# REFINING OLFACTORY RECEPTOR REPRESENTATIONS FROM BINDING GRAPH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Predicting which olfactory receptors a given odorant molecule activates is a central challenge in computational olfaction. Current methods treat each molecule-receptor pair in isolation, representing receptors with frozen embeddings from general-purpose protein language models. The measured response profile of a receptor across many odorants, which directly encodes its functional selectivity, is never explicitly incorporated into the receptor representation. We introduce Olfactory Binding Graph (**OlfaGraph**), a three-stage procedure that explicitly models this relational evidence. First, experimentally confirmed receptor responses and non-responses are organized into a signed bipartite graph over odorants and receptors. Second, a two-layer graph neural network propagates information across these relations to produce response-aware receptor embeddings that encode the binding evidence observed during training. Third, the adapted embeddings are combined with the original protein and molecular features in a downstream predictor that generalizes to entirely novel odorants. Across three datasets spanning mammalian and insect olfaction, OlfaGraph ranks among the strongest methods in both seen- and cold-molecule settings, using only 1.6 million trainable parameters. Ablations reveal a domain-specific limitation: general-purpose protein embeddings offer little advantage over far simpler descriptors. When general-purpose encoders fall short for a task-relevant protein family, the accumulated experimental receptor-odorant measurements themselves can ground a more effective representation. The code for OlfaGraph is available at `https://anonymous.4open.science/r/Olfaction-Binding-5FE7`.

## 1 INTRODUCTION

Modeling how biological systems transform chemical information into olfactory perception remains a fundamental challenge. Odor perception begins when odorant molecules bind olfactory receptors (ORs), a family of G-protein-coupled receptors (GPCRs) expressed in the nose. Humans have approximately 350–400 functional OR types, and each odorant activates a subset of them, forming a combinatorial code that is subsequently processed by the brain to produce an odor percept Buck (2004); Malnic et al. (2004); Keller & Vosshall (2004). This poses a key problem for computational olfaction: predicting which ORs a molecule activates based on its chemical structure. Molecular docking, a widely used method for predicting how molecules bind to proteins, has limited applicability to olfaction: few OR structures have been experimentally resolved Billesbølle et al. (2023); Choi et al. (2023), while predicted receptor structures generally lead to less accurate docking results Karelina et al. (2023). Therefore, machine learning offers a promising alternative by learning odorant–receptor responses directly from molecular and receptor representations Wang et al. (2026).

A central challenge in odorant-receptor response prediction is generalization to entirely novel odorants, a regime known as the *cold-molecule setting*. The vast chemical space of potential odorants makes this a demanding test: without training responses for a new molecule, a model must transfer knowledge from previously observed odorant-receptor interactions. The receptors, however, remain known, and their measured responses to other odorants provide evidence of their selectivity. Incorporating this evidence into receptor embeddings offers a promising path toward richer receptor representations, enabling transfer to entirely unseen molecules. However, existing approaches Hladiš et al. (2023); Kroll et al. (2024); Chithrananda et al. (2026) largely overlook this opportunity. They commonly represent receptors using frozen protein language model embeddings Lin et al. (2023);



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.

Hayes et al. (2025), which capture general protein properties but are not tailored to olfaction-related tasks. Our experiments show that simple protein descriptors and even one-hot receptor identities can match or exceed pre-trained protein language model embeddings in several settings of response prediction (Section 5.1), suggesting that generic representations offer limited benefits for olfactory modeling. Moreover, existing approaches typically model each odorant-receptor pair independently, without explicitly leveraging the information shared across measured interactions. Yet, the response profile of a receptor across multiple odorants can provide informative evidence about its functional properties.

To address the identified limitations, we propose Olfactory Binding Graph (**OlfaGraph**), a graph-based approach for learning olfaction-specific receptor representations from measured odorant-receptor response relations. OlfaGraph consists of three components: (i) a signed bipartite graph that represents measured responses and non-responses between odorants and receptors; (ii) a graph neural network that propagates information across these relations and adapts pre-trained protein representations to olfactory response patterns; and (iii) a downstream predictor that combines the resulting response-aware receptor embeddings with the original protein and molecular features. By constructing receptor representations solely from training-set response relations, OlfaGraph can be applied directly to cold molecules without requiring any measurements for the query odorant. Across mammalian and insect datasets and both seen- and cold-molecule settings, the resulting response-aware representations consistently improve upon general-purpose protein representations, as well as upon models that predict molecule–receptor binding via learned, pair-specific interaction architectures.

## 2 Related Work

**From odor perception to receptor responses.** A growing line of research aims to predict how a molecule will smell directly from its chemical structure Keller et al. (2017); Lee et al. (2023); Tom et al. (2025); Xie et al. (2025); Su et al. (2026); Chithrananda et al. (2026). Such works map molecular representations to human-provided odor descriptors, but these labels are shaped by individual experience, language, and culture Majid (2021), making them inherently subjective and noisy. A more biologically grounded and tractable alternative is to predict the activation of olfactory receptors directly: measurable, species-specific, and upstream of perception. Several recent approaches target this particular task, including Hladiš et al. Hladiš et al. (2023), ProSmith Kroll et al. (2024), and LORAX McConachie et al. (2026). However, all these methods make predictions for each odorant-receptor pair in isolation: the observed response of a given pair is used only as a training label, meaning that relational patterns across other odorant-receptor pairs are never explicitly exploited. Since in most experimental setups a single receptor is tested against many different odorants Lalis et al. (2024); Carey et al. (2010); Hallem & Carlson (2006), its representation is the natural place to accumulate such cross-pair information. Moreover, in the cold-molecule setting, prediction quality depends entirely on the receptor representation, since no odorant-specific information is available to compensate for it. Despite this, all these response-prediction methods keep the protein representation frozen. LORAX is a partial exception: it explored fine-tuning the protein encoder alongside the molecular encoder, but found this hurt performance, and reverted to a frozen protein encoder as its default configuration. The frozen protein representations used by these response-prediction baselines are typically drawn from pre-trained protein language models, including ESM-1b Rives et al. (2021), ESM-2 Lin et al. (2023), ESM3 Hayes et al. (2025), and ProtT5 Elnaggar et al. (2021). Although these representations capture broad protein properties, they are trained without any knowledge of olfactory responses. Such off-the-shelf representations offer no guarantee of effectiveness in this specialized olfactory setting. We close these gaps by moving beyond frozen, per-pair protein representations, learning instead a protein representation that explicitly captures response-driven relational patterns across odorant-receptor pairs.

**Learning representations from interaction graphs.** Drug-target interaction prediction faces a similar problem: determining whether a molecule and a protein interact, based on a set of experimentally measured molecule-protein pairs. In this domain, graph-based approaches such as NeoDTI Wan et al. (2019), GeNNius Veleiro et al. (2024), and $H^2$GnnDTI Jing et al. (2025) encode known interactions as relations within a graph and learn representations on this basis. This principle provides a route to address the protein-representation gap in olfaction: the set of known interaction partners of a node is itself informative about that node. However, these methods cannot

be directly transferred to the olfaction setting. Predominantly, these drug approaches construct their graphs from confirmed positive interactions alone, treating the absence of a known interaction as missing information rather than as evidence in its own right; some also rely on auxiliary data (e.g., disease networks) with no analogue in olfaction. We therefore build on the shared insight above with OlfaGraph, which uses experimentally confirmed responses and non-responses as graph relations to produce olfaction-adapted receptor embeddings, without requiring additional task-specific data sources.

# 3 METHODOLOGY

We explore **odorant–receptor response prediction**, i.e., whether a given odorant activates a given olfactory receptor and, when applicable, how strongly. Existing approaches encode each molecule and each receptor with off-the-shelf pre-trained models and learn their interaction from labeled pairs in isolation. However, the binding profile of a receptor across many odorants is itself informative about that receptor: two receptors that respond to similar sets of molecules likely share functional properties that no sequence-only encoder captures. We make this relational evidence explicit. Our method, OlfaGraph, proceeds in three stages. First, we organize the experimentally measured responses into a signed bipartite graph over odorants and receptors (Section 3.1). Second, we train a graph neural network on this graph to produce receptor embeddings that encode observed binding evidence (Section 3.2). Third, we use the adapted embeddings in a downstream predictor that supports both previously observed and entirely novel odorants (Section 3.3).

**Task formulation.** Each training example is a pair $(m, r, y)$ consisting of an odorant $m$, an olfactory receptor $r$, and an experimentally measured target $y$. When $y$ is binary, it indicates whether the receptor responds to the odorant; when $y$ is continuous, it quantifies the response strength. The task is therefore classification in the binary case and regression in the continuous case.

