

Decoding Smell from Receptor Structure

Hiroaki Matsunami

hiroaki.matsunami@duke.edu

Duke University School of Medicine <https://orcid.org/0000-0002-8850-2608>

Hsiu-Yi Lu

Duke University

Aashutosh Vihani Vihani

<https://orcid.org/0000-0001-5979-101X>

Maira Nagai

Duke University

Xiaoyang Hu

Duke University

Dan Takase

Duke University

Conan Juan

Duke University

Claire De March

Université Paris-Saclay <https://orcid.org/0000-0001-6461-5547>

Biological Sciences - Article

Keywords:

Posted Date: June 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9947738/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

There is **NO** Competing Interest.

All animal procedures were approved and overseen by the Duke University Institutional Animal Care and Use Committee (IACUC) and conducted in accordance with Duke University's Division of Laboratory Animal Resources (DLAR) policies and applicable institutional and federal guidelines.

Title: Decoding Smell from Receptor Structure

Authors: Hsiu-Yi Lu^{1,†,*}, Aashutosh Vihani^{2,†}, Maira H. Nagai¹, Xiaoyang S. Hu¹, Dan Takase¹, Conan Juan¹, Claire A. deMarch¹, Hiroaki Matsunami^{1,3,*}

Affiliations:

¹ Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Research Drive, Durham, NC 27710 USA

² Department of Neurobiology, Neurobiology graduate program, Duke University Medical Center, Durham, NC 27710, USA

³ Department of Neurobiology, Duke University School of Medicine, Research Drive, Durham, NC 27710 USA

† These authors contributed equally to this work

*Corresponding author. Email: hyjusticelu@gmail.com & Hiroaki.matsunami@duke.edu

Abstract: Olfaction enables animals to detect and discriminate an immense diversity of volatile chemicals, yet how individual odorant receptors determine ligand selectivity remains poorly understood (1-4). Recent cryo-electron microscopy structures of mammalian odorant receptors(5-11), together with advances in protein structure prediction(12-14), now provide an opportunity to test whether receptor structure can explain odor recognition at scale. Here, we combine AlphaFold3-predicted odorant receptor structures with ESM2-derived protein sequence embeddings and large-scale sequencing linked to in vivo olfactory sensory neuron activation across a chemically diverse odorant panel. Using these data, we develop a structure-informed deep learning framework that predicts odorant receptor-ligand interactions and maps relationships between receptor space and chemical space. The learned receptor representation organizes odorant receptors by functional similarity rather than primary sequence alone, while feature attribution identifies binding-cavity subregions that contribute to chemical recognition. Together, these results show that odorant receptor ligand selectivity is encoded in three-dimensional receptor features and establish a predictive framework for interpreting odorant receptor coding.

Main Text:
Introduction

Animals rely on multiple sensory systems to navigate and interpret their environment, each specialized to encode distinct types of physical or chemical information. Among these, olfaction presents one of the least understood and most complex encoding challenges. In color vision, the task is to discriminate between wavelengths of light and is achieved through the tuning of a small number of cone opsins with distinct spectral sensitivities, which are a subfamily of class A G protein-coupled receptors (GPCRs), is central to how humans and other animals perceive light as distinct colors (15, 16). Odorant receptors (ORs) in vertebrates are also class A GPCRs; however, in contrast to vision, olfaction must distinguish among diverse chemical molecules, requiring large repertoires of odorant receptors (ORs). ORs ranging from hundreds to thousands to encode this high-dimensional chemical space, which constitute the largest subfamily of class A GPCRs (1), with humans expressing approximately 400 intact ORs and mice over 1,000 (3, 17, 18). Mammals devote a substantial fraction of their genomes to ORs, to navigate the environment (4, 19), which reflects the evolutionary importance of chemical sensing across species. Despite their central role in odor perception, a fundamental unresolved question is how an individual OR determines which chemicals it binds and how it responds across chemical space. Addressing this question is essential for understanding the logic by which chemical information is encoded at the receptor level, yet progress has been limited by long-standing technical barriers to obtaining high-resolution OR structures.

Recent cryo-EM studies have resolved structures of mammalian ORs, first with OR51E2 in complex with propionate (5-11). Notably, these cryo-EM structures revealed that odorants

primarily occupy a conserved binding cavity location, suggesting a canonical structural locus for odorant recognition across ORs (5, 6, 20). Functional studies further demonstrate that mutations altering the shape or chemical properties of this cavity can shift ligand preference, suggesting ligand selectivity is primarily influenced by the geometry and physicochemical features of the binding pocket.

In parallel, advances in computational biology have transformed access to protein structural and evolutionary information. Deep learning-based structure prediction methods such as AlphaFold3 (12, 13, 21), have provided the potential for accurate *in silico* predictions for the entire OR repertoire, offering a complementary resource where experimental structures are lacking. Additionally, other approaches like protein large language models (LLMs), such as the Evolutionary Scale Model (ESM) (22) extend this progress by embedding sequence-based evolutionary information into learned vector spaces that infer functional and biophysical constraints. Together, these approaches overcome a major limitation by providing receptor-level structural information at scale, creating an opportunity to pursue structure-informed modeling of OR-ligand interactions.

Many existing efforts to model olfactory perception and OR-ligand interactions have focused on ligand descriptors, psychophysical measurements, and receptor sequence features, yielding important insights into odor coding and chemical space (23-30). These approaches offer interpretable representations of odorant chemistry and perceptual outcomes. However, the limited availability of high-resolution receptor structures has limited the ability to connect ligand binding to three-dimensional features of the receptor binding cavity. Given that the receptor

structure plays a central role in determining ligand specificity and activation (5, 6), integrating structural information is essential for directly linking chemical recognition to receptor function.

To address this gap, we developed an end-to-end modeling framework that integrates AlphaFold-predicted OR structures with ESM2-derived sequence embeddings within convolutional neural networks to predict OR-ligand interactions. Central to this effort is the use of pS6-IP-Seq, an *in vivo* assay that captures activation of olfactory sensory neurons through enrichment of phosphorylated ribosomal protein S6-associated transcripts (31, 32). While numerous studies have reported OR-ligand interactions, these datasets are typically small, heterogeneous in experimental design, and difficult to integrate across studies. In contrast, collection of pS6-IP-Seq studies enables systematic, parallel measurement of receptor activation across chemically diverse odorants within a unified experimental framework, making it particularly well suited for machine learning applications.

By integrating large-scale pS6-IP-Seq activation data with predicted receptor structures and sequence-derived embeddings, we set out to ask whether receptor structure representation can be explicitly leveraged to model and interpret OR-ligand interactions. Specifically, we aimed to determine the structural features of the receptor that encodes for the ligand selectivity and receptor tuning. Through this approach, we seek to establish a structure-informed framework for linking receptor feature to odorant activation, and to assess whether focusing on a canonical binding cavity can provide a mechanistic basis for understanding olfactory coding at the receptor level.

Results

pS6-IP-Seq as a Strategy to Identify Odorant Activated OR In Vivo

The phosphorylation of the ribosomal protein S6 is a well-established marker of neuronal activity (33). Similar to the induction of immediate early genes such as c-Fos, pS6 reflects recent neuronal activation, but with the added advantage of being physically associated with mRNA molecules. As a result, phosphorylated ribosomal S6 (pS6) complex can be immunoprecipitated and coupled to RNA sequencing, enabling a direct molecular readout of transcriptional profiles from activated olfactory sensory neurons (OSNs) following odor exposure (Fig. 1A). Given, each mature OSN expresses a single olfactory receptor gene (34-40), we can directly assign OR identities to individual OSNs. This organization allows pS6-IP-Seq to link odor-evoked neuronal activity to specific OR identities in vivo. Consistent with this framework, previous studies have demonstrated the utility of pS6-IP-Seq in inferring odorant-OR interactions in vivo freely behaving mice (31, 41, 42).

To ensure broad coverage of chemical space, we selected a panel of 48 odorants spanning diverse functional groups, molecular scaffolds, and physicochemical properties from panel of small molecules commonly found in foods and fragrances (43). In addition to maximizing global chemical diversity, the panel was intentionally designed to include closely related chemical pairs, such as enantiomers and homologous series differing by carbon chain length, enabling assessment of receptor sensitivity to subtle structural variations. Projection of these odorants into a reduced chemical feature space using Morgan fingerprint based chemical features (44), as well as clustering within an Open Principal Odor Map (OpenPOM) derived perceptual space (27, 45),

generated from message passing neural networks trained on human odor perception, confirmed that the selected panel captures both structural and perceptual diversity across odorants (Fig. 1C and fig. S1B).

Defining Activation Score to Quantify OR Responses

Traditional identification of activated olfactory receptors in pS6-IP experiments often repurposes differential expression criteria developed for gene-level analyses, such as false discovery rate (FDR) thresholds (e.g., $\text{FDR} < 0.05$) or minimum log fold-change ($\log\text{FC} > 1$). While effective for detecting broadly regulated transcripts, these metrics are not optimized for classifying receptor activation and introduce systematic biases when applied to ORs. LogFC is strongly influenced by transcript abundance and sequencing depth: ORs with low counts can exhibit inflated fold changes due to small denominators, whereas highly expressed receptors frequently display more modest logFC values despite reproducible enrichment. As a result, logFC-based selection preferentially highlights sparsely expressed ORs while deprioritizing abundantly expressed receptors, confounding biological activation with sampling effects (Fig. 1E). In contrast, FDR-based selection introduces a complementary but distinct bias. Because FDR reflects statistical confidence rather than effect size, it preferentially prioritizes receptors with stable baseline expression and low variance, even when the magnitude of enrichment is small. Consequently, receptors exhibiting consistent but weak changes can be classified as activated, whereas biologically meaningful but variable responses may be excluded. Together, reliance on either logFC or FDR alone emphasizes different statistical regimes rather than receptor activation, limiting their interpretability for OR classification.

