (-1.14%). On DREAM2024, ESM-2(3B) is inferior to the 650 M variant, with the PLCC decreasing from 0.7842 to 0.7770 (-0.92%) and the KRCC decreasing from 0.5648 to 0.5554 (-1.66%). We attribute this to the relatively small scale of both DREAM benchmarks, where higher-dimensional embeddings may make it more difficult for the cross-attention module to learn OR-mixture binding from limited supervision. Therefore, we adopt ESM-2 (650M) as the default backbone for MixORNet due to its efficiency and effectiveness.

Different training loss. We ablate three alternative losses against the default settings. We find that the mean absolute error (MAE) loss yields the weakest DREAM2024 performance (PLCC 0.7057, KRCC 0.4762), as its uniform gradient signal lacks an incentive to learn the distribution structure of perceptual distances. In contrast, MSE loss improves DREAM2024 PLCC to 0.7193 by penalizing large errors more strongly. PLCC loss alone achieves the highest DREAM2024 KRCC by directly optimizing the rank-order correlation, but its scale invariance causes a catastrophic collapse in DREAM2025 label prediction (cos-dist 0.4866, versus MixORNet's 0.1489), as the predicted label vectors lose magnitude calibration. The $\mathrm{MSEtoPLCC}$ loss resolves this tension by first using MSE to calibrate the absolute prediction scale and then transitioning to PLCC to refine the rank-order cor-

Table A2: Comparison with official submission leaderboards on DREAM2025 (perceptual label prediction) and DREAM2024 (perceptual distance prediction).

Method	Avg PLCC (↑)	cos-dist (↓)	Rank (↓)
PL21	0.8043	0.1280	1
PL21	0.8053	0.1283	1
@yuanfang.guan	0.7928	0.1372	3
IM-Scent	0.7798	0.1448	4
@yuanfang.guan	0.7712	0.1477	5
PL21	0.7741	0.1495	5
Nylix	0.7664	0.1530	7
@yuanfang.guan	0.7554	0.1578	8
@jgburk	0.7394	0.1690	9
IM-Scent	0.7268	0.1825	10
MixORNet	0.7895	0.1412	3.5

(a) DREAM2025 (label prediction)

Method	MSE (↓)	MSE Rank (↓)	PLCC (↑)	PLCC Rank (↓)	Avg Rank (↓)
UMich CASI (1st)	0.007874	3	0.5011	2	2.5
ChemSenSim Lab (1st)	0.008578	5	0.5012	1	3
belfaction (1st)	0.007810	2	0.4302	5	3.5
D2Smell (1st)	0.0062	1	0.6075	1	1
ruhallah1020	0.009114	7	0.4668	3	5
PL21	0.007525	1	0.3581	10	5.5
Olfactory24UNIBA	0.008501	4	0.4038	7	5.5
Tamarin	0.008868	6	0.4032	8	7
who_nose	0.013576	12	0.4549	4	8
Songxuebo	0.013520	11	0.3725	9	10
MixORNet	0.0110	7.5	0.6007	1.5	4.5

(b) DREAM2024 (perceptual distance estimation)

relation, thereby achieving the best DREAM2024 PLCC (0.7842) and confirming that the warm-up is essential for jointly satisfying the scale-calibration and rank-discrimination requirements.

Different feature comparison measures. For Task 2, we ablate three distance metrics: MAE, MSE, and cosine distance to compare the OR fingerprints. Since the distance metric is used only for perceptual distance estimation, we report the results only for DREAM2024. As shown in Table A1, the MAE distance yields a PLCC of 0.7552 and a KRCC of 0.5339, whereas the MSE distance is 0.7587 and 0.5392. The cosine distance improves further to 0.7718 the PLCC and 0.5504 KRCC but still lags behind the default Tanimoto distance, which achieves the best performance of 0.7842 the PLCC and 0.5648 KRCC. These results are consistent with prior evidence that the continuous Tanimoto distance is particularly well suited for comparing bounded probability vectors, such as OR-activation fingerprints (Rácz et al., 2018).

A.2 Performance on official DREAM leaderboards.

We further compare MixORNet against official leaderboard submissions on DREAM2025 and DREAM2024 using the official train-validation-test splits provided by the competition organizers. All leaderboard results are sourced directly from the official websites. Specifically, for DREAM2025 (Sage Bionetworks, 2025), we train our MixORNet on the official training set and evaluate it on the validation sets, as the test set annotations are currently unavailable. The results are compared against the 1st-10th teams on the validation leaderboard[1] shown in Table A2a. For DREAM2024 (Satarifard et al., 2025), we use the official train-validation test split offered by the organizer in their GitHub repository[2]. The results are also compared against 1st–10th teams on the official test leaderboard[3] shown in Table A2b. The reported rank for MixORNet indicates its potential placement on the corresponding leaderboard without changing the original rank. Notably, for DREAM2024, the organizers assigned four 1st-place teams: team UMich CASI, ChemSenSim Lab, belfaction, and D2Smell. We denote them with '(1st)'.