## 3.1 OLFACTORY BINDING GRAPH

**Biological motivation.** Olfactory coding is combinatorial: each odorant activates a subset of receptors, and each receptor responds to many odorants (Buck, 2004). The resulting response matrix is naturally a bipartite structure. We encode it as a graph so that a message passing step can propagate binding evidence between receptors that share odorant partners.

**Odorant-receptor graph.** We define a signed bipartite graph $\mathcal{G} = (\mathcal{M} \cup \mathcal{R},\ \mathcal{E}^+,\ \mathcal{E}^-)$, where $\mathcal{M}$ is the set of odorant nodes and $\mathcal{R}$ the set of receptor nodes. An edge $(m, r) \in \mathcal{E}^+$ indicates that receptor $r$ responds to odorant $m$; an edge $(m, r) \in \mathcal{E}^-$ indicates a confirmed non-response. For binary targets the mapping is direct: $y = 1$ yields a positive edge, $y = 0$ a negative edge. For standardised continuous targets, we assign a positive edge when $y > 0$ and a negative edge when $y \leq 0$, where zero corresponds to the dataset mean after standardisation. Pairs without an experimental measurement contribute no edge.

**Edge construction.** For binary-response data, to let receptors exchange information through shared odorants, we select odorants that collectively cover many receptor pairs: two receptors are covered together when both are tested with the same selected odorant, regardless of their response values. We greedily add the odorant covering the most previously uncovered receptor pairs until the graph contains $K$ selected odorants that connect receptors. We construct edges exclusively from training-set measurements. This restriction is deliberate: it ensures that the receptor embedding $\mathbf{z}_{\text{prot}}$ can be computed before any test molecule is observed, which is a prerequisite for the cold-molecule setting. Odorant selection restricts only the graph topology used for message passing. All training pairs, including those whose odorant was not selected, retain their original targets and contribute to the training loss. For continuous-response data, we retain all odorants with training measurements. Further details on graph construction and design choices are provided in Appendix C.

## 3.2 GRAPH NEURAL NETWORK FOR OLFACTION-SPECIFIC PROTEIN REPRESENTATIONS

We learn olfaction-specific receptor representations by adapting frozen pre-trained protein embeddings $\mathbf{x}_{\text{prot}}$, using response-profile evidence alongside pre-trained molecular embeddings $\mathbf{x}_{\text{mol}}$. Unless stated otherwise, receptor nodes are initialized with ESM3 (Hayes et al., 2025) protein em-

beddings, and odorant nodes with ChemBERTa (Chithrananda et al., 2020) molecular embeddings computed from SMILES strings. Ablations of alternative protein and molecular representations are provided in Section 5.1 and Appendix D, respectively.

**Input projection.** For a node $v$ of type $t \in \{\mathrm{mol}, \mathrm{prot}\}$, we project the input embedding into a shared $d$-dimensional hidden space:

$$\mathbf{h}_v^{(0)} = \sigma\big(\mathbf{P}_t\, \mathbf{x}_v + \mathbf{c}_t\big), \tag{1}$$

where $\mathbf{P}_t$ and $\mathbf{c}_t$ are trainable, type-specific parameters that map the differing input dimensions to $d$, and $\sigma$ is the LeakyReLU nonlinearity.

**Signed message passing.** At layer $\ell \in \{1, 2\}$, we compute separate GraphSAGE updates for each sign $s \in \{+, -\}$, corresponding to the response relations $\mathcal{E}^+$ and $\mathcal{E}^-$:

$$\mathbf{a}_{v,s}^{(\ell)} = \mathbf{W}_{t,s,\mathrm{self}}^{(\ell)}\, \mathbf{h}_v^{(\ell-1)} \;+\; \mathbf{W}_{t,s,\mathrm{nbr}}^{(\ell)}\, \frac{1}{|\mathcal{N}_s(v)|} \sum_{u \in \mathcal{N}_s(v)} \mathbf{h}_u^{(\ell-1)} \;+\; \mathbf{b}_{t,s}^{(\ell)}, \tag{2}$$

where $\mathcal{N}_s(v)$ is the set of neighbors connected to $v$ through edges of type $s$, and $\mathrm{self}$ and $\mathrm{nbr}$ denote self-feature and mean-neighbor transformations, respectively. An empty neighborhood contributes zero. The trainable parameters $\mathbf{W}_{t,s,\mathrm{self}}^{(\ell)}$, $\mathbf{W}_{t,s,\mathrm{nbr}}^{(\ell)}$, and $\mathbf{b}_{t,s}^{(\ell)}$, are independent across layers $\ell$, edge types $s$, and receiving node types $t$, distinguishing $\mathrm{mol} \to \mathrm{prot}$ from $\mathrm{prot} \to \mathrm{mol}$ propagation.

**Combining edge types.** We merge the positive and negative updates by subtraction:

$$\mathbf{h}_v^{(1)} = \sigma\big(\mathbf{a}_{v,+}^{(1)} - \mathbf{a}_{v,-}^{(1)}\big), \qquad \mathbf{h}_v^{(2)} = \mathbf{a}_{v,+}^{(2)} - \mathbf{a}_{v,-}^{(2)}. \tag{3}$$

Subtraction provides a biologically motivated contrast between learned response relations, reflecting that receptor selectivity is characterized by both odorants that activate it and those that elicit no response. The first layer applies LeakyReLU after subtraction, while the second retains the signed difference without activation, and in both layers the combined representation is $L_2$-normalized. We denote the encoder's output protein and molecule embeddings by $\mathbf{z}_{\mathrm{prot}} = \mathbf{h}_{\mathrm{prot}}^{(2)}$ and $\mathbf{z}_{\mathrm{mol}} = \mathbf{h}_{\mathrm{mol}}^{(2)}$, respectively.

**Two-hop receptive field.** Two GraphSAGE layers give each receptor a receptive field that spans $\mathrm{prot} \to \mathrm{mol} \to \mathrm{prot}$ paths. After one hop, a receptor aggregates features from its directly connected odorants. After the second hop, it additionally receives information from other receptors that share those odorants, capturing receptor–receptor similarity induced by overlapping binding profiles. We fix the depth at two because one hop cannot propagate receptor-level information and deeper layers risk over-smoothing on graphs of the considered sizes.

**Training objective.** During graph representation learning, a three-layer MLP head maps the concatenation $[\mathbf{z}_{\mathrm{mol}} \| \mathbf{z}_{\mathrm{prot}}]$ to a scalar prediction. We train the graph encoder and this prediction head jointly over all training pairs, using binary cross-entropy for binary targets and mean squared error for continuous targets. After training, we discard the MLP head and use the learned protein embeddings as features for the downstream predictor described next.

### 3.3 Downstream prediction

After graph training, we freeze the response-aware receptor embeddings $\mathbf{z}_{\mathrm{prot}}$ and construct the feature vector for each query odorant-receptor pair:

$$\mathbf{f}_{\mathrm{prot,mol}} = \big[\mathbf{x}_{\mathrm{prot}} \,\|\, \mathbf{x}_{\mathrm{mol}} \,\|\, \mathbf{z}_{\mathrm{prot}}\big]. \tag{4}$$

We retain both the original protein embedding $\mathbf{x}_{\mathrm{prot}}$ and the graph-adapted embedding $\mathbf{z}_{\mathrm{prot}}$ because the former preserves general sequence information that may be diminished during graph adaptation. We use a gradient-boosted tree model, XGBoost Chen & Guestrin (2016), to map $\mathbf{f}_{\mathrm{prot,mol}}$ to the final odorant-receptor response prediction.

### 3.4 Evaluation setup

We assume a fixed set of known receptors and evaluate generalization across odorants in two settings: *seen molecules* and *cold molecules*.

**Seen molecules.** In the *seen-molecule setting*, we hold out individual receptor-odorant measurements, allowing the same odorant to appear in both training and evaluation data. The task is to predict held-out responses from partially observed odorant-receptor interactions.

**Cold molecules.** In the *cold-molecule setting*, the query odorant has no measured responses in the training set. Equation (4) supports this setting directly: a new odorant requires only its pre-trained embedding $\mathbf{x}_{\mathrm{mol}}$, which is combined with the the readily available receptor features $\mathbf{x}_{\mathrm{prot}}$ and $\mathbf{z}_{\mathrm{prot}}$. The graph-adapted receptor embedding $\mathbf{z}_{\mathrm{prot}}$ is learned from training measurements without using any responses to the query odorant. No retraining or graph update is needed at inference time.

# 4 Experiments and Results

## 4.1 Datasets

We evaluate on three datasets of experimentally measured odorant-receptor responses, using the benchmark versions provided by LORAX McConachie et al. (2026). These datasets span different species and measurement densities: **M2OR** provides binary labels for a partially measured panel of mammalian receptor-odorant pairs, while **Mosquito** and **Fly** provide continuous responses for fully measured panels of insects' receptors. For the insect datasets, we use the standardized continuous targets and formulate prediction as regression. Standardization uses the mean and standard deviation of each entire dataset, so zero represents the dataset's mean response.