To address these limitations, we developed a composite metric, termed the Activation Score, which integrates both enrichment magnitude and statistical confidence without relying on a single criterion alone.

$$Activation\ Score = -\log(FDR) * \log FoldChange \left(\frac{Odor}{No\ Odor} \right)$$

By jointly considering effect size and reliability, the Activation Score captures activation signals that reduces sensitivity to transcript abundance, variance structure, and arbitrary threshold selection. The advantage of this approach is illustrated in volcano plots of acetophenone pS6-IP responses (Fig. 1B), where increasing Activation Score thresholds yield a more balanced identification of responsive ORs compared with conventional FDR- or logFC-based filtering (Fig. 1E).

Furthermore, because different odorants elicit varying overall activation strengths at fixed experimental concentrations, direct comparison of raw activation scores across odors can introduce additional bias. To mitigate this, we normalized activation scores by z-scoring within each odor experiment, generating an odor-specific activation z-score. This normalization ensures that OR detection reflects relative responsiveness within an odor condition rather than absolute signal magnitude. Using this normalized metric, we defined activated ORs as those with an activation z-score greater than 2 for all downstream analyses.

Comparison of the top 1.06% of OR-odorant pairs, corresponding to the number of pairs exceeding an activation z-score of 2. Whereas FDR- and logFC-based criteria were skewed toward small subsets of odorants, activation z-score-based selection preserved a more diverse

distribution of odor categories (Fig. 1E), indicating reduced bias toward ligands that produce exceptionally strong or abundant signals (46).

Analysis of OR responses identified using the activation z-score revealed broad coverage of receptor space. Projection of activated ORs into a sequence-based receptor space defined by Grantham distances demonstrated that activated receptors were widely distributed rather than concentrated within specific OR families or classes (Fig. 1D). This result was further supported by phylogenetic analysis, in which activated ORs were dispersed throughout the OR tree (Fig. 1F). Together, these analyses indicate that pS6-IP enables deorphanization across a diverse and representative fraction of the ORs.

Consistent with this diversity, clustering of OR-odorant responses across all 48 odor's pS6-IPseq revealed that receptors did not group according to family identity, which is defined by overall amino acid similarities (47), but rather according to shared ligand responsiveness (Fig. 1G). Conversely, ORs with high sequence similarity did not necessarily respond to the same odorants, highlighting the dissociation between phylogenetic proximity and functional tuning.

Together, the odor-normalized Activation Score z-scored provide a more balanced and interpretable framework for quantifying OR activation from pS6-IP data. This metric reduces bias associated with transcript abundance and differential response strength, forming a robust foundation for the downstream structural modeling and machine learning analyses presented in this study.

Encoding Odorant Receptor Structure

Having established a large-scale, chemically diverse dataset of in vivo OR activation, we next asked whether receptor structure contains information to predict ligand selectivity.

Specifically, we sought to determine whether ORs could be represented in a structure-informed space that reflects their likelihood to respond to a given chemical stimuli, thereby enabling an interpretable model of receptor coding.

We hypothesized that the three-dimensional architecture of the receptor, particularly features within the ligand-binding cavity encodes constraints that shapes chemical recognition. Under this hypothesis, structural representations alone should organize ORs into a space that reflects their patterns of chemical responsiveness, even without incorporating ligand features.

To establish a reliable structural foundation for this analysis, we first evaluated the ability of AlphaFold3 to generate accurate, activation-relevant OR conformations. As a test case, we modeled OR51E2 in complex with a mini Gas protein using AlphaFold3 multimer. Notably, inclusion of the mini-Gas protein alone was sufficient to stabilize an active-like receptor conformation, without providing the ligand (fig. S3). Comparison of the predicted structure with the recently solved cryo-EM structure of OR51E2 (5) revealed less than 1 Å RMSD overall backbone deviation, indicating close agreement between predicted and experimental models (Fig. 2A and fig. S3). This result is consistent with prior observations that AlphaFold predictions tend toward inactive-like states when modeled in isolation but can be shifted toward active conformations through inclusion of signaling partners such as G proteins (Fig. 2A and B)(5). Importantly, similar structural concordance was observed across additional OR cryo-EM structures spanning distinct classes and subfamilies, with AlphaFold3 models showing

consistently low backbone deviations. Together, these results support the use of AlphaFold3 with OR with miniGas to generate active-like structural models for diverse ORs (Fig. 2B and fig. S3).

We next assessed whether these predicted structures were suitable for interrogating ligand selectivity by examining the architecture of the binding cavity. Analysis of residues lining the predicted ligand-binding pocket showed that AlphaFold3 accurately recapitulates not only the overall receptor fold but also the positioning of key side chains that define cavity shape (Fig. 2C). The close correspondence between predicted and experimentally observed side-chain orientations indicates that AlphaFold3 models preserve the structural features critical for defining the binding pocket, supporting their use for downstream structural encoding and machine learning analyses.

Feature encoding of primary sequence

To examine how sequence-derived representations capture OR similarity, we leveraged embeddings generated by the Evolutionary Scale Model (ESM), a protein language model trained on large-scale protein sequences to learn contextual and evolutionary patterns directly from sequence data (22). Unlike traditional representations based on primary amino acid identity or predefined substitution matrices, ESM embeddings encode each residue in the context of its surrounding sequence enabling the representation to capture higher-order relationships. ESM-based embeddings have been shown to correlate with protein structure, mutational effects, and functional similarity across diverse protein families (14, 48). To compare against a baseline measure of OR similarity from primary sequence, we computed pairwise distances using Grantham's amino acid substitution matrix. These distances capture differences in

physicochemical properties such as polarity, volume, and composition, and thus provide a familiar framework for comparing receptor families. Using multidimensional scaling, we projected the pairwise Grantham distances into two dimensions (Fig. 2D). The resulting clusters closely mirrored known OR class and family designations, confirming that physicochemical substitution metrics reflect evolutionary relationships across the OR repertoire as expected.

Full-length OR sequences were encoded using ESM, and the resulting embeddings were projected into principal component space for visualization (Fig. 2D). Compared with Grantham distance-based clustering, which primarily reflects amino acid similarity and substitution patterns, ESM embeddings showed weaker segregation by canonical OR families, suggesting that they capture sequence features beyond simple residue identity (Fig. 2D-E).

To directly compare these representations, we computed pairwise distances using both ESM embeddings and Grantham metrics. While the resulting distance matrices showed broad agreement at the level of OR class and family organization (Fig. 2E), they also revealed distinct patterns unique to each approach. Correlation analysis confirmed a moderate but incomplete relationship between the two measures (fig. S2), indicating that ESM embeddings encode additional evolutionary and contextual information not captured by traditional substitution-based metrics. Together, these results highlight that ESM provides a complementary, higher-order representation of OR sequence space.

Integration of Structural Voxels and ESM Embeddings

With validated structural models and residue-level embeddings in hand, we developed a unified spatial learning framework that integrates three-dimensional receptor architecture with

sequence-derived features into a single representation (Fig. 3A). For each OR, the cavity space was calculated using pyKVFinder (49) from the AlphaFold3-predicted structure generated in complex with mini-G α . These cavities were then filtered against the canonical binding cavity reference, yielding a single, standardized ligand-binding region for each receptor. Amino acid residues encasing this cavity space were extracted as the structural representation of the OR.

The resulting cavity-centered receptor structures were discretized into a three-dimensional voxel grid of size 32 x 32 x 32 at 1 Å resolution, with the grid centered on the predicted ligand-binding cavity (Fig. 3B). For each receptor, residues contributing to the binding cavity were first identified based on their spatial proximity, and each of their amino acid residue's atomic coordinates were mapped onto the voxel grid according to their three-dimensional positions. Each occupied voxel was annotated with sequence-derived features obtained from ESM2 embeddings corresponding to the residue mapped to that spatial location. This voxelization strategy yields a spatially aligned, feature-aware representation of each receptor that jointly encodes binding cavity structure and residue identity for three-dimensional convolutional learning.

To provide functional supervision for training, we leveraged the pS6-IP-Seq data described above. Using the activation z-score metric greater than or equal to 2, which equates to roughly 0.05 of population in normal distribution, each OR was assigned to binary activated or inactive for each individually tested odors. This labeling scheme transforms the experimental response data into a multi-label classification task: given the structural and embedding representation of an OR, predict which odor it is likely to respond to. In this way, the pipeline

effectively provides the first attempt at building an interpretable olfactory map directly from structural information.

Convolutional Neural Network Odor Response Model

To predict OR odor responses, we trained a convolutional neural network (CNN) using voxelized binding cavity representations of 349 ORs responding to at least one odorant. For each receptor, five independent AlphaFold3 structural predictions were included, and voxel-level data augmentation was applied to increase robustness to minor structural variation. The dataset was randomly partitioned into training, validation, and test sets using a 70/15/15 split at the receptor level to prevent information leakage across folds. The model was optimized using the Adam optimizer with binary cross-entropy loss in a multi-label classification setting, reflecting the fact that individual receptors may respond to multiple odorants.

Across training runs, both training and validation losses decreased with no evidence of divergence between the two curves, indicating stable optimization without overfitting (fig. S4). Improvements in loss were accompanied by concordant increases in validation F1-score, demonstrating that optimization gains translated into improved classification performance rather than trivial fitting to dominant classes (fig. S4).

The final Spatial LLM CNN achieved consistently high predictive performance across the odor panel. Multi-class ROC analysis yielded a macro-averaged AUCROC of 0.959, indicating strong overall discriminative ability across odors (Fig. 3C). However, because the dataset is heavily class-imbalanced, with substantially more inactive than active OR-odorant pairs, ROC-AUC alone can overestimate performance due to the abundance of true negatives

(50). To address this potential limitation, we evaluated additional performance metrics that are more sensitive to prediction confidence and class balance. First, we assessed whether model output probabilities reflected experimentally measured activation strength by correlating predicted confidence scores with pS6-IPseq activation values across receptors for each odorant (Fig. 3D). This analysis revealed strong correspondence between model confidence and biological activation, with 41 of 48 odorants exhibiting correlation coefficients greater than 0.6 and a mean correlation of 0.71 across odors (Fig. 3E and table S1). Second, precision-recall performance remained robust, with 32 of 48 odorants achieving F1-scores above 0.6 and a mean F1-score of 0.67, indicating balanced sensitivity and precision across most odor categories (Fig. 3D, fig. S4 and table S1).