On DREAM2025, we find that MixORNet achieves an average PLCC of 0.7895 and a cos-dist of 0.1412, placing at an average leaderboard rank of 3.5, surpassing the team IM-Scent and trailing only PL21 (Avg PLCC 0.8043, 0.8053, rank 1) and the third best @yuanfang.guan submission (Avg PLCC 0.7928, rank 3). With respect to DREAM2024, MixORNet achieves a PLCC of 0.6007 (rank 1.5), effectively matching and surpassing three of the four first-place methods by substantial margins, UMich CASI (0.5011), ChemSenSim Lab (0.5012), and belfaction (0.4302). Only D2Smell (0.6075) scores higher. MixORNet's overall average rank is 4.5 because the MSE is 0.0110, and the ranking is 7.5. This is because we use the $\mathrm{MSEtoPLCC}$ loss, where the MSE loss just serves as a warm-up objective. In conclusion, these leaderboard results demonstrate that explicitly modeling OR-activation fingerprints yields competitive performance on established mixture perception benchmarks. Such results have clear scientific significance: we introduce a fundamentally new paradigm grounded in receptor-level biology rather than in pure molecular physicochemical properties alone, and demonstrate that the OR-activation fingerprints for scent mixtures can yield superior performance on two challenges.

[1] https://www.synapse.org/Synapse:syn53470621/wiki/627131

[2] https://github.com/Satarifard/DREAM-olfactory-mixtures-prediction-challenge/

[3] https://www.synapse.org/Synapse:syn66484079/tables/

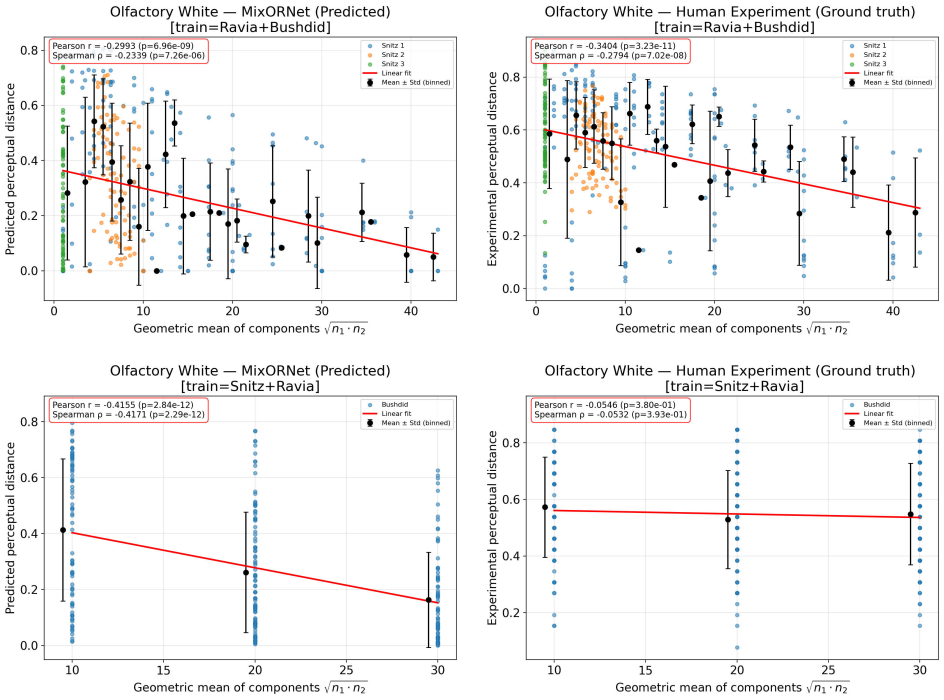


Figure A1: Reproducing the '*olfactory white*' phenomena on the Snitz Snitz et al. (2013) and Bushdid Bushdid et al. (2014) subsets in DREAM2024 Satarifard et al. (2025).