**M2OR.** M2OR Lalis et al. (2024) collects receptor responses from studies using different testing procedures and odorant concentrations, providing binary labels for responsive and non-responsive receptor–odorant pairs. Following LORAX, we use a benchmark subset of 46,563 response records across 1,237 receptor sequences and 596 odorants, and formulate prediction as a classification task. Unmeasured pairs remain unlabeled and are excluded from supervision.

**Mosquito.** The dataset Carey et al. (2010) tests 50 functional mosquito (*Anopheles gambiae*) receptors against 110 odorants, resulting in a complete panel of 5,500 receptor-odorant pairs. Each receptor is expressed in a fruit-fly nerve cell lacking its own odorant receptor. Responses measure the average change in electrical firing rate during odorant stimulation, in spikes per second. Positive and negative responses indicate increases and decreases in nerve-cell firing relative to baseline, respectively.

**Fly.** The authors Hallem & Carlson (2006) provide a complete response matrix for 24 fruit-fly (*Drosophila melanogaster*) receptors and 110 odorants, comprising 2,640 receptor-odorant pairs. Electrical recordings measure responses in spikes per second and capture both increases and decreases in nerve-cell firing.

## 4.2 Baselines

We compare our approach against a set of *interaction baselines*, all trained to predict odorant-receptor response by explicitly modeling the interaction between each molecule-protein pair.

**Hladiš.** This approach Hladiš et al. (2023) incorporates a projected protein embedding into each atom's features and combines bond-based message passing with attention across atoms. Learned attention weights pool the atom representations into a joint protein-molecule vector.

**MolOR.** We use the receptor response prediction component of MolOR Chithrananda et al. (2026), which combines atom embeddings from a graph convolutional network trained from scratch with pre-trained per-residue protein embeddings. Bidirectional cross-attention generates weights for pooling atoms and residues separately, and the pooled vectors are concatenated to form the interaction representation.

**ProSmith.** ProSmith Kroll et al. (2024) uses a multimodal Transformer to jointly process uncompressed, per-token molecular and per-residue protein embeddings, with the final `[CLS]` token encoding their interaction. We use this `[CLS]` token as the interaction representation, omitting the original three-model boosting ensemble over various representation combinations to isolate the effectiveness of the learned interaction representation itself.

**LORAX.** LORAX McConachie et al. (2026) fine-tunes the ChemBERTa molecular encoder using low-rank adaptation (LoRA) for receptor-response prediction while keeping the protein encoder frozen. Bidirectional cross-attention exchanges information between molecular token embeddings and per-residue protein embeddings, and the updated embeddings are mean-pooled separately and concatenated into an interaction representation.

For all baselines described above, we take the final pooled or joint vector as the interaction representation between the given molecule and protein, and denote this vector $\mathrm{cls}$. To ensure a fair comparison across methods, we replace each baseline's original protein encoder (ProtBERT, ESM-2, or otherwise) with the same frozen ESM3 embeddings, isolating the effect of the interaction-modeling architecture from differences in the underlying protein representation.

Along with the aforementioned baselines, we consider **XGBoost-base**, a simple non-interaction baseline that relies solely on the predictive power of concatenated, independently-derived representations. In its setting, we concatenate ChemBERTa molecule embedding with ESM3 protein embedding and apply an XGBoost Chen & Guestrin (2016) model to the resulting vector to predict odorant-receptor response.

### 4.3 Experimental Setup

**Implementation details.** OlfaGraph uses two GraphSAGE layers with a hidden dimension of 256 and a prediction MLP with hidden widths 256 and 128 and dropout 0.3. We train for 900 epochs with Adam (learning rate $3 \times 10^{-3}$, weight decay $10^{-4}$), clipping the gradient norm at 1.0. The downstream XGBoost predictor uses 400 trees, maximum depth 6, learning rate 0.1, and row and feature subsampling ratios of 0.8.

**Model size and computational resources.** OlfaGraph uses approximately $1.6$M trainable parameters. This is comparable to MolOR ($1.6$M) and Hladiš ($1.4$M), and smaller than LORAX ($6.7$M) and ProSmith ($43.8$M). We train OlfaGraph on a single NVIDIA A100 (80 GB).

**Data split.** We use five train–validation–test partitions, splitting by measured pairs for seen molecules and by molecules for cold molecules. We report means $\pm$ 95% confidence intervals across the five test sets.

**Metrics.** We use AUROC, AUPRC, MCC, and F1 for classification, with MCC accounting for both classes under label imbalance Chicco & Jurman (2020). For regression, we use $R^2$, RMSE, Pearson, and Spearman correlations to capture explained variance, prediction error, and response ordering.

### 4.4 Results

We evaluate OlfaGraph's response-aware protein embeddings against representations learned by molecule-protein interaction models and off-the-shelf pretrained protein embeddings for odorant-receptor response prediction. Table 1 reports results in the seen-molecule and cold-molecule settings across three datasets. Interaction baselines augment the original ESM3 protein and ChemBERTa molecular embeddings with their learned interaction representation, yielding $[\mathbf{x}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}} \| \mathrm{cls}]$, while OlfaGraph adds its response-aware protein embedding, resulting in $[\mathbf{x}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}} \| \mathbf{z}_{\mathrm{prot}}]$. For every feature configuration, we train an XGBoost classifier or regressor on the concatenated embeddings; XGBoost-base denotes the configuration that uses only the original protein and molecular embeddings, $[\mathbf{x}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}}]$. For a more detailed analysis of the contribution of each feature component, see Appendix A. We base our comparative discussion on aggregate ranks to summarize performance across metrics with different scales and across evaluation settings. The results support the following findings:

**OlfaGraph ranks among the top-performing methods most consistently across datasets and settings.** OlfaGraph ranks first or second in five of six dataset–setting combinations, more often than any baseline. Its largest average-rank advantages over the next-best method occur for seen molecules on Mosquito ($1.55$ versus $2.75$) and Fly ($1.45$ versus $2.55$), where it also leads all four metrics. Averaging the reported ranks across all six combinations gives OlfaGraph the best overall rank of $2.35$, followed by XGBoost-base at $2.96$. This consistency spans mammalian and insect datasets, binary and continuous targets, and partially and fully measured response panels.

Table 1: Odorant-receptor response prediction on M2OR, Mosquito, and Fly in the seen-molecule and cold-molecule settings. All methods use XGBoost over concatenated features: $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}}]$ for XGBoost-base, combining original ESM3 protein and ChemBERTa molecular embeddings; $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}} \| \text{cls}]$ for interaction baselines, adding the learned interaction representation; and $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}} \| \mathbf{z}_{\text{prot}}]$ for OlfaGraph, adding the response-aware protein embedding. **Bold** and underline mark the best and second-best results per dataset and setting.