We further examined the relationship between predictive performance and the number of positive training ORs available per odorant. Correlation strength increased with the number of activated ORs used for training, revealing a significant positive association between data availability and model performance (Fig. 3D). Notably, this trend was not deterministic, several odorants with relatively few positive examples nonetheless achieved high correlation and F1-scores, suggesting that certain odor response patterns are more easily inferred or generalized between receptors. Together, these complementary assessments confirm that the high ROC-AUC values reflect discriminative performance rather than artifacts of class imbalance. The Spatial LLM framework therefore provides an approach for predicting OR-odorant interactions from structural representations.

Feature Embeddings

To interpret the learned representation of OR-odorant interactions, we extracted receptor embeddings from the final fully connected layer of the trained CNN. These vectors provide compact, high-level encodings of receptor structure after transformation through the convolutional architecture. Embeddings for each OR were retained alongside experimentally
5 determined odor response labels for downstream analyses.

We first examined the global organization of the learned representation using UMAP dimensionality reduction for visualization. The resulting two-dimensional embedding, which we term smeLLMap, provides a spatial map of the OR landscape in which each point corresponds to a single receptor and its position reflects similarity in the learned odor-response representation, with colors indicating ligand selectivity (Fig. 4A and fig S5). At a global level, receptors do not
10 strongly segregate by canonical family assignments (Fig. 4B). Consistent with this, functional organization emerges as the dominant structure. When receptors responding to individual odorants are visualized, they form distinct and coherent clusters within smeLLMap (Fig. 4C and fig. S5), indicating that the learned representation groups ORs according to shared ligand
15 selectivity rather than phylogenetic relatedness. To quantify the relationship between embedding proximity and sequence similarity, we computed Grantham distance between 5 k nearest neighbors (kNN) derived from smeLLMap in comparison with those defined by amino acid-based neighbors for each ORs. Nearest neighbors in the smeLLMap exhibited significantly higher Grantham distances than amino acid derived neighbors (Fig. 4B). This result demonstrates
20 that local organization in the smeLLMap is not solely driven by sequence similarity, but instead reflects alternative features captured by the model.

We next asked whether model embedding-based similarity better reflects shared ligand selectivity than primary amino acid sequence similarity. To address this, we computed pairwise Grantham distances between ORs responding to the same odor and compared these distributions with Euclidean distances measured in the smeLLMap. Subsequently, for 41 of 48 odorants, responding ORs show higher similarity between learned embedding than amino acid similarity (Fig. 4D and table S2). Together, these patterns and difference indicates that the learned embedding representation captures functional similarities between receptors that are not apparent from primary amino acid sequence similarity alone, highlighting the added value of the CNN-derived structural embedding over sequence-based metrics.

Chemical and Perception Inference

We next asked whether the smeLLMap reflects relationships between odorants, despite the model having no explicit access to ligand chemical features during training. To address this, we first examined whether receptors responding to chemically similar odorants occupy similar regions in the learned latent space. For each odorant, the set of responding OR embeddings was treated as a distribution, and pairwise distances between odorants were quantified using Wasserstein distance (Fig. 5A and fig. S6). Odorants with high structural similarity exhibited significantly smaller distances between their corresponding receptor distributions. This was observed for closely related compounds such as stereoisomeric pairs, as well as homologous series including heptanal and octanal, which differ by a single carbon unit (Fig. 5A and fig. S6). In contrast, chemically dissimilar odorants, such as octanal and (+)-carvone, showed substantially larger distances between their respective receptor distributions. These results

indicate that smeLLMap can infer chemical similarity through receptor organization, even in the absence of explicit ligand descriptors.

While these relationships highlight the ability of the model to recover chemical similarities between enantiomer and chain isomer odorants, we next asked whether this organization extends beyond chemical features to reflect perceptual similarity. To test this, for each pair of odorants we computed the distance between their responding OR distributions in smeLLMap and compared these values to the corresponding distances derived from OpenPOM (27, 45) space. Across all 48 odorants, we observed a significant positive correlation between receptor embedding distance and perceptual distance ($r^2=0.112$, $p=6.372e-31$) (Fig. 5B and fig. S6A), indicating that odorants perceived as more similar by humans panels tend to be represented by similar ORs represented with regions of receptor space. Notably, this relationship was not clear when using amino acid sequence-based Grantham distances between responding ORs ($r^2=0.007$, $p=2.283e-3$) (Fig. 5B and fig. S6B), demonstrating that primary sequence similarity alone does not capture perceptual organization.

To further visualize how perceptual features are represented across receptor space, we projected individual OpenPOM-derived odor descriptors onto smeLLMap. By mapping specific odor descriptors, such as floral, roasted and sulfurous qualities, revealed localized clusters of receptors preferentially associated with descriptors (Fig. 5C). We next examined whether this organization could be observed at a global level by integrating the full set of 138 OpenPOM descriptors. Utilizing dimensionality reduction of these perceptual features followed by projection onto smeLLMap, with each principal component axis as a representation RGB color space revealed further clustering of colored gradients. Likely indicating that perceptual features

may also be organized and inferred by the OR space, highlights a structured encoding in which different regions of receptor space are tuned to distinct combinations of perceptual features. To quantitatively assess whether local neighborhoods in smeLLMap reflect shared perceptual tuning, we computed kNN relationships and compared the OpenPOM (27, 45) representations of neighboring receptors. Because our analyses relate mouse receptor activation to human perception derived features, they should be interpreted with appropriate caution. Nevertheless, receptors exhibited significantly greater similarity in perceptual representation to their nearest neighbors than random (Fig. 5D), supporting the idea that smeLLMap captures perceptually relevant organization.

Together, these results demonstrate that smeLLMap captures not only chemical similarity but also perceptual relationships between odorants. This organization reflects encoding in which receptor structure gives rise to functional groupings that align with both molecular features and human odor perception. This finding suggests that perceptual similarity may emerge, at least in part, from the structure of receptor activation patterns encoded within the OR binding cavity.

Feature Importance

While the smeLLMap captures the global organization of receptor-odorant relationships, we next sought to identify which specific structural features most strongly influenced individual model predictions. To this end, we applied Integrated Gradients (IG), which attributes a model's output to input voxels by measuring how the prediction changes as structural information is gradually introduced from a baseline representing the absence of receptor structure. Voxels with large attribution magnitudes, either positive or negative, exert the strongest influence on the

prediction, yielding a three-dimensional importance map over the binding cavity that highlights features that promote or disfavor ligand binding (Fig. 6A). Importantly, each voxel in our representation encodes both spatial location and residue-level biochemical features. As a result, voxel attribution reflects the joint contribution of three-dimensional positioning and physicochemical identity, rather than isolating either factor in abstraction. This integrated representation is advantageous, as ligand recognition in ORs depends on the precise spatial arrangement of chemically specific residues within the binding pocket. Attribution scores therefore capture how geometry and chemistry act together to shape receptor selectivity (Fig. 6A).

Notably, when attributions were aggregated across multiple receptors responding to the same odorant, importance did not simply track residue identity. Instead, the structural conformation of those residues appeared to influence their contribution to ligand binding. The same residue position could contribute positively or negatively depending on whether its side chain projected inward to constrict the cavity or outward to expand it, likely excluding or permitting ligand access (Fig. 6A). Together, these results indicate that odorant selectivity is governed not only by which residues are present, but by how their three-dimensional positioning modulates cavity accessibility.

We next asked whether this combined embedding and voxel-level attribution framework could be used to guide odorant-receptor activation. As a case study, we focused on tert-butyl mercaptan (TBM), a sulfur-containing odorant widely used as a natural gas additive due to its strong detectability.

Guided by the IG attributions, we engineered targeted mutations in Or2t43 and assessed their responses to TBM (Fig. 6B). Although Or2t43 is robustly activated by TBM in its native form, mutation of either C5.42 or C5.43 resulted in a complete loss of activation, consistent with disruption of the predicted sulfur-binding coordination site. In contrast, substitution of D4.53 with alanine, reducing side-chain size and polarity, or with asparagine, maintaining steric occupancy while altering charge, both increased TBM responsiveness to TBM. These reciprocal effects are consistent with the directionality of the model's attribution scores and indicate that the highlighted residues exert functionally meaningful influence on ligand sensitivity. In addition, the model identified residue D4.53 as negatively associated with TBM responsiveness, suggesting a potential inhibitory role at this position.

To further examine the structural basis of these effects, we generated AlphaFold3 models of the corresponding Or2t43 mutants. Comparison of the predicted structures revealed that mutation of D4.53 to asparagine results in a repositioning of the side chain within the binding cavity. Notably, the mutated N4.53 side chain extends into a spatial region immediately adjacent to, but distinct from, the original asparagine position, consistent with the observed increase in TBM responsiveness. These structural changes align with the attribution-derived importance map, providing support that local rearrangements within the binding cavity can modulate ligand sensitivity in a predictable manner.

Together, these results demonstrate that our interpretable Spatial LLM framework not only captures the global organization of odor-receptor relationships but also provides mechanistic insight at residue-level resolution. By linking structural context, learned

embeddings, and voxel-level attribution, the model enables rational prediction of functionally relevant mutations and supports targeted modulation of receptor sensitivity to specific odorants.

Discussion

This study demonstrates that OR-ligand selectivity can be learned and predicted through a structure-centered framework that explicitly models the three-dimensional organization and sequence-informed composition of the receptor binding cavity. By integrating in vivo pS6-IP-Seq activation measurements with spatial learning, we show that the canonical binding cavity contains sufficient information to organize receptor responses across chemically diverse odorants (Fig. 3B,D and 4D). These findings indicate that the determinants of ligand selectivity are encoded within the binding pocket itself, even without explicit modeling of alternative or allosteric interaction sites, thereby addressing a long-standing challenge in linking receptor variation to interpretable chemical recognition.