A.3 More Olfactory White Reproducing Results

We present two additional leave-one-subset-out evaluations on the Snitz (Snitz et al., 2013) and Bushdid (Bushdid et al., 2014) subsets of DREAM2024 (Satarifard et al., 2025) in Figure A1, complementing the results of Ravia in the main text. Across all three splits, MixORNet produces a statistically significant negative trend between mixture complexity and predicted perceptual distance, with p-value significance $p < 10^{-5}$ in every configuration, demonstrating that olfactory white reproduction is robust to diverse held-out subset settings. On the Snitz subset (train: Ravia+Bushdid), MixORNet yields Pearson $r = -0.2993$ ($p = 6.96 \times 10^{-9}$) and Spearman $\rho = -0.2339$ ($p = 7.26 \times 10^{-6}$), closely matching the significant human ground-truth correlations ($r = -0.3404$, $p = 3.23 \times 10^{-11}$; $\rho = -0.2794$, $p = 7.02 \times 10^{-8}$). The Bushdid subset (train: Snitz+Ravia) yields the most striking result. MixORNet produces the strongest olfactory white signal, with Pearson $r = -0.4155$ ($p = 2.84 \times 10^{-12}$) and Spearman $\rho = -0.4171$ ($p = 2.29 \times 10^{-12}$), whereas the human ground-truth ratings essentially have no correlation with complexity ($r = -0.0546$, $p = 0.38$; $\rho = -0.0532$, $p = 0.39$). We attribute this divergence to the Bushdid triangle discrimination paradigm, which measures upper bounds on discriminability rather than subjective similarity and introduces individual noise that hinders complexity-driven convergence. MixORNet recovers the olfactory white trend without any information about the measurement paradigm, demonstrating that the OR-activation fingerprint captures the latent psychophysical structure of the phenomenon even when individual human judgments fail to reveal it.

A.4 Dataset Details

DREAM 2025 and DREAM 2024. The 51 perceptual labels of DREAM2025 are listed as follows. Each mixture is rated on these descriptors by trained panelists via the rate-all-that-apply (RATA) method, with each descriptor paired with visual and odour references during training. Only subjects who achieved a test-retest correlation $R > 0.35$ on a 20-odourant pretest were admitted to the panel. The DREAM2025 organizers did not disclose how scores were aggregated across panelists or how mixtures were physically prepared.

```
Green, Cucumber, Herbal, Mint, Woody, Pine, Floral, Powdery,
Fruity, Citrus, Tropical, Berry, Peach, Sweet, Caramellic,
Vanilla, BrownSpice, Smoky, Burnt, Roasted, Grainy, Meaty,
Nutty, Fatty, Coconut, Waxy, Dairy, Buttery, Cheesy, Sour,
Fermented, Sulfurous, and Garlic.  Onion, Earthy, Mushroom,
Musty, Ammonia, Fishy, Fecal, Rotten.  Decay, Rubber,
Phenolic, Animal, Medicinal, Cooling, Sharp, Chlorine,
Alcoholic, Plastic, Ozone, Metallic
```

The training and validation sets of DREAM2024 (Satarifard et al., 2025) aggregate mixture pairs from three prior studies: Snitz et al. (2013), Ravia et al. (2020), and Bushdid et al. (2014). The three subsets differ in the experimental paradigm. Bushdid et al. (2014) quantified the upper bound of olfactory discrimination via a triangle test, presenting mixtures at three dilutions to control the odour intensity. The perceptual distance was defined as the proportion of successful discriminations. Ravia et al. (2020) constructed olfactory metamers: mixture pairs with disjoint molecular compositions but perceptually indistinguishable odours, and applied continuous similarity ratings. Snitz et al. (2013) collected direct similarity ratings to establish a chemical descriptor-based odour similarity metric. Since these paradigms differ substantially, the DREAM2024 organizers unified all the measurements onto a continuous 0–1 scale via linear normalization, and they acknowledge that this unification cannot fully correct paradigm-specific biases or systematic differences arising from distinct evaluation methods, indicating an inherent limitation of the unified dataset. The DREAM2024 test set (Satarifard et al., 2025) contained 92 unique mixtures forming 46 pairs, each containing 10 nonoverlapping intensity-matched components drawn from Bushdid et al. (2014). A total of 34 participants were recruited and completed the triangle discrimination tasks, in which odd-smelling vials from three participants were identified in random order. Specifically, two vials contained the same mixture, and one contained a different mixture. A mandatory 90-second rest period between trials minimized olfactory fatigue. Each pair of mixtures yielded 136–140 measurements.