| Dataset | Method | Seen molecules | | | | | Cold molecules | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC↑ | AUPRC↑ | MCC↑ | F1↑ | Rank↓ | AUROC↑ | AUPRC↑ | MCC↑ | F1↑ | Rank↓ |
| M2OR | LORAX | 0.891±0.008 | 0.753±0.033 | 0.629±0.037 | 0.712±0.029 | 3.80 | 0.829±0.021 | 0.665±0.059 | 0.523±0.049 | 0.614±0.047 | 4.25 |
| | ProSmith | 0.877±0.011 | 0.742±0.035 | 0.617±0.032 | 0.704±0.025 | 5.25 | 0.826±0.032 | **0.675±0.056** | 0.556±0.041 | 0.656±0.042 | 2.00 |
| | MolOR | 0.891±0.010 | 0.761±0.030 | 0.630±0.026 | 0.711±0.019 | 3.70 | **0.837±0.016** | 0.667±0.057 | 0.526±0.024 | 0.629±0.033 | 3.40 |
| | Hladiš | **0.902±0.008** | **0.780±0.032** | **0.654±0.025** | **0.731±0.020** | **1.20** | 0.815±0.033 | 0.666±0.059 | 0.509±0.061 | 0.605±0.059 | 5.25 |
| | XGBoost-base | 0.896±0.012 | 0.759±0.032 | 0.637±0.026 | 0.719±0.021 | 2.80 | 0.820±0.035 | 0.661±0.056 | 0.527±0.048 | 0.618±0.049 | 4.25 |
| | **OlfaGraph (ours)** | 0.889±0.011 | 0.746±0.033 | 0.629±0.029 | 0.712±0.023 | 4.25 | 0.832±0.030 | 0.668±0.054 | **0.562±0.045** | **0.657±0.045** | **1.85** |
| | | $R^2$↑ | RMSE↓ | Pearson↑ | Spearman↑ | Rank↓ | $R^2$↑ | RMSE↓ | Pearson↑ | Spearman↑ | Rank↓ |
| Mosquito | LORAX | 0.612±0.031 | 0.635±0.029 | 0.784±0.021 | 0.772±0.017 | 4.75 | 0.451±0.061 | 0.741±0.056 | 0.681±0.043 | 0.723±0.012 | 3.50 |
| | ProSmith | 0.616±0.031 | 0.631±0.019 | 0.786±0.020 | 0.776±0.019 | 4.80 | 0.447±0.078 | 0.743±0.068 | 0.680±0.057 | **0.725±0.026** | 3.30 |
| | MolOR | 0.642±0.031 | 0.610±0.039 | 0.801±0.019 | 0.772±0.012 | 3.95 | 0.359±0.151 | 0.796±0.087 | 0.658±0.043 | 0.687±0.057 | 4.35 |
| | Hladiš | 0.656±0.028 | 0.598±0.030 | 0.811±0.018 | 0.779±0.018 | 2.75 | 0.378±0.134 | 0.786±0.099 | 0.646±0.089 | 0.670±0.039 | 4.75 |
| | XGBoost-base | 0.656±0.021 | 0.598±0.035 | 0.813±0.013 | 0.770±0.017 | 3.20 | **0.483±0.074** | **0.718±0.066** | **0.700±0.054** | 0.717±0.017 | **2.10** |
| | **OlfaGraph (ours)** | **0.664±0.020** | **0.591±0.035** | **0.817±0.012** | **0.781±0.017** | **1.55** | 0.471±0.084 | 0.726±0.072 | 0.689±0.064 | 0.711±0.023 | 3.00 |
| Fly | LORAX | 0.505±0.048 | 0.721±0.038 | 0.714±0.031 | 0.788±0.013 | 5.40 | 0.436±0.063 | 0.748±0.030 | 0.676±0.041 | **0.760±0.021** | 3.60 |
| | ProSmith | 0.511±0.037 | 0.717±0.032 | 0.719±0.025 | 0.791±0.012 | 4.80 | 0.462±0.111 | 0.728±0.067 | 0.689±0.072 | 0.750±0.019 | 3.20 |
| | MolOR | 0.548±0.060 | 0.688±0.054 | 0.742±0.039 | 0.814±0.025 | 3.05 | 0.385±0.142 | 0.778±0.078 | 0.652±0.073 | 0.715±0.034 | 4.60 |
| | Hladiš | 0.546±0.063 | 0.690±0.056 | 0.740±0.043 | 0.811±0.011 | 3.75 | 0.416±0.086 | 0.761±0.053 | 0.659±0.050 | 0.710±0.039 | 4.75 |
| | XGBoost-base | 0.561±0.052 | 0.679±0.049 | 0.749±0.035 | 0.811±0.010 | 2.55 | 0.484±0.060 | 0.715±0.030 | 0.699±0.040 | 0.750±0.016 | 2.85 |
| | **OlfaGraph (ours)** | **0.563±0.052** | **0.677±0.049** | **0.751±0.035** | **0.816±0.010** | **1.45** | **0.492±0.062** | **0.710±0.033** | **0.704±0.042** | 0.754±0.014 | **2.00** |

**Complex interaction models offer limited gains over off-the-shelf pre-trained features.** XGBoost-base outperforms each interaction-based baseline by average rank in at least four of the six settings. This raises the question of whether complex interaction modeling at the atom and residue level is necessary when pooled pre-trained features achieve comparable predictive performance.

**Olfaction-specific protein adaptation improves upon general-purpose pre-trained representations.** OlfaGraph improves all four reported metrics over XGBoost-base, which uses the original, unadapted protein embeddings, in four of six settings. In contrast, adding learned interaction representations improves the average rank over these pre-trained features in only one or two settings per baseline. These results highlight the value of augmenting generic pre-trained features with protein representations informed by measured olfactory response profiles.

**Response-aware protein embeddings improve cold-molecule prediction.** OlfaGraph achieves the best average rank on two of three cold-molecule benchmarks. On M2OR, it improves MCC from $0.527$ to $0.562$ ($+6.6\%$) and F1 from $0.618$ to $0.657$ ($+6.3\%$) over XGBoost-base. ProSmith achieves similar M2OR performance, but uses $43.8$M parameters compared with OlfaGraph's $1.6$M, highlighting OlfaGraph's parameter efficiency in the cold-molecule setting.

## 5 Ablation Study

### 5.1 Comparison of protein representations

Discovering the stronger average performance of off-the-shelf pre-trained protein features compared with learned interaction representations in Section 4.4, we focus on the isolated effect of protein representation on odorant-receptor response prediction in Table 2. We concatenate each protein representation with the original molecular embedding and train an XGBoost predictor on $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}}]$, replacing $\mathbf{x}_{\text{prot}}$ with $\mathbf{z}_{\text{prot}}$ for OlfaGraph. Then, we compare four categories of protein representations: (i) graph-learned embeddings from OlfaGraph with ESM3, random-vector, or ProtT5 node initialization; (ii) pretrained protein language model embeddings from ESM3, ESM-1b, and ProtT5; (iii) one-hot receptor identities; and (iv) sequence-derived descriptors, comprising amino-acid and dipeptide frequencies (AAC and kmer2) Chen et al. (2018), composition–transition–distribution features (CTD) Dubchak et al. (1995), pseudo-amino-acid composition (PseAAC) Chou (2001), mean-pooled BLOSUM62 substitution scores Henikoff & Henikoff (1992), and means and standard deviations of seven physicochemical amino-acid properites (AAindex) Kawashima et al.

Table 2: Comparison of receptor representations: (i) OlfaGraph embeddings with different node initializations, (ii) pre-trained protein language model embeddings, (iii) one-hot receptor encoding, (iv) sequence descriptors, and (v) receptor-only and molecule-only controls. Except for these controls, XGBoost predictor uses $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}}]$, with $\mathbf{z}_{\text{prot}}$ replacing $\mathbf{x}_{\text{prot}}$ for OlfaGraph. Seen and Cold denote the seen-molecule and cold-molecule settings. Dim. denotes the receptor representation dimension; # recept. means the number of receptors. Mean rank averages ranks across the six dataset-setting combinations. **Bold** and underline mark the best and second-best results per dataset and setting.

| | | | **M2OR** (AUROC↑) | | **Mosquito** ($R^2$ ↑) | | **Fly** ($R^2$ ↑) | |
|---|---|---|---|---|---|---|---|---|
| **Receptor representation** | **Dim.** | **Mean rank↓** | Seen | Cold | Seen | Cold | Seen | Cold |
| OlfaGraph (ESM3 initialization) | 256 | 6.00 | 0.876±0.024 | 0.831±0.032 | 0.665±0.029 | 0.473±0.092 | 0.561±0.064 | **0.495±0.079** |
| OlfaGraph (Random initialization) | 256 | 7.50 | 0.840±0.007 | 0.799±0.033 | 0.667±0.011 | 0.476±0.098 | 0.562±0.067 | 0.491±0.079 |
| OlfaGraph (ProtT5 initialization) | 256 | **4.33** | 0.877±0.014 | **0.842±0.024** | **0.682±0.017** | 0.478±0.101 | **0.572±0.064** | 0.493±0.077 |
| ESM3 | 1536 | 5.33 | **0.896±0.015** | 0.820±0.043 | 0.656±0.026 | 0.483±0.092 | 0.561±0.064 | 0.484±0.074 |
| ESM-1b | 1280 | 6.00 | 0.894±0.017 | 0.817±0.041 | 0.656±0.014 | 0.486±0.105 | 0.560±0.067 | 0.484±0.071 |
| ProtT5 | 1024 | 5.92 | **0.896±0.016** | 0.818±0.038 | 0.657±0.033 | 0.480±0.102 | 0.561±0.063 | 0.483±0.075 |
| One-hot | # recept. | 8.50 | 0.794±0.017 | 0.778±0.034 | 0.544±0.043 | **0.504±0.102** | 0.564±0.048 | 0.485±0.074 |
| kmer2 | 400 | 6.33 | 0.891±0.014 | 0.810±0.039 | 0.645±0.022 | 0.495±0.105 | 0.552±0.061 | 0.487±0.077 |
| CTD | 147 | 5.67 | 0.890±0.014 | 0.811±0.037 | 0.651±0.018 | 0.502±0.112 | 0.566±0.068 | 0.482±0.077 |
| PseAAC | 25 | 8.25 | 0.888±0.016 | 0.806±0.048 | 0.617±0.020 | 0.483±0.102 | 0.546±0.064 | 0.486±0.081 |
| BLOSUM | 20 | 8.33 | 0.884±0.019 | 0.802±0.045 | 0.604±0.032 | 0.481±0.096 | 0.555±0.053 | 0.491±0.073 |
| AAC | 20 | 9.42 | 0.885±0.017 | 0.799±0.049 | 0.607±0.021 | 0.479±0.098 | 0.543±0.065 | 0.489±0.081 |
| AAindex | 14 | 9.42 | 0.884±0.016 | 0.808±0.044 | 0.588±0.031 | 0.489±0.101 | 0.548±0.055 | 0.481±0.074 |
| Receptor one-hot only | # recept. | 14.00 | 0.777±0.019 | 0.772±0.031 | 0.205±0.024 | 0.225±0.040 | 0.259±0.040 | 0.267±0.056 |
| Molecular features only | – | 15.00 | 0.651±0.025 | 0.543±0.054 | 0.075±0.014 | 0.032±0.026 | 0.097±0.022 | 0.060±0.030 |