A key conceptual advance of this work is the demonstration that functional similarity among ORs is better captured by learned structural embeddings than by primary sequence similarity alone (Fig. 4B). While traditional sequence-based metrics and protein language model embeddings reflect evolutionary relationships, they fail to fully explain why distantly related receptors can converge on similar ligand selectivity. In contrast, the CNN-derived embedding organizes receptors according to shared response profiles, revealing functional neighborhoods that spans beyond phylogenetic relationship. This result suggests that ligand selectivity is governed by higher-order structural constraints, such as cavity geometry and residue positioning, that are not apparent from sequence comparisons.

Notably, this organization extends beyond chemical similarity to reflect perceptual relationships between odorants (Fig. 5A-D and fig. S6). Despite being trained without explicit ligand features, the learned embedding correlates with perceptual distances derived from OpenPOM (45), a model trained on chemical molecules to predict odor perception, whereas amino acid sequence-based distances do not (Fig. 5B). This correspondence suggests that receptor binding cavity encode features that can be used to infer odor perceptual similarity. At the same time, these comparisons should be interpreted with appropriate caution. The receptor structures and activation data in this study are derived from mouse, whereas OpenPOM reflects human perceptual judgments. Rather than implying a direct mapping between mouse receptor activity and human perception, this result points to shared, cross-species constraints in how chemical features are transformed into receptor-level representations. Such constraints may arise from conserved structural principles governing receptor-ligand interactions. Future work integrating human OR repertoires, alongside matched perceptual datasets, will be important to determine the extent to which receptor-level features directly encode perceptual dimensions. Establishing this link will help clarify whether the observed alignment reflects a generalizable organizing principle or emerges from partial overlap between species-specific receptor spaces.

Importantly, the interpretability of this representation allows functional predictions to be traced back to specific structural features (Fig. 6A-C). Integrated Gradients attribution revealed that both positively and negatively contributing voxels shape ligand responsiveness, highlighting residues that promote binding as well as those that disfavor activation. The observation that identical residues can exert opposite effects depending on their three-dimensional orientation underscores that odorant recognition is fundamentally a spatial phenomenon (Fig. 6A). These

results support a model in which receptor tuning arises from the precise arrangement of residues within the binding cavity, rather than from residue identity alone.

Given the dynamic nature of OR activation, the ability of static AlphaFold3 predicted structures to support prediction and attribution is notable. While receptor conformational and ligand-induced rearrangements undoubtedly contribute to signaling, our results indicate that determinants of selectivity are already encoded in the active-state architecture of the binding pocket structure and ESM features (Fig. 3E). Although this framework does not explicitly model OR secondary or allosteric binding sites, the predictive performance suggests that such interactions, when present, either act through modulation of the canonical cavity or play a secondary role in determining primary ligand selectivity. Alternatively, the current framework may also suggest the frequency of binding outside of canonical binding cavity occur in relatively limited subset of receptors such that the overall statistical framework remains significant. Additionally, by encoding amino acid features with ESM embeddings may also enables the feature to capture long-range dependencies that could be outside of the selected binding cavity itself (fig. S2). Future incorporation of receptor dynamics, alternative conformational states, or ligand-induced structural shifts may further refine predictions, but may not be required to recover core selectivity principles.

Beyond mechanistic insight, this study provides a scalable strategy for interrogating OR function at a systems level. The use of pS6-IP-Seq, coupled with an activation metric designed to minimize transcript abundance and variance biases, enabled broad and relatively unbiased sampling of receptor-odorant interactions in vivo (Fig. 1A-F). The resulting dataset supports a model that generalizes across receptors and odorants, including cases with sparse activation data

(Fig. 3D), highlighting the importance of carefully designed functional metrics in enabling downstream computational modeling.

While the present study focuses on single odorants at fixed concentrations, several extensions could substantially broaden its scope. Expanding chemical coverage would improve representation of sparsely sampled ligand classes, while incorporating concentration dependence could reveal how cavity features govern sensitivity and dynamic range. Natural olfactory stimuli are typically complex mixtures (4, 19), and extending the framework to model mixture responses or competitive interactions between ligands represents an important next step. Although caution is warranted when interpreting pS6-IP-Seq data, as odorants used for stimulation may be metabolized within the olfactory mucus before binding to ORs (51), integrating explicit ligand representations alongside receptor structures may further enable dissection of receptor-ligand complementarity and cross-reactivity.

The implications of this work extend beyond olfaction. ORs belong to the GPCR superfamily, which encompasses a wide range of receptors involved in sensory perception, neurotransmission, and pharmacological signaling. Demonstrating that structural features alone can organize receptor function in a highly diverse GPCR family suggests that similar approaches may be applicable to other receptors lacking extensive experimental characterization. In this context, the framework introduced here provides a generalizable strategy for linking receptor structure to ligand selectivity in an interpretable and predictive manner.

In summary, the central of this study is that OR function is encoded in the three-dimensional structure of the binding cavity in a way that can be learned, interpreted, and

experimentally validated. By incorporating sequence-centric descriptions to structure-informed representations, this work contributes toward a predictive, mechanistic understanding of chemical sensing. As structural models, functional data, and computational methods continue to converge, this approach lays the groundwork for quantitative and ultimately generative models of olfactory coding.

5

Materials and Methods

Phosphorylated S6 ribosomal immunoprecipitation sequencing (pS6-IP-Seq)

Mice used for pS6-IP were ~3 weeks old, mixed sex, and littermates. Mice were killed by CO₂ asphyxiation and cervical dislocation. Olfactory tissue was rapidly dissected in Buffer B (2.5 mM HEPES KOH pH 7.4, 0.63% glucose, 100 µg/mL cycloheximide, 5 mM sodium fluoride, 1 mM sodium orthovanadate, 1 mM sodium pyrophosphate, 1 mM β-glycerophosphate, in Hank's balanced salt solution). Tissue pieces were then minced in 1.35 mL Buffer C (150 mM KCl, 5 mM MgCl₂, 10 mM HEPES KOH pH 7.4, 0.100 µM Calyculin A, 2 mM DTT, 100 U/mL RNasin, 100 µg/mL cycloheximide, protease inhibitor cocktail, 5 mM sodium fluoride, 1 mM sodium orthovanadate, 1 mM sodium pyrophosphate, 1 mM β-glycerophosphate) and subsequently transferred to homogenization tubes for steady homogenization at 250 rpm three times and at 750 rpm nine times at 4 °C. Samples were then transferred to a 1.5 mL LoBind tube (Eppendorf 022431021) and clarified at 2000xg for 10 min at 4 °C. The low-speed supernatant was transferred to a new tube on ice, and 90 µL of NP40 (Sigma 11332473001) and 90 µL of 1,2-diheptanoyl-sn-glycero-3-phosphocholine (DHPC, Avanti Polar Lipids 850306P, 100 mg/0.69 mL) were added to this solution. This solution was mixed and then clarified at a max speed (17,000xg) for 10 min at 4 °C. The resulting high-speed supernatant was transferred to a new tube where 20 µL was saved and transferred to a tube containing 350 µL buffer RLT. To the remainder of the sample, 1.3 µL of 100 µg/mL cycloheximide, 27 µL of phosphatase inhibitor cocktail (250 mM sodium fluoride, 50 mM sodium orthovanadate, 50 mM sodium pyrophosphate, 50 mM β-glycerophosphate) and 6 µL of anti-pS6 antibody (Cell Signaling

D68F8) were added. The sample was gently rotated for 90 min at 4 °C. To prepare beads, 100 µL of beads (Invitrogen 10002D) was washed three times with 900 µL of buffer A (150 mM KCl, 5 mM MgCl₂, 10 mM HEPES KOH pH 7.4, 10% NP40, 10% BSA), and once with 500 µL of buffer C. Sample homogenate was added to the beads and incubated with gentle rotation for 60 min at 4 °C. Following incubation, beads were washed with four times with 700 µL of buffer D (350 mM KCl, 5 mM MgCl₂, 10 mM HEPES KOH pH 7.4, 10% NP40, 2 mM DTT, 100 U/mL RNAsin, 100 µg/mL cycloheximide, 5 mM sodium fluoride, 1 mM sodium orthovanadate, 1 mM sodium pyrophosphate, 1 mM β-glycerophosphate). During the final wash, beads were moved to room temperature, wash buffer was removed, and 350 µL of buffer RLT was added. Beads were incubated in buffer RLT for 5 min at room temperature. Buffer RLT containing immunoprecipitated RNA was then eluted and stored at −80 °C until clean up using a kit (Qiagen 74004). cDNA was generated using 11 rounds of amplification with 10 ng RNA input. DNA libraries were prepared using a half-sized Nexterra XT DNA Library Preparation Kit (Illumina 15032354) protocol as per the manufacturer's guidelines. Libraries were sequenced on either HiSeq 2000/2500 (50 base pair single read mode) or NextSeq 500 (75 base pair single read mode) with 6–12 pooled indexed libraries per lane.

RNA-Seq alignment, quantification, and differential expression analysis

Reads were aligned against a modified GRCm38.p6 (M25) reference, in which we deleted ENSMUSG00000116179 (Olfcr290), using STAR (52) with --outFilterMultimapNmax 10. Reads mapping to Olfcr290 were inferred from ENSMUSG00000070459, with the rationale that this gene model included ENSMUSG00000116179 plus untranslated regions. Gene-level

read quantification was done using RSEM (53). Differential expression analysis was performed against all genes using EdgeR (54). Gene nomenclature was retrieved from BioMart (55). Intact *Olf* genes with identifiable sequences were filtered, and p-values were then re-corrected by FDR. Only ORs exhibiting odor response to at least one of the tested odorants ($\log_2FC > 0$ and FDR < 0.05) were considered. A total of 555 ORs responded across the 72 different odorants at various concentrations. A total of 375 ORs were responsive to unique odorants at the lowest tested concentrations. Raw and processed RNA-Seq datasets generated as part of this study are available from NCBI GEO at accession GSE185415.