M2OR and HORDE. The two databases share a common focus on ORs but differ substantially in terms of their scale, construction methodology, and research focus. HORDE (Olender et al., 2013) is a gene-centered resource built on an automated computational pipeline that aims to explore functional genes and pseudogenes of ORs. In the latest version, HORDE (Olender et al., 2013) catalogues 853 human OR entries, including 386 intact genes and 467 pseudogenes. HORDE also classifies all ORs into 18 families and 454 subfamilies, and its primary focus is on characterizing the genomic landscape of ORs, including gene nomenclature, chromosomal organization, interindividual variation, evidence of expression, and pseudogene identification. It applied a probabilistic classifier (CORP) (Menashe et al., 2006) for detecting functionally silent ORs that retain an intact open reading frame. In contrast, M2OR (Lalis et al., 2024) is built through systematic manual curation of 42 published studies spanning 25 years of OR bioassay research. M2OR operates at a relatively large experimental scale, aggregating 75,050 bioassay measurements across 51,395 distinct OR-molecule pairs, involving 1,246 unique receptor sequences from 11 mammalian species and 768 odourant compounds. The M2OR used in our manuscript is a subset of the entire database related to human olfaction. M2OR focuses on the functional recognition spectrum of ORs, especially in terms of which receptors respond to which molecules, alongside rich experimental metadata, including cell lines, G-protein types, assay concentrations, EC50 values, and stereochemical properties of ligands. In summary, HORDE addresses the question of what OR genes exist and how they are organized in the genome, whereas M2OR addresses the question of which ORs are functionally activated by which odourant molecules. The two databases thus operate at complementary levels of the OR knowledge hierarchy, with HORDE providing the genomic foundation and M2OR capturing the functional interaction layer built upon it.

A.5 Implementation details

MixORNet. There are three important feature dimensions in MixORNet. First, the D-MPNN backbone of (Anonymous, 2026) yields single-molecule embeddings with 2048 dimensions, which are computationally infeasible to process with limited GPU memory. We projected them into a working dimension of $d_{mol} = 512$ with a linear projection before feeding them into the mixture encoder. Second, the mixture encoder uses $d_{mol} = 512$ across all three attention layers, with 4 attention heads in layers 1–2 and 1 in layer 3. The dropout rate is 0.1 after each layer. The PNA readout

Table A3: Comparison between the original MolSets Zhang et al. (2024) and our modified version used as a baseline. **Left:** retained designs based on the original architecture. **Right:** modified parts.

Designs	Retained
Graph construction	✓
GNN backbone	✓
Global meanpool	✓
Activations: tanh	✓
MLP prediction head	✓

(a) Retained designs.

Designs	Original MolSets	Modified (ours)
Physical inputs	MW, frac, salt, molality	Graphs only
Encoders	2 (Solution+salt)	1 (Mixture)
Attention	Single-head	Multihead
Output modes	Regression only	Regression + distance
Applicability	Narrow (fixed prediction)	Broad (multitask)

(b) Modified designs.

concatenates the mean, standard deviation, min, and max statistics across valid molecular nodes, producing an $4 \times 512 = 2048$ dimensional vector, which is then projected to $d_h = 512$ via an MLP with ELU activation. Finally, the dimension of the ESM-2 per-residue embeddings is 1280 for ESM-2(650 M). We also reduce this dimension to $d_h = 512$ via a linear projection before conducting the molecule-gated cross-attention. The final MLP for OR-activation prediction comprises a single hidden layer with ReLU activation, BatchNorm1d, and a linear output layer. More details can be found in our released anonymous report code repository[4].

Classic baseline methods. We directly utilized the realization proposed by Tom et al. (2025), including the Snitz baseline Snitz et al. (2013) and XGBOOST with diverse feature embeddings. Details can be found in their open-source repository[5].

MolSets. MolSets (Zhang et al., 2024) was originally designed to predict the ionic conductivity of polymer electrolyte systems. It comprises two components: a solution network and a salt network. The solution component consists of a single-molecule SAGEConv-based GNN encoder (three layers with LayerNorm and ReLU, followed by global mean pooling and a linear layer concatenating molecular weight) and a single-head self-attention module that captures molecular interactions, with attention outputs aggregated via mass-fraction weighting. The structurally identical salt encoder produces a salt embedding that is concatenated with the solution embedding and fed into the prediction head. To adapt MolSets to olfactory tasks, we adopt a GNN-based approach to process odour mixtures, remove the salt encoder, restrict inputs to component molecular graphs, replace the prediction head with the same multilabel regression head as MixORNet, and replace single-head self-attention with multihead self-attention. The full details of the modifications are provided in Table A3b.

Chemeleon. CheMeleon (Green et al., 2026) adopts self-supervised pretraining but operates as a single-molecule encoder. It is included as a baseline to assess whether the naive aggregation of pretrained single-molecule features can match methods designed specifically for mixed tasks. The backbone is kept unchanged and equipped with two task-specific heads: the same MLP head as MixORNet for label prediction and a feature-summation head with cosine distance for perceptual distance estimation. The full backbone is fine-tuned end-to-end on both tasks.

POMMix and MixScentNet. We evaluate POMMix (Tom et al., 2025) and MixScentNet (Anonymous, 2026) without any modification and test them on the same training/testing splits as MixORNet does.

[4] https://anonymous.4open.science/r/MixORNet/

[5] https://github.com/chemcognition-lab/pom-mix