(2007). Protein-one-hot-only and molecular-features-only controls assess prediction from either input alone. We draw the following conclusions:

**OlfaGraph learns useful receptor representations on the basis of response relations.** Using response relations as the sole source of receptor-specific information, OlfaGraph with fixed random-vector initialization improves M2OR AUROC over one-hot receptor features from $0.794$ to $0.840$ ($+5.8\%$) for seen molecules and from $0.778$ to $0.799$ ($+2.7\%$) for cold molecules. The gain is particularly pronounced on Mosquito with seen molecules, where $R^2$ rises from $0.544$ to $0.667$ ($+22.6\%$), suggesting that OlfaGraph captures olfaction-relevant receptor characteristics using only response relations and the features of connected odorants, without access to protein sequences

**OlfaGraph initialized with pre-trained protein embeddings achieves the best overall rank.** OlfaGraph initialized with ProtT5 ranks first among all evaluated representations, with a mean rank of $4.33$, compared with $5.92$ for the original ProtT5 embeddings. This improvement includes an AUROC increase from $0.818$ to $0.842$ ($+2.9\%$) on M2OR with cold molecules and an $R^2$ increase from $0.657$ to $0.682$ ($+3.8\%$) on Mosquito with seen molecules. On M2OR, pre-trained protein language model features improve OlfaGraph over random initialization in both seen- and cold-molecule settings. For further analysis of the relative contributions of response relations and pre-trained protein embeddings, see Appendix B. Together, these results support ProtT5 as a promising initialization for olfaction-specific graph learning and suggest that protein information from pretrained models complements response relations of mammalian receptors in M2OR.

**Simple receptor features match pre-trained protein embeddings.** On the insect datasets, sequence descriptors and one-hot receptor identities achieve performance comparable to pre-trained protein language model embeddings. For instance, CTD matches or exceeds all three pre-trained alternatives. These results show that the complexity of pre-trained protein language model embeddings does not guarantee more accurate receptor-response prediction than simple receptor features.

### 5.2 OLFAGRAPH DESIGN CHOICES

We ablate OlfaGraph's *message-passing design* and compare alternative *graph encoders* to assess their contributions to prediction performance (Table 3).

**Message-passing design.** We ablate OlfaGraph's two-layer signed GraphSAGE encoder. *One message-passing layer* retains direct molecule-receptor interactions but removes receptor-receptor

Table 3: Ablation of OlfaGraph's message-passing design (upper block) and comparison of graph encoders (lower block). XGBoost predicts responses from the feature vector $[\mathbf{z}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}}]$. Mean ranks are computed within each block; **bold** marks the best mean per column in that block.

| | | **M2OR** (AUROC↑) | | **Mosquito** ($R^2$ ↑) | | **Fly** ($R^2$ ↑) | |
|---|---|---|---|---|---|---|---|
| **Architecture** | **Mean rank↓** | Seen | Cold | Seen | Cold | Seen | Cold |
| OlfaGraph (full model) | **2.17** | 0.876±0.024 | **0.831±0.032** | 0.665±0.029 | 0.473±0.092 | **0.561±0.064** | **0.495±0.079** |
| One message-passing layer | 3.50 | 0.891±0.012 | 0.808±0.037 | 0.657±0.035 | 0.476±0.099 | 0.558±0.066 | 0.491±0.076 |
| Positive-edge message passing | 2.50 | 0.864±0.011 | 0.815±0.032 | **0.673±0.024** | 0.480±0.098 | 0.559±0.068 | 0.493±0.083 |
| Unsigned message passing | 3.58 | 0.874±0.017 | 0.812±0.034 | 0.595±0.045 | **0.484±0.103** | 0.539±0.071 | 0.492±0.078 |
| No message passing | 3.25 | **0.892±0.013** | 0.812±0.043 | 0.651±0.030 | 0.483±0.097 | 0.552±0.061 | 0.490±0.072 |
| GraphSAGE (OlfaGraph default) | **1.58** | **0.876±0.024** | **0.831±0.032** | 0.665±0.029 | 0.473±0.092 | **0.561±0.064** | **0.495±0.079** |
| GAT | 3.67 | 0.799±0.008 | 0.789±0.053 | 0.655±0.034 | 0.471±0.092 | 0.548±0.067 | 0.490±0.072 |
| GraphConv | 2.00 | **0.876±0.021** | 0.830±0.034 | **0.668±0.035** | 0.473±0.085 | 0.557±0.072 | 0.493±0.080 |
| GIN | 2.75 | 0.856±0.015 | **0.831±0.039** | 0.659±0.039 | **0.477±0.088** | 0.546±0.061 | 0.488±0.081 |

exchange through shared odorants. *Positive-edge message passing* retains two layers but removes negative edges: non-responses in binary datasets and responses at or below the dataset mean in continuous datasets. *Unsigned message passing* retains all edges as a single relation, removing sign-specific transformations and their subtractive combination. *No message passing* retains OlfaGraph's trainable input projections and feeds the separately projected protein and molecular embeddings directly to the same prediction MLP, without graph-based exchange. All variants use the same training pairs, targets, loss, and downstream evaluation protocol; removing negative edges does not remove their pairs from supervision.

*OlfaGraph benefits from two-hop signed message passing.* Table 3 supports the importance of two-hop propagation and signed relations: the full encoder achieves the best mean rank (2.17), compared with one-layer (3.50) and unsigned (3.58) variants. Positive-edge propagation remains competitive (2.50), but retaining negative edges improves the overall ranking; removing message passing also worsens the mean rank (3.25), highlighting the value of explicitly incorporating the relational structure of measured odorant–receptor responses. Together, these results support the complete OlfaGraph design, with each component contributing to its aggregate performance.

**Graph encoder.** We also evaluate the choice of graph encoder by comparing GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2017), GraphConv (Morris et al., 2019), and GIN (Xu et al., 2018) in Table 3. We can conclude that *GraphSAGE is the strongest encoder for OlfaGraph.* The default GraphSAGE implementation achieves the best reported average rank (1.5), while GAT ranks lowest (3.7), with notably lower AUROC on M2OR. GraphConv and GIN remain competitive, but neither improves the overall ranking over GraphSAGE. These results support GraphSAGE as a strong default choice for OlfaGraph and suggest that attention-based aggregation hinders effective representation learning.

## 6 Conclusion

Predicting how olfactory receptors respond to odorant molecules is a central problem in computational olfaction, and one we address in this work. To this end, we introduce OlfaGraph, an approach that learns olfaction-specific receptor representations from a signed graph of measured odorant–receptor responses. We find that general-purpose protein language model embeddings offer no benefit over simple protein descriptors, motivating adapting these widely-used, typically frozen, generic representations to the olfactory domain. OlfaGraph does so by incorporating response evidence directly into receptor representations, achieving the best overall ranking across the evaluated mammalian and insect datasets. Importantly, its design natively supports prediction for odorants entirely unseen during training, with leading performance in most cold-molecule settings. Our findings further show that response relations alone can produce informative receptor embeddings, even without protein sequence information. For mammalian receptors, in turn, the response graph complements rather than replaces pre-trained sequence embeddings. Thus, our results establish that accounting for measured response relations is key to producing effective, olfaction-specific protein representations that complement general-purpose ones and improve generalization to unseen odorants.

## AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing, LaTeX table formatting, and providing feedback on experimental design. We also used generative AI tools to assist with code development. The authors reviewed the AI-assisted content and take full responsibility for the final manuscript, including its claims, results, and references.

## REFERENCES

Christian B Billesbølle, Claire A de March, Wijnand JC van der Velden, Ning Ma, Jeevan Tewari, Claudia Llinas Del Torrent, Linus Li, Bryan Faust, Nagarajan Vaidehi, Hiroaki Matsunami, et al. Structural basis of odorant recognition by a human odorant receptor. *Nature*, 615(7953):742–749, 2023.