Source of odorants

The following odors and concentrations were used for molecular response profiling and model training: 1% p-Cresol (Sigma C85751), 1% 2-phenylethanol (Sigma 77861), 1% guaiacol (Sigma G10903), 10% (+)-2-octanol (Sigma O4504), 10% (-)-2-octanol (Sigma 147990), 1% linalool (Sigma L2602), 1 M (+)-menthol (Sigma 224464), 1 M (-)-menthol (Sigma M2780), 1% citronellol (Sigma W230915), 1% 2-methyl-2-pentenal (Sigma 294667), 1% benzaldehyde (Sigma W212717), 1% heptanal (Sigma W254002), 1% octanal (Sigma O5608), 1% trans-cinnamaldehyde (Sigma C80687), 0.01% anisaldehyde (Sigma A88107), 0.01% citral (Sigma W230316), 1% ethyl butyrate (Sigma W242713), 1% isoamyl acetate (Sigma 306967), 1% methyl salicylate (Sigma W274502), 1% diacetyl (Sigma W237027), 1% 2-hexanone (Sigma 103004), 1% 2-heptanone (Sigma 537683 (41)), 0.01% acetophenone (Sigma W200910), 1% 4-methylacetophenone (Sigma W267708), 1% (+)-carvone (Sigma 22070), 1% (-)-carvone (Sigma 22060), 1% β -damascone (Sigma W324300), 1% β -ionone (Sigma W259525), 1% pyridine

(Sigma 270970), 1% 2,5-dimethylpyrazine (Sigma 175420 (41)), 1% 2-ethyl-3-methylpyrazine (Sigma W315508), 0.01% 2-butene-1-thiol (CheMall Corp OR116574), 1% tert-butyl mercaptan (2-methyl-2-propanethiol; Sigma 109207), 0.01% cyclopentanethiol (Sigma W326208), 1% 3-methyl-1-butanethiol (Sigma W385808), 1% dimethyl trisulfide (Sigma W327506), 1% 2-methyl-2-thiazoline (Sigma M83406), 1% 2,4,5-trimethylthiazole (nTMT; Sigma 219185), 0.01% 2,4,5-trimethyl-4,5-dihydrothiazole (TMT, synthesized (56)), 0.01% 2-sec-butyl-4,5-dihydrothiazole (SBT; synthesized (41)), 1% ethyl tiglate (Sigma W246000), 1% isopropyl tiglate (Sigma W322903), and 1% hexyl tiglate (Sigma W500909).

Chemical space estimation

To estimate chemical space, we first identified 4680 small molecules commonly found in foods and fragrances from <http://www.thegoodscentscompany.com/> (43). Smiles strings for these molecules and the 48 in the test odor set were then downloaded from PubChem, and Morgan-algorithm based fingerprints were calculated using Rdkit (44, 57). Chemical space was estimated by PCA dimensionality reduction on all molecules.

To complement this structure-based representation with a perceptual framework, we additionally embedded all molecules using OpenPOM(27, 45), a message-passing neural network trained to predict human odor perception from chemical structure. The resulting OpenPOM feature vectors provide a high-dimensional representation of perceptual odor space derived from human panel data (58). These embeddings were used to assess perceptual diversity of the odor panel, perform pairwise distance comparisons, and relate receptor embedding organization to perceptual similarity in downstream analyses.

Receptor alignment and space estimation

Mouse ORs were aligned to one another using the MAFFT E-INS-I method with manual refinements (59). The resulting alignment file was subjected to ModelTest-NG to identify ideal amino acid substitution models (60). Receptor pairwise similarity was calculated by summing amino acid differences at each position by Grantham's amino acid distances (61). Multidimensional scaling on all ORs.

Structural Model Generation

All mouse olfactory receptor (OR) protein sequences were compiled together with mini-Gα peptide definitions into standardized JSON configuration files. For each receptor, five independent structural predictions of the OR–miniGα complex were generated using AlphaFold-Multimer (13), capturing variability in receptor conformation under identical modeling conditions. In parallel, five ORs with respective ligands with available cryo-EM structures (OR51E2, consOR1, consOR2, consOR4, consOR51) were processed through the same workflow to serve as experimentally grounded structural references. Following structural prediction, all OR complexes were placed into a common coordinate frame using TM-align (62), ensuring consistent spatial orientation of transmembrane helices and intracellular domains across receptors. Internal solvent-accessible pockets were then identified using pyKVFinder (49), which outputs cavity coordinates. This procedure was applied uniformly to predicted models, cryo-EM structures, and later to ligand-bound complexes.

Canonical Binding Cavity Identification

To determine the biologically relevant ligand binding region, we incorporated receptor–odorant activation data derived from pS6-IP assays. For each experimentally activated receptor–ligand pair, AlphaFold3 was used to model receptor–miniGα–ligand complexes, followed by
 5 cavity detection with pyKVFinder. Ligand-occupied pockets were spatially compared across cryo-EM receptors, and regions demonstrating $\geq 50\%$ overlap across ligand-bound cavities were defined as the canonical binding cavity. This region represents the reproducibly utilized ligand-engagement space across structurally diverse ORs.

The canonical binding cavity mask was then projected back onto ligand-free OR models
 10 for the entire repertoire. For each receptor, pyKVFinder-defined cavities were filtered relative to the canonical binding cavity; regions with $< 80\%$ spatial overlap were discarded. This filtering process preserved a single, standardized binding cavity per receptor while removing allosteric, non-functional or receptor-specific peripheral pockets. Individual receptor’s cavity surrounding amino acid then defined by less than 5 Å. The resulting structures consist of coordinates of each
 15 OR binding cavity and amino acid coordinates.

Voxelization and Feature Encoding

For each receptor, all residues spatially lining the canonical binding cavity were extracted to construct a localized structural representation. The cavity volume of each receptor was then discretized into a $32 \times 32 \times 32$ voxel grid, maintaining 1 Å spatial resolution, such that each
 20 voxel encodes a fixed physical region of three-dimensional space surrounding the binding

pocket. For every voxel containing protein atoms, the residue occupying that spatial position was assigned as the voxel occupant. To provide a rich biochemical and evolutionary descriptor beyond atomic coordinates, the reduced ESM sequence embedding vector corresponding to that residue was inserted as the voxel feature representation. Voxels not occupied by residues were assigned a null background vector. This procedure resulted in a structured, spatially aligned, and biologically informed voxel map for each receptor. In this final representation, every OR is encoded as a voxelized binding cavity populated with residue-level embedding features, providing a standardized structural input for downstream convolutional neural network training and odor selectivity modeling.

Heterologous luciferase assay

Hana3A cells; which stably express G_{olf} , RTP1, RTP2, and REEP1; were grown in minimum essential medium eagle (MEM; Corning 10-010-CV) containing 10% Fetal Bovine Serum (FBS; vol/vol; Gibco 16000-044), penicillin-streptomycin (Sigma-Aldrich P4333), and amphotericin B (Gibco 15290018). Cells were cultured and incubated at 37°C, 5% CO₂, and saturated humidity for use with the Dual-Glo Luciferase Assay (Promega E2980)(63, 64). Cells were plated at 20-25% confluence on poly-D-lysine-coated 96-well plates (Corning 3843) overnight. After overnight incubation, cells were transfected with 6 mL of MEM containing 10% FBS, 0.5 µg SV40-RL (Promega E2980), 1 µg CRE-Luc (Promega E2980), 0.5 µg mouse RTP1s, 0.25 µg M3 muscarinic receptor, 0.5 µg of Rho-tagged receptor plasmid DNA, and 20 µg Lipofectamine 2000 (Invitrogen 11668019) per plate. Transfection medium was divided equally among the wells so that each OR-odorant combination could be conducted in triplicates.

The following day, cells were incubated with 25 μ L of odorant solution diluted in CD-293 (Gibco 11913-019) containing 30 μ M CuCl₂ (Sigma-Aldrich C-6641) and 2 mM glutamine (Gibco 25030-081) for 3.5 hours. cAMP-driven firefly Luciferase luminescence (Luc) was used to assess OR activation, and SV40-driven *Renilla* Luciferase luminescence (Ren) was used to control for variation in cell viability within wells. Cell luminescence was read by a POLARstar OPTIMA (BMG Labtech) luminometer, and normalized response values were calculated using the formula (Luc-400)/(Ren-400). ORs were considered responsive *in vitro* if ANOVA *p*-value was < 0.05 and ANOVA with post-hoc Dunnet's test correction *p*-adjusted was < 0.05 for at least 2 of the tested odor concentrations using the R package DescTools (v0.99.42). Log-logistic 4-parameter dose response curves were fit to the data using the R package drc (v3.0-1). *In vitro* responses were compared to *in vivo* responses by subtracting mean ligand-independent activity (luciferase values of ORs with no odor stimulation) from each of the ligand stimulated data points and summing. Scaled summed (+)-enantiomer responses were divided by scaled summed (-)-enantiomer responses and log₂ transformed for comparison to log₂FC (+)/(-) *in vivo* enrichments.

Data analysis

Analysis were performed in Python 3.8. Custom scripts for data analysis and visualization built using open-source python libraries (pandas, numpy, matplotlib, plotly, sklearn, scipy, seaborn and itertools). Scripts to replicate data analysis are available on github (https://github.com/Justice-Lu/OR_learning).

Statistical analysis

Experimental data are presented as mean $\pm$ standard deviation (SD), based on results from at least three independent experiments. Statistical analysis was conducted in Python 3.11.13 using SciPy v1.16.0. Paired comparisons (e.g., samples with different condition measured in the same experiment) were analyzed using the paired Student's t-test (`ttest_rel`). For unpaired comparisons between two independent groups, the non-parametric Wilcoxon rank-sum test (`ranksums`) was applied. A table with all the statistical comparison values and description can be found at Table 3. A two-tailed p-value less than 0.05 was considered statistically significant. Significance levels are denoted as follows: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p \geq 0.05$ (n.s., not significant).

References and Notes

1. C. Bushdid, M. O. Magnasco, L. B. Vosshall, A. Keller, Humans Can Discriminate More than 1 Trillion Olfactory Stimuli. *Science* **343**, 1370-1372 (2014).
2. L. Buck, R. Axel, A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175-187 (1991).
3. Y. Niimura, A. Matsui, K. Touhara, Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res* **24**, 1485-1496 (2014).
4. T. Tsukahara *et al.*, A transcriptional rheostat couples past activity to future sensory responses. *Cell* **184**, 6326-6343.e6332 (2021).
5. C. B. Billesbølle *et al.*, Structural basis of odorant recognition by a human odorant receptor. *Nature* **615**, 742-749 (2023).
6. C. A. de March *et al.*, Engineered odorant receptors illuminate the basis of odour discrimination. *Nature* **635**, 499-508 (2024).
7. C. Choi *et al.*, Understanding the molecular mechanisms of odorant binding and activation of the human OR52 family. *Nat Commun* **14**, 8105 (2023).
8. X. Han *et al.*, Mechanistic insights into fatty acid odor detection mediated by class II olfactory receptors. *Cell*, (2026).
9. X.-Y. Ge *et al.*, Identification of Or5v1/Olfr110 as an oxylipin receptor and anti-obesity target. *Cell*, (2026).
10. T. Wang *et al.*, Structural decoding of reversible covalent linkage of odorants in human olfactory receptor OR6A2. *Cell*, (2026).
11. M. Gil *et al.*, Structural basis of odorant recognition by a mammalian class II odorant receptor. *Sci Adv* **12**, eab9026 (2026).
12. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
13. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
14. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
15. G. Wald, The Receptors of Human Color Vision: Action spectra of three visual pigments in human cones account for normal color vision and color-blindness. *Science* **145**, 1007-1016 (1964).
16. J. Schnapf, T. Kraft, D. Baylor, Spectral sensitivity of human cone photoreceptors. *Nature* **325**, 439-441 (1987).
17. L. Buck, R. Axel, A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175-187 (1991).
18. T. Bozza *et al.*, Mapping of class I and class II odorant receptors to glomerular domains by two distinct types of olfactory sensory neurons in the mouse. *Neuron* **61**, 220-233 (2009).
19. L. F. Horgue *et al.*, Transcriptional adaptation of olfactory sensory neurons to GPCR identity and activity. *Nature Communications* **13**, 2929 (2022).
20. J. del Marmol, M. A. Yedlin, V. Ruta, The structural basis of odorant recognition in insect olfactory receptors. *Nature* **597**, 126-131 (2021).

21. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
22. Z. Lin *et al.*, Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022**, 500902 (2022).
- 5 23. J. Lötsch, D. Kringel, T. Hummel, Machine Learning in Human Olfactory Research. *Chem Senses* **44**, 11-22 (2019).
24. E. J. Mayhew *et al.*, Transport features predict if a molecule is odorous. *Proc Natl Acad Sci U S A* **119**, e2116576119 (2022).
- 10 25. R. M. Khan *et al.*, Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *J Neurosci* **27**, 10015-10023 (2007).
26. Y. Nozaki, T. Nakamoto, Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. *PLoS One* **13**, e0198475 (2018).
- 15 27. B. K. Lee *et al.*, A principal odor map unifies diverse tasks in olfactory perception. *Science* **381**, 999-1006 (2023).
28. S. Shuvaev, K. Tran, K. Samoilova, C. Mascart, A. Koulakov, in *2024 58th Asilomar Conference on Signals, Systems, and Computers*. (2024), pp. 536-543.
29. M. Schmuker, G. Schneider, Processing and classification of chemical data inspired by insect olfaction. *Proceedings of the National Academy of Sciences* **104**, 20285-20289 (2007).
- 20 30. K. Snitz *et al.*, Predicting odor perceptual similarity from odor structure. *PLoS computational biology* **9**, e1003184 (2013).
31. Y. Jiang *et al.*, Molecular profiling of activated olfactory neurons identifies odorant receptors for odors in vivo. *Nature neuroscience* **18**, 1446-1454 (2015).
- 25 32. A. Vihani *et al.*, Encoding of odors by mammalian olfactory receptors. *bioRxiv*, 2021.2012.2027.474279 (2021).
33. Zachary A. Knight *et al.*, Molecular Profiling of Activated Neurons by Phosphorylated Ribosome Capture. *Cell* **151**, 1126-1137 (2012).
- 30 34. P. Mombaerts, Odorant receptor gene choice in olfactory sensory neurons: the one receptor–one neuron hypothesis revisited. *Current opinion in neurobiology* **14**, 31-36 (2004).
- 35 35. S. Serizawa, K. Miyamichi, H. Sakano, One neuron-one receptor rule in the mouse olfactory system. *Trends Genet* **20**, 648-653 (2004).
36. A. Chess, I. Simon, H. Cedar, R. Axel, Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**, 823-834 (1994).
37. B. Malnic, J. Hirono, T. Sato, L. B. Buck, Combinatorial receptor codes for odors. *Cell* **96**, 713-723 (1999).
38. K. Touhara *et al.*, Functional identification and reconstitution of an odorant receptor in single olfactory neurons. *Proc Natl Acad Sci U S A* **96**, 4040-4045 (1999).
- 40 39. K. Monahan, A. Horta, S. Lomvardas, LHX2-and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565**, 448-453 (2019).
40. H. J. Shayya *et al.*, ER stress transforms random olfactory receptor choice into axon targeting precision. *Cell* **185**, 3896-3912.e3822 (2022).

41. A. Vihani *et al.*, Semiochemical responsive olfactory sensory neurons are sexually dimorphic and plastic. *Elife* **9**, e54501 (2020).
42. Y. Isogai *et al.*, Multisensory logic of infant-directed aggression by males. *Cell* **175**, 1827-1841. e1817 (2018).
- 5 43. A. Ravia *et al.*, A measure of smell enables the creation of olfactory metamers. *Nature* **588**, 118-123 (2020).
44. H. L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation* **5**, 107-113 (1965).
- 10 45. A. A. Barsainyan, R. Kumar, P. Saha, M. Schmuker. (2023).
46. H. He, E. A. Garcia, Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1263-1284 (2009).
47. T. Olender, T. E. Jones, E. Bruford, D. Lancet, A unified nomenclature for vertebrate olfactory receptors. *BMC Evolutionary Biology* **20**, 42 (2020).
- 15 48. W. Yeung *et al.*, Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Briefings in Bioinformatics* **24**, (2023).
49. J. V. d. S. Guerra *et al.*, pyKVFinder: an efficient and integrable Python package for biomolecular cavity detection and characterization in data science. *BMC Bioinformatics* **22**, 607 (2021).
- 20 50. E. Richardson *et al.*, The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* **5**, (2024).
51. C. Ijichi *et al.*, Metabolism of Odorant Molecules in Human Nasal/Oral Cavity Affects the Odorant Perception. *Chem Senses* **44**, 465-481 (2019).
- 25 52. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
53. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
54. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **26**, 139-140 (2010).
- 30 55. D. Smedley *et al.*, BioMart–biological queries made easy. *BMC genomics* **10**, 22 (2009).
56. X. S. Hu *et al.*, Concentration-dependent recruitment of mammalian odorant receptors. *eneuro* **7**, (2020).
- 35 57. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).
58. A. Keller, L. B. Vosshall, Olfactory perception of chemically diverse molecules. *BMC Neurosci* **17**, 55 (2016).
59. K. Katoh, K.-i. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**, 511-518 (2005).
- 40 60. D. Darriba *et al.*, ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular biology and evolution* **37**, 291-294 (2020).
61. R. Grantham, Amino acid difference formula to help explain protein evolution. *science* **185**, 862-864 (1974).

62. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).
63. H. Saito, M. Kubota, R. W. Roberts, Q. Chi, H. Matsunami, RTP family members induce functional expression of mammalian odorant receptors. *Cell* **119**, 679-691 (2004).
- 5 64. H. Zhuang, H. Matsunami, Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. *Nat Protoc* **3**, 1402-1413 (2008).

Acknowledgments:

10 We thank Mengjue Jessica Ni for expert technical assistance. We appreciate Rohit Singh, Kapil Devkota and John Pearson for helpful discussions on data analysis. We also thank Duke Sequencing and Genomic Technologies core for sequencing and processing services.

Funding:

15 National Institutes of Health grants DC022770, DC020353, DC016224, DC014423 (HM)

Author contributions:

Conceptualization: HL, AV, HM

Methodology: HL, AV, HM

Investigation: HL, AV, MN, XH, DT, CJ, CdM, HM

20 Visualization: HL, HM

Funding acquisition: HM

Project administration: HM

Supervision: HM

Writing - original draft: HL, AV, HM

25 Writing - review & editing: HL, AV, MN, XH, DT, CJ, CdM, HM

Competing interests:

H.M. has received royalties from Chemcom, research grants from Givaudan and consultant fees from Kao.