Linda B Buck. Olfactory receptors and odor coding in mammals. *Nutrition reviews*, 62(suppl_3): S184–S188, 2004.

Allison F Carey, Guirong Wang, Chih-Ying Su, Laurence J Zwiebel, and John R Carlson. Odorant reception in the malaria mosquito anopheles gambiae. *Nature*, 464(7285):66–71, 2010.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14):2499–2502, 2018.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

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. *Cell Systems*, 17(9):101711, 2026. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2026.101711.

Chulwon Choi, Jungnam Bae, Seonghan Kim, Seho Lee, Hyunook Kang, Jinuk Kim, Injin Bang, Kiheon Kim, Won-Ki Huh, Chaok Seok, et al. Understanding the molecular mechanisms of odorant binding and activation of the human or52 family. *Nature Communications*, 14(1):8105, 2023.

Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

Inna Dubchak, Ilya Muchnik, Stephen R Holbrook, and Sung-Hou Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, 92(19):8700–8704, 1995.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Elissa A Hallem and John R Carlson. Coding of odors by a receptor repertoire. *Cell*, 125(1): 143–160, 2006.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the national academy of sciences*, 89(22):10915–10919, 1992.

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.

Yueying Jing, Dongxue Zhang, and Limin Li. H2gnndti: hierarchical heterogeneous graph neural networks for drug–target interaction prediction. *Bioinformatics*, 41(4):btaf117, 2025.

Masha Karelina, Joseph J Noh, and Ron O Dror. How accurately can one predict drug binding modes using alphafold models? *Elife*, 12:RP89386, 2023.

Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, 36(suppl_1):D202–D205, 2007.

Andreas Keller and Leslie B Vosshall. Human olfactory psychophysics. *Current Biology*, 14(20): R875–R878, 2004.

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.

Alexander Kroll, Sahasra Ranjan, and Martin J Lercher. A multimodal transformer network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships. *PLoS computational biology*, 20(5):e1012100, 2024.

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.

Asifa Majid. Human olfaction at the intersection of language, culture, and biology. *Trends in Cognitive Sciences*, 25(2):111–123, 2021.

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.

Grant D. McConachie, Emily Duniec, Florence V. Guerina, Meg Younger, and Brian DePasquale. Low rank adaptation of chemical foundation models generate effective odorant representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4602–4609, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Yanyi Su, Hongshuai Wang, Zhifeng Gao, and Jun Cheng. Nose: neural olfactory-semantic embedding with tri-modal orthogonal contrastive learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19615–19647, 2026.

Gary Tom, Cher Tian Ser, Ella Miray Rajaonson, Stanley Lo, Hyun Suk Park, Brian Lee, and Benjamin Manuel Sanchez. Does this smell the same? learning representations of olfactory mixtures using inductive biases. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.

Uxía Veleiro, Jesús de la Fuente, Guillermo Serrano, Marija Pizurica, Mikel Casals, Antonio Pineda-Lucena, Silve Vicent, Idoia Ochoa, Olivier Gevaert, and Mikel Hernaez. Gennius: an ultrafast drug–target interaction inference method based on graph neural networks. *Bioinformatics*, 40(1): btad774, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Fangping Wan, Lixiang Hong, An Xiao, Tao Jiang, and Jianyang Zeng. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, 35(1):104–111, 2019.

Yujing Wang, Jun Zhan, Fan Lü, Hua Zhang, Junjie Qiu, Bin Xu, and Pinjing He. Machine learning–driven mapping of the odorant–olfactory receptor–olfactory perception network: A comprehensive review. *Sensing and Bio-Sensing Research*, pp. 101018, 2026.

Hongxin Xie, Jiande Sun, Fanfu Xue, Zifei Han, Shanshan Feng, and Qi Chen. Multi-hierarchical fine-grained feature mapping driven by feature contribution for molecular odor prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1046.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

# Appendix

The appendix supports the ablation study of the main text. Appendix A explains which feature sets the interaction baselines and OlfaGraph are compared on, and why. Appendix B interpolates the receptor input between receptor identity and the ESM3 embedding, separating what the response-aware protein embedding owes to the measurements from what it owes to the sequence. Appendix C studies how the graph is built, i.e. which odorants carry messages. Appendix D repeats the main comparison with other molecular embeddings.

## A Feature configurations

ProSmith (Kroll et al., 2024) and LORAX (McConachie et al., 2026) read their learned interaction representation out with a second stage of gradient boosting: three XGBoost models are fitted on $[\mathrm{cls}]$, on $[\mathbf{x}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}]$ and on $[\mathbf{x}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}\|\mathrm{cls}]$ combinations, the hyperparameters of each are additionally tuned by random search on the validation set, and the final prediction is a weighted mean of the three, with the weights chosen on validation as well. We adopt neither the tuning nor the ensemble, using only XGBoost with default hyperparameters on one feature set. We assume that tuned and ensembled predictors are stronger readers, and a stronger reader extracts more from whatever representation it is given; applied to every method, they narrow the differences between representations that this paper sets out to measure, and move every method towards a common ceiling that reflects the predictor's power more than the features' one.

Out of three feature sets above, the interaction baselines are reported with the full concatenation $[\mathbf{x}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}\|\mathrm{cls}]$, and Table 4 shows why. Reduced to their learned interaction representation alone ($[\mathrm{cls}]$), all four baselines fall behind XGBoost-base on nearly every metric in both settings. Reporting the baselines in that reduced form ($[\mathrm{cls}]$) would understate them.

Table 4: Interaction baselines on M2OR using only their learned interaction representation, $[\mathrm{cls}]$, against XGBoost-base, $[\mathbf{x}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}]$. **Bold** = best in column, underline = second; ties at three decimals share the mark.

| | Seen molecules | | | | | Cold molecules | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | AUROC↑ | AUPRC↑ | MCC↑ | F1↑ | Rank↓ | AUROC↑ | AUPRC↑ | MCC↑ | F1↑ | Rank↓ |
| LORAX, [cls] | 0.873±0.012 | 0.714±0.033 | 0.587±0.024 | 0.681±0.019 | 3.20 | 0.821±0.018 | 0.626±0.064 | 0.475±0.048 | 0.592±0.048 | 2.95 |
| ProSmith, [cls] | 0.829±0.020 | 0.648±0.043 | 0.462±0.061 | 0.591±0.046 | 5.00 | 0.801±0.019 | 0.610±0.049 | 0.440±0.038 | 0.589±0.035 | 4.00 |
| MolOR, [cls] | 0.877±0.008 | 0.731±0.032 | 0.599±0.016 | 0.689±0.012 | 2.45 | **0.834±0.023** | 0.647±0.054 | 0.493±0.030 | 0.615±0.033 | 2.00 |
| Hladiš, [cls] | 0.871±0.008 | 0.720±0.029 | 0.584±0.032 | 0.674±0.024 | 3.30 | 0.768±0.058 | 0.587±0.079 | 0.405±0.091 | 0.507±0.067 | 4.35 |
| XGBoost-base | **0.896±0.012** | **0.759±0.032** | **0.637±0.026** | **0.719±0.021** | **1.05** | 0.820±0.035 | **0.661±0.056** | **0.527±0.048** | **0.618±0.049** | **1.70** |

OlfaGraph is reported in the main results in the matching shape $[\mathbf{x}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}\|\mathbf{z}_{\mathrm{prot}}]$: the same two original embeddings, plus one learned vector each. Where representations rather than complete pipelines are compared — the ablation study in the main text (Tables 2 and 3) and Appendices B, C and D — OlfaGraph is used in the reduced form $[\mathbf{z}_{\mathrm{prot}}\|\mathbf{x}_{\mathrm{mol}}]$, in which the response-aware protein embedding replaces the original one. We regard this, rather than $\mathbf{z}_{\mathrm{prot}}$ on its own, as the counterpart of a baseline's $\mathrm{cls}$. The learned interaction representation is computed from the pair, so the molecule has already entered it; $\mathbf{z}_{\mathrm{prot}}$ is computed for the receptor alone and, used by itself, would leave the predictor with no description of the molecule it is asked about. Adding the molecular embedding back restores that symmetry, while the original protein embedding stays out, so that any difference from XGBoost-base is attributable to the response-aware protein embedding.

## B Between identity and sequence

### B.1 Interpolating the receptor input

OlfaGraph initializes each receptor node with its ESM3 embedding. The response-aware protein embedding therefore combines two kinds of information: the sequence, through this initial embedding, and the measurements, through message passing along the positive and negative edges, which bring each receptor the molecular embeddings of the odorants it responds to and of those it does not. To separate the two, we interpolate the receptor input between the ESM3 embedding and a vector that identifies the receptor and carries nothing else. Each receptor $r$ is assigned a fixed random unit

vector $\mathbf{u}_r$ of the same dimension, drawn once and keyed to the receptor, so that it is the same in every split and every run, and the receptor input at dial position $\alpha \in [0, 1]$ is

$$\mathbf{x}_r(\alpha) = \alpha\, \mathbf{x}_r + (1 - \alpha)\, \mathbf{u}_r. \tag{5}$$

In the implementation both terms are centered on the training receptors, the random vectors are rescaled to the spread of the ESM3 embeddings, and the mixture is renormalized so that its spread does not change along the dial.

At $\alpha = 1$ the input is the ESM3 embedding itself, the configuration used throughout the paper. At $\alpha = 0$ the random vectors are linearly independent with probability one, and since the encoder's first operation is affine (Eq. (1)), this end is equivalent to one-hot initialization: the graph knows only which receptor a node is. Intermediate values of $\alpha$ move smoothly from this identity-only input towards ESM3, which also identifies each receptor but adds what a protein model is expected to provide: a geometry of protein space in which related receptors lie close to one another.

Figure 1 shows the dial in both forms of Appendix A, each as a paired difference from XGBoost-base. In the full form $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}} \| \mathbf{z}_{\text{prot}}]$ the predictor receives the original ESM3 embedding at every $\alpha$: whatever the graph is initialized with, the sequence reaches the predictor anyway, so the contrast between bare identity and the geometry supplied by the protein model is largely masked. In the reduced form $[\mathbf{z}_{\text{prot}} \| \mathbf{x}_{\text{mol}}]$, $\mathbf{z}_{\text{prot}}$ is the only receptor information the predictor receives, so moving $\alpha$ moves exactly the share of that information which comes from the sequence.

Each point in Figure 1 is the difference between OlfaGraph and XGBoost-base taken within each held-out split and then averaged over splits, rather than the difference between their averages. The splits differ considerably in overall difficulty; this variation is shared by both models, and without the pairing it would dominate the intervals and hide differences between the models that are consistent from split to split.

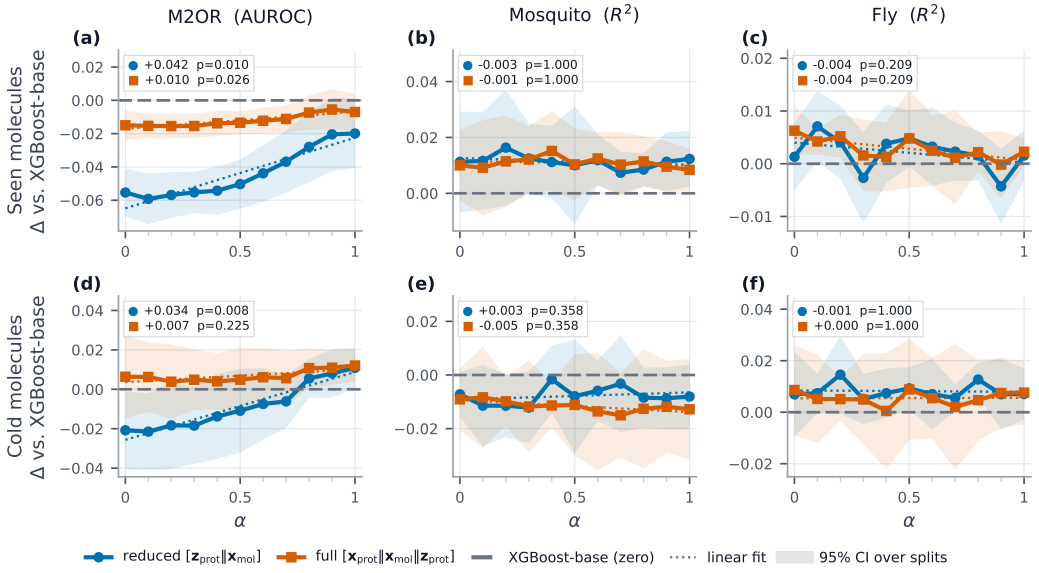


Figure 1: Difference between OlfaGraph and XGBoost-base as a function of $\alpha$ (Eq. (5)) on M2OR, Mosquito and Fly (columns) in the seen-molecule and cold-molecule settings (rows), for the reduced form $[\mathbf{z}_{\text{prot}} \| \mathbf{x}_{\text{mol}}]$ and the full form $[\mathbf{x}_{\text{prot}} \| \mathbf{x}_{\text{mol}} \| \mathbf{z}_{\text{prot}}]$. Zero is the level of XGBoost-base. Bands are 95% confidence intervals over held-out splits; dotted lines are linear fits, and each box gives their slopes with the Holm-corrected $p$-values for a slope of zero.

On M2OR the reduced-form curves rise significantly with $\alpha$ in both settings (Figure 1a,d): the protein model adds to what the measurements provide, and for cold molecules it lifts OlfaGraph from below XGBoost-base to above it. On Mosquito and Fly the slopes do not differ significantly from zero (Figure 1b,c,e,f): identity alone already reaches the level of ESM3, so the geometry from

protein models does not carry additional information here. The full-form curves are much flatter throughout.

### B.2 THE IDENTITY-ONLY END

Table 5 sets the $\alpha = 0$ end against XGBoost-base and against the same predictor with a one-hot receptor block in place of the ESM3 embedding. Both know each receptor only by its identity and are trained on the same measurements; they differ only in how that identity is used, read directly by XGBoost or first passed through the graph of measured responses. OlfaGraph at $\alpha = 0$ is the random-initialization row of Table 2, and its advantage over the one-hot predictor was already noted in Section 5.1; here both identity-only predictors are read as paired differences from XGBoost-base.

On Fly, and for cold molecules on Mosquito, neither differs significantly from XGBoost-base: there the ESM3 embedding adds nothing that receptor identity does not already provide. For seen molecules on Mosquito the two part ways. XGBoost over the one-hot block falls clearly behind XGBoost-base, whereas OlfaGraph at $\alpha = 0$, given the same identity and the same measurements, does not. What falls short there is not the identity but a predictor that reads it directly; passed through the graph of measured responses, the same identity is enough. On M2OR, in contrast, both identity-only predictors fall significantly below XGBoost-base in both settings. OlfaGraph halves the gap left by the one-hot predictor but does not close it: on this dataset the protein language model embedding carries information that the measurements do not replace, in line with the rising reduced-form curves on M2OR in Figure 1. An identity-only representation can therefore match a protein language model embedding on the insect panels, in one setting only when read through the graph, but not on M2OR.

Table 5: Two predictors that know each receptor only by its identity, compared with XGBoost-base, $[\mathbf{x}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}}]$: OlfaGraph at $\alpha = 0$ in the reduced form $[\mathbf{z}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}}]$, and XGBoost with a one-hot receptor block in place of $\mathbf{x}_{\mathrm{prot}}$. Differences are taken within each held-out split and averaged over the 5 splits; positive values favor the identity-only predictor. Mean $\pm$ 95% CI over splits. $^*$: paired $t$-test, Holm-corrected over the two differences in a row, $p < 0.05$.

| | | | **Difference from XGBoost-base** | |
|---|---|---|---|---|
| **Dataset** | **Setting** | XGBoost-base | OlfaGraph, $\alpha$=0 | XGBoost, one-hot |
| M2OR (AUROC) | Seen molecules | 0.896±0.015 | −0.055±0.014$^*$ | −0.102±0.017$^*$ |
| | Cold molecules | 0.820±0.043 | −0.021±0.020$^*$ | −0.041±0.028$^*$ |
| Mosquito ($R^2$) | Seen molecules | 0.656±0.026 | +0.011±0.018 | −0.112±0.042$^*$ |
| | Cold molecules | 0.483±0.092 | −0.007±0.008 | +0.020±0.018 |
| Fly ($R^2$) | Seen molecules | 0.561±0.064 | +0.001±0.006 | +0.003±0.030 |
| | Cold molecules | 0.484±0.074 | +0.007±0.017 | +0.001±0.032 |

## C GRAPH CONSTRUCTION

OlfaGraph's edges are the measured training pairs themselves, each signed by its label (Section 3.1); no similarity rule enters, so the only degree of freedom is which odorants carry messages. It is set by a cut $q$, which keeps the odorants whose number of training measurements reaches the $q$-th quantile of that number and so fixes how many survive, $K$, and by a criterion that decides which $K$ odorants these are. We compare seven criteria: the number of training measurements (coverage); the summed inverse document frequency of the receptors an odorant is measured on (IDF coverage); the binary entropy of its fraction of positive responses (balance); coverage times balance (entropy); the product of its positive and negative counts (discriminative pairs); balance times IDF coverage (composite); and the greedy maximization of covered receptor pairs described in Section 3.1 (greedy pair cover). $K$ odorants drawn at random serve as a control. At $q = 0$ nothing is cut and every criterion yields the complete graph. All results on M2OR use greedy pair cover at $q = 0.99$. On Mosquito and Fly every odorant is measured against every receptor, so the cut has nothing to act on and the graph is always complete.