Data, code, and materials availability:

5 All code used for data processing, structural modeling, and machine learning is publicly available at GitHub: https://github.com/Justice-Lu/OR_learning/tree/master. Raw and processed RNA-sequencing data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE185415. Additional data supporting the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Materials

10 Materials and Methods

Supplementary Text

Figs. S1 to S6

Tables S1 to S2

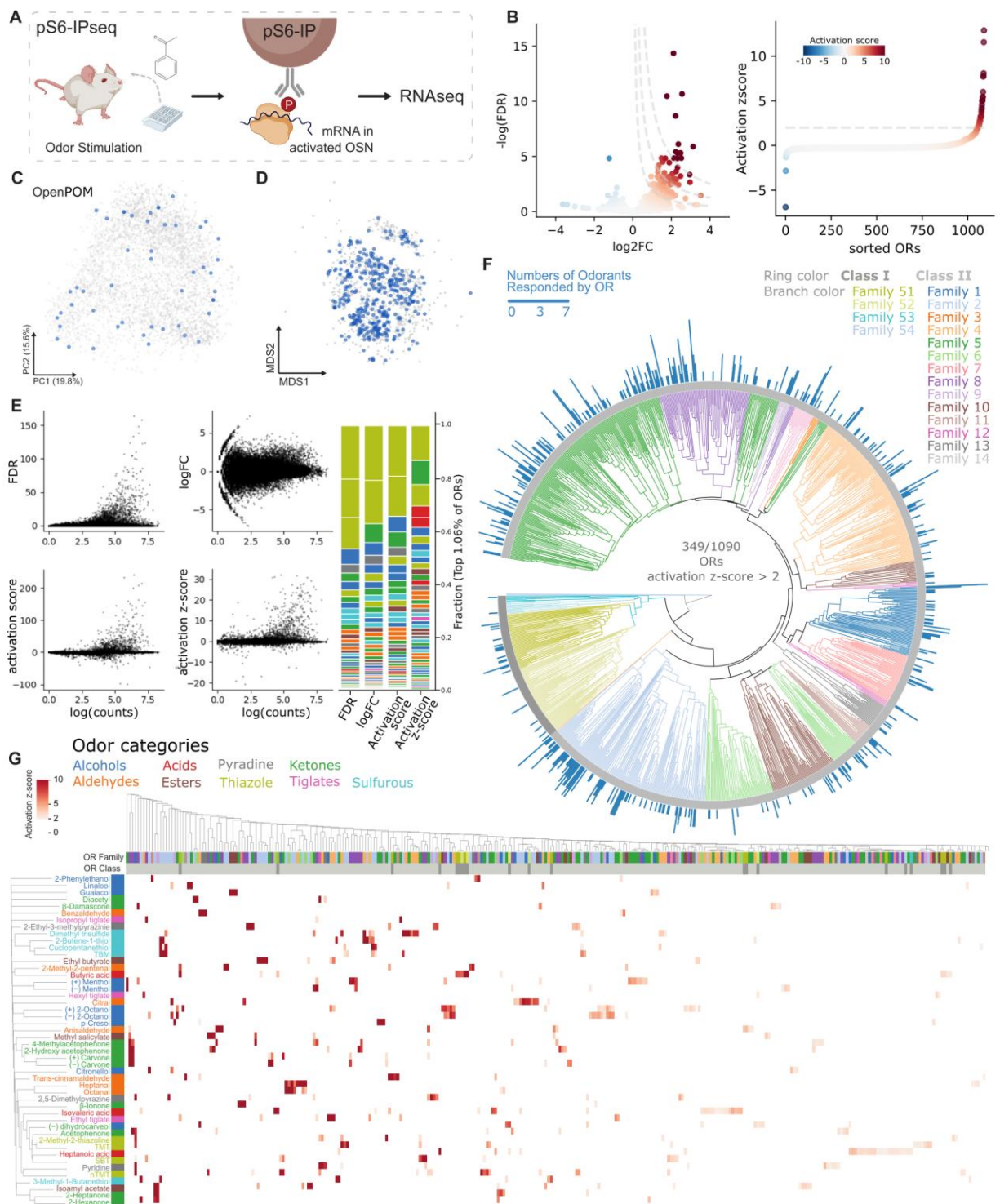


Figure 1. Defining odor-evoked olfactory receptor activation using pS6-IP

A. Schematic of the pS6-IP-Seq experimental workflow. Odor exposure induces phosphorylation of ribosomal protein S6 in activated olfactory sensory neurons (OSNs). pS6-IP-Seq enriches mRNAs expressed in activated OSNs, which are subsequently profiled by RNA sequencing to identify odor-responsive olfactory receptors (ORs). **B.** Quantification of OR activation in response to 100% acetophenone stimulation. Left: Volcano plot showing enrichment of OR mRNAs, with points colored by activation score. Right: ORs ranked by activation z-score, with colors indicating activation score. The horizontal dashed line denotes the activation z-score threshold (> 2) used to define activated receptors. **C.** Chemical perceptual space representations of tested odorants. Principal component analysis (PCA) of 4680 small molecules based on OpenPOM. The 48 odorants tested by pS6-IP, each activating at least one OR, are highlighted in blue. **D.** ORs space representations of responding receptors. Multidimensional scaling (MDS) visualization of intact ORs based on pairwise Grantham distances. ORs responding to at least one of the 48 odorants are shown in blue ($n = 349$). **E.** Scatter plots of different metric's (FDR, logFC, activation score, activation z-score) relationship biases with receptor counts. Right: Histogram showing the fraction of identified top 1% odor-ligand pairs selected using different response metrics. **F.** Circular phylogenetic tree of ORs, with branch colors indicating OR family and shaded ranges denoting OR class. The outer ring displays a bar plot representing the number of odorants to which each OR responds, as defined by activation z-score threshold > 2 . **G.** Heatmap summarizing activation z-scores across all 48 pS6-IP odor experiments with activation z-score threshold > 2 . Rows represent odorants and columns represent ORs, with hierarchical clustering applied to both dimensions.

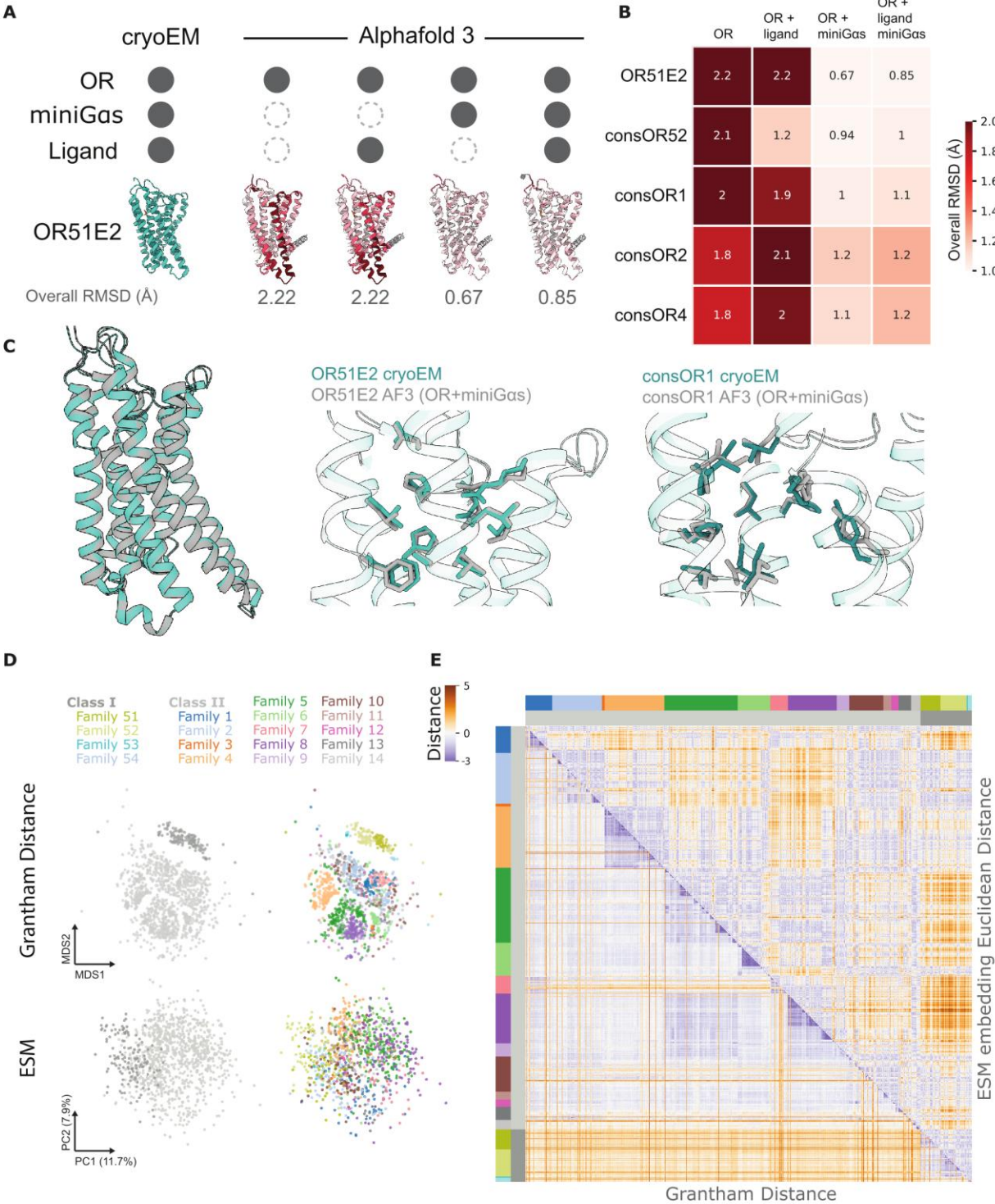


Figure 2. Structural accuracy of AlphaFold3 models and comparison of sequence- and language model-based receptor representations.

A. Structural comparison between AlphaFold3-predicted models and the cryo-EM structure of OR51E2. Columns show different multimer configurations used for structure prediction. Structures are colored by per-residue RMSD relative to the cryo-EM reference, with overall RMSD values reported below each model. B. Heatmap summarizing overall RMSD values between AlphaFold3 predictions and experimentally determined cryo-EM structures across five available OR structures. C. Structural overlays of predicted and experimental OR structures. Left: superposition of OR51E2 cryo-EM and AlphaFold3 (OR+miniG_{as}) models. Middle: Magnified view of binding cavity residues for OR51E2, comparing cryo-EM and AlphaFold3 models. Right: Magnified view of binding cavity residues for consOR1, comparing cryo-EM and AlphaFold3 models. D. Low-dimensional visualization of OR similarity based on sequence-derived and language model-derived representations colored by class and family. Top: Multidimensional scaling (MDS) of pairwise Grantham distances. Bottom: Principal component analysis (PCA) of ESM-2 sequence embeddings. E. Pairwise distance heatmap showing Grantham distance based similarities (lower triangle) and Euclidean distances computed from ESM-2 embeddings (upper triangle) across ORs.