Figure 2 shows the sweep on M2OR, read on the validation split. With the complete graph, about 580 odorants, OlfaGraph falls clearly below XGBoost-base for seen molecules and slightly below it for cold ones. Cutting the graph raises it in both settings, and at $q = 0.99$, where about six odorants carry messages, it reaches XGBoost-base for seen molecules and lies slightly above it for cold ones. The reported configuration is ahead of the complete graph, significantly so for seen molecules, and no cell of the grid is significantly ahead of it after correction for the size of the grid; for cold molecules it is the best cell. At the strongest cut the remaining criteria and the random control lie within split-to-split variation of one another, except for three label-based criteria (entropy, discriminative pairs, composite), which fall behind for seen molecules.

It is notable how little of the molecular side the graph needs: about six odorants, roughly one percent of those with training measurements, and the complete graph does clearly worse. This suggests that OlfaGraph uses the odorant nodes not as a record of every measured interaction but as a small shared basis, a handful of hub odorants through which receptors exchange information about one another. Each receptor is then described by how it responds to these few reference odorants and, through them, by which other receptors respond alike. The criterion that performs well at this cut fits this reading: greedy pair cover chooses the odorants so that as many receptor pairs as possible are measured on a common odorant, that is, so that as many receptors as possible can learn about one another through the same small set of odorants.

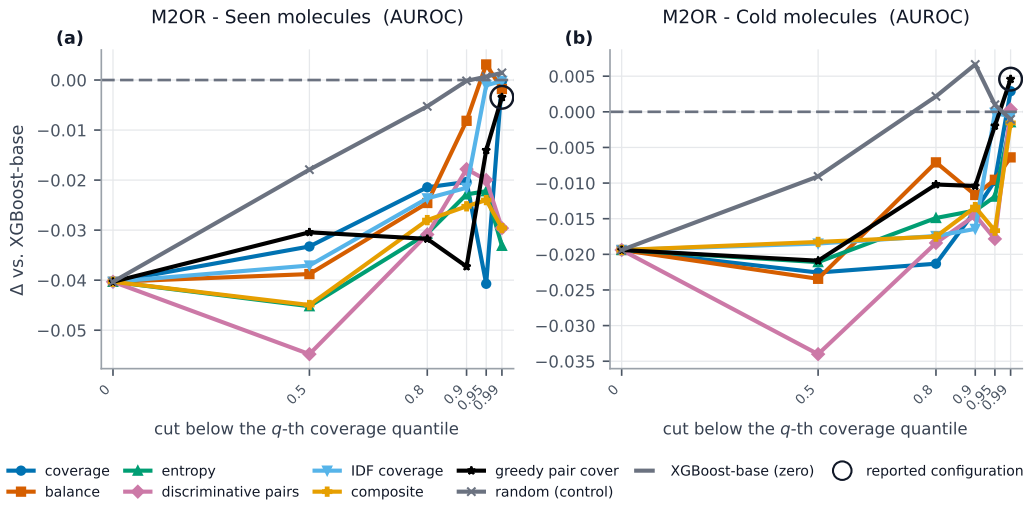


Figure 2: Difference between OlfaGraph and XGBoost-base on M2OR as a function of the cut $q$, for seen (a) and cold (b) molecules, on the validation split; one curve per odorant-selection criterion. Differences are taken within each held-out split and averaged over splits. Zero is the level of XGBoost-base; the circle marks the reported configuration, greedy pair cover at $q = 0.99$.

## D MOLECULAR REPRESENTATION

Throughout the paper the odorant side is fixed to ChemBERTa (Chithrananda et al., 2020), the encoder that both ProSmith (Kroll et al., 2024) and LORAX (McConachie et al., 2026) build on. McConachie et al. (2026) report that, as long as the odorant representation is kept frozen, its choice has little effect on predictive performance: on Mosquito, Fly and M2OR, embeddings from chemical foundation models perform on par with graph-based encoders and classical descriptors. As an additional check in our setting, we repeat the comparison with two further odorant embeddings of a different kind, a pretrained GIN (Xu et al., 2018) and the ECFP fingerprint. Among the baselines, only the model of Hladiš et al. (2023) is included: the available implementations of ProSmith and LORAX are built around Transformer encoders that operate on per-site molecular representations, which pooled embeddings such as GIN and ECFP do not provide, whereas the model of Hladiš et al. is not of this kind. In the comparison of McConachie et al., that model appears only on M2OR, through the values reported in the original publication; we reimplemented it from scratch and trained

it on every dataset, split and odorant embedding considered here, so that it is compared with the other rows split by split.

Across the three embeddings the methods keep much the same relative positions as with ChemBERTa (Table 6). On M2OR, OlfaGraph trails XGBoost-base for seen molecules and is on par with it or ahead for cold molecules; on Mosquito it is ahead for seen molecules; on Fly the methods are indistinguishable. The embedding shifts the absolute level more than it shifts the gap between methods, with ECFP the weakest input throughout. The model of Hladiš et al. is strongest for seen molecules on M2OR but falls behind for cold molecules on the insect datasets.

Table 6: OlfaGraph in the reduced form $[\mathbf{z}_{\mathrm{prot}} \| \mathbf{x}_{\mathrm{mol}}]$, XGBoost-base and Hladiš with each molecular embedding, in the seen-molecule and cold-molecule settings; the metric is named beside each dataset. Mean $\pm$ std over 5 held-out splits, and in parentheses the mean place among the three methods within a split. **Bold** = best for its dataset, setting and embedding, taken on the mean. $^*$ = OlfaGraph is ahead of that row, paired $t$-test over splits, Holm-corrected within the group. *Mean place* averages the place over the three embeddings.

| | | **Seen molecules** | | | **Cold molecules** | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Embedding** | OlfaGraph | XGBoost-base | Hladiš | OlfaGraph | XGBoost-base | Hladiš |
| M2OR (AUROC) | ChemBERTa | 0.876±0.017 (3.00) | 0.896±0.012 (2.00) | **0.902±0.008 (1.00)** | **0.830±0.027 (1.00)** | 0.820±0.035 (2.20)$^*$ | 0.815±0.033 (2.80)$^*$ |
| | GIN | 0.881±0.015 (3.00) | 0.898±0.009 (1.80) | **0.900±0.011 (1.20)** | **0.829±0.033 (1.60)** | 0.824±0.037 (2.40) | 0.824±0.046 (2.00) |
| | ECFP | 0.863±0.013 (3.00) | 0.880±0.010 (2.00) | **0.891±0.009 (1.00)** | 0.832±0.023 (2.00) | **0.838±0.018 (1.40)** | 0.825±0.035 (2.60) |
| | *Mean place* | 3.00 | 1.93 | **1.07** | **1.53** | 2.00 | 2.47 |
| Mosquito ($R^2$) | ChemBERTa | **0.668±0.021 (1.20)** | 0.656±0.021 (2.40) | 0.656±0.028 (2.40) | 0.475±0.078 (1.80) | **0.483±0.074 (1.20)** | 0.378±0.134 (3.00) |
| | GIN | **0.665±0.030 (1.60)** | 0.664±0.021 (2.20) | 0.660±0.026 (2.20) | **0.515±0.060 (1.20)** | 0.499±0.073 (2.00) | 0.387±0.107 (2.80) |
| | ECFP | 0.602±0.047 (1.80) | 0.564±0.033 (3.00)$^*$ | **0.628±0.029 (1.20)** | **0.395±0.082 (1.80)** | 0.392±0.080 (2.40) | 0.340±0.141 (1.80) |
| | *Mean place* | **1.53** | 2.53 | 1.93 | **1.60** | 1.87 | 2.53 |
| Fly ($R^2$) | ChemBERTa | **0.562±0.054 (1.20)** | 0.561±0.052 (1.80) | 0.546±0.063 (3.00) | **0.491±0.064 (1.20)** | 0.484±0.060 (1.80) | 0.416±0.086 (3.00) |
| | GIN | 0.545±0.053 (1.60) | **0.547±0.054 (1.80)** | 0.534±0.058 (2.60) | 0.458±0.063 (2.00) | **0.467±0.066 (1.40)** | 0.389±0.081 (2.60) |
| | ECFP | **0.530±0.035 (2.00)** | 0.523±0.026 (2.40) | 0.522±0.066 (1.60) | 0.353±0.075 (2.20) | **0.356±0.070 (2.00)** | 0.344±0.163 (1.80) |
| | *Mean place* | **1.60** | 2.00 | 2.40 | 1.80 | **1.73** | 2.47 |