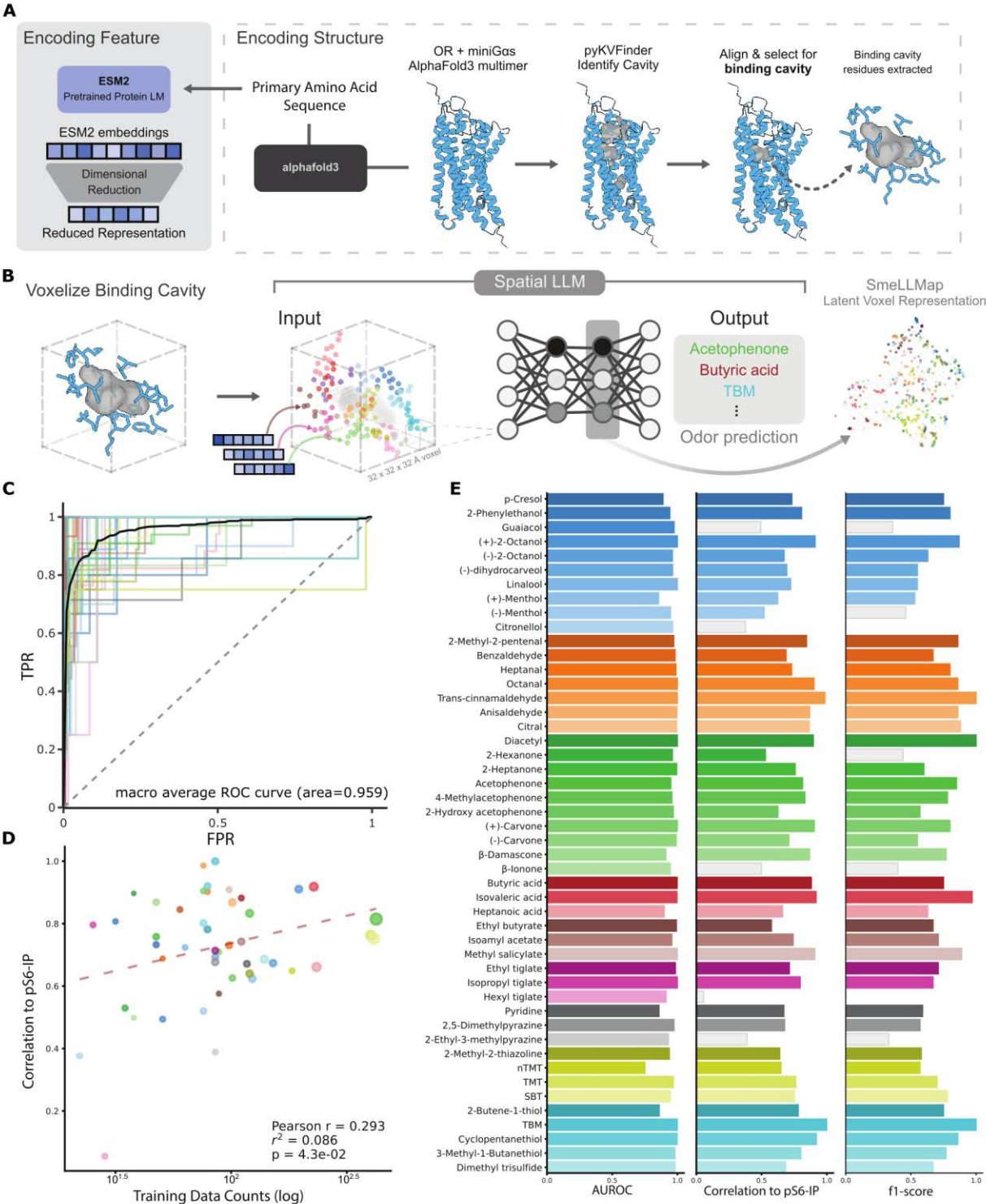


Figure 3. Spatial LLM architecture and model performance.

A. Schematic of the data pre-processing pipeline. For sequence encoding, full-length OR sequences were input to ESM-2 to generate residue-wise embeddings, which were subsequently reduced in dimensionality by principal component analysis (PCA). In parallel, the same sequences were used to predict OR structures (OR+miniGas) using AlphaFold3. Predicted structures were analyzed with pyKVFinder to identify internal cavities, which were filtered against predefined canonical binding cavity coordinates. Residues surrounding the retained binding cavity were extracted to generate a structural representation for each OR. **B.** Spatial LLM modeling pipeline. OR structural representations were voxelized into three-dimensional grids. Voxel occupancy encoded spatial structure, while reduced ESM-2 features were assigned to voxels based on residue identity, yielding a spatially and feature-aware representation of each OR. These voxel tensors were used as input to a convolutional neural network (CNN) trained to predict binary OR activation labels derived from pS6-IP activation z-scores. The output of the final fully connected layer was additionally extracted as a latent embedding, termed ‘smeLLMap’, representing learned receptor features. **C.** Receiver operating characteristic (ROC) curve summarizing CNN classification performance across all OR-odorant pairs with a macro-averaged curve plotted in black. **D.** CNN model prediction correlation to pS6-IP labels for individual odorants, plotted against the number of OR-ligand pairs in training data. Circle size is proportion to the number of test-set molecules. **E.** Model performance metrics evaluated per odorant, including area under the ROC curve (AUROC), correlation between prediction and pS6-IP activation values, and F1 score.

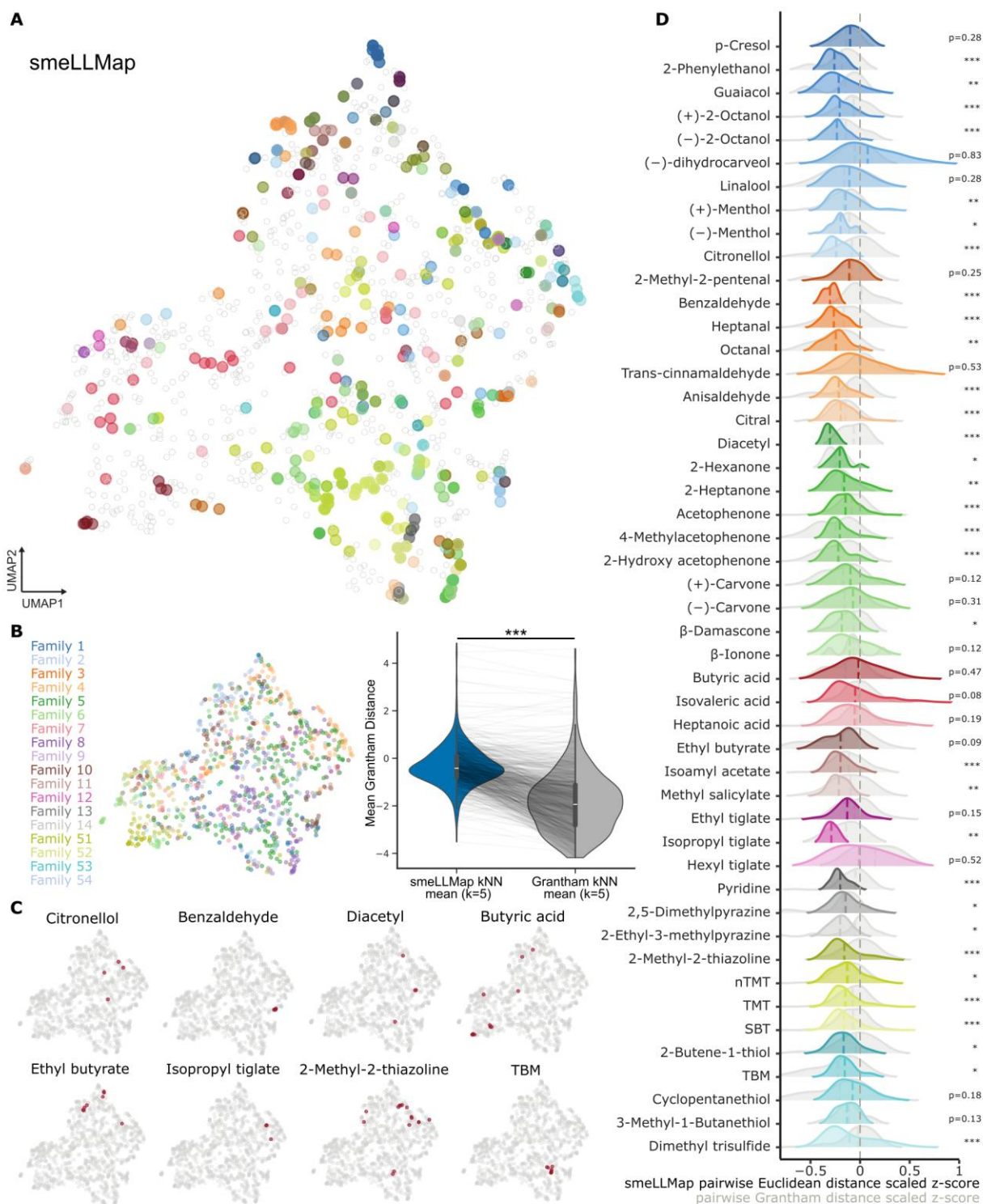


Figure 4. SmeLLMap captures odor-specific relationships among odorant receptors.

A. Left: smeLLMap visualization of OR embeddings using UMAP dimensionality reduction. Each point represents a single OR; receptors with no detected responses are shown in hollow grey, while responding receptors are overlaid by odorant colors to which they respond based on pS6-IP activation labels. Right: individual smeLLMap subplots highlighting ORs response to a single odorant. **B.** Left: smeLLMap representation of ORs by family. Right: Violin plots show Grantham distance between neighbor distributions across receptors in smeLLMap. line indicates matched random pairs. Paired t-test, $p = 1.05 \times 10^{-35}$. **C.** smeLLMap representation of individual ORs activated (labeled in red) by specific odorant clusters. **D.** Violin plots of pairwise Euclidean distances between CNN-derived embeddings of ORs responding to the same odorant. Statistical significance was assessed using permutation tests with randomly sampled OR groups matched for group size per odorants. Grey violins denote the corresponding pairwise Grantham distance distributions for the same OR sets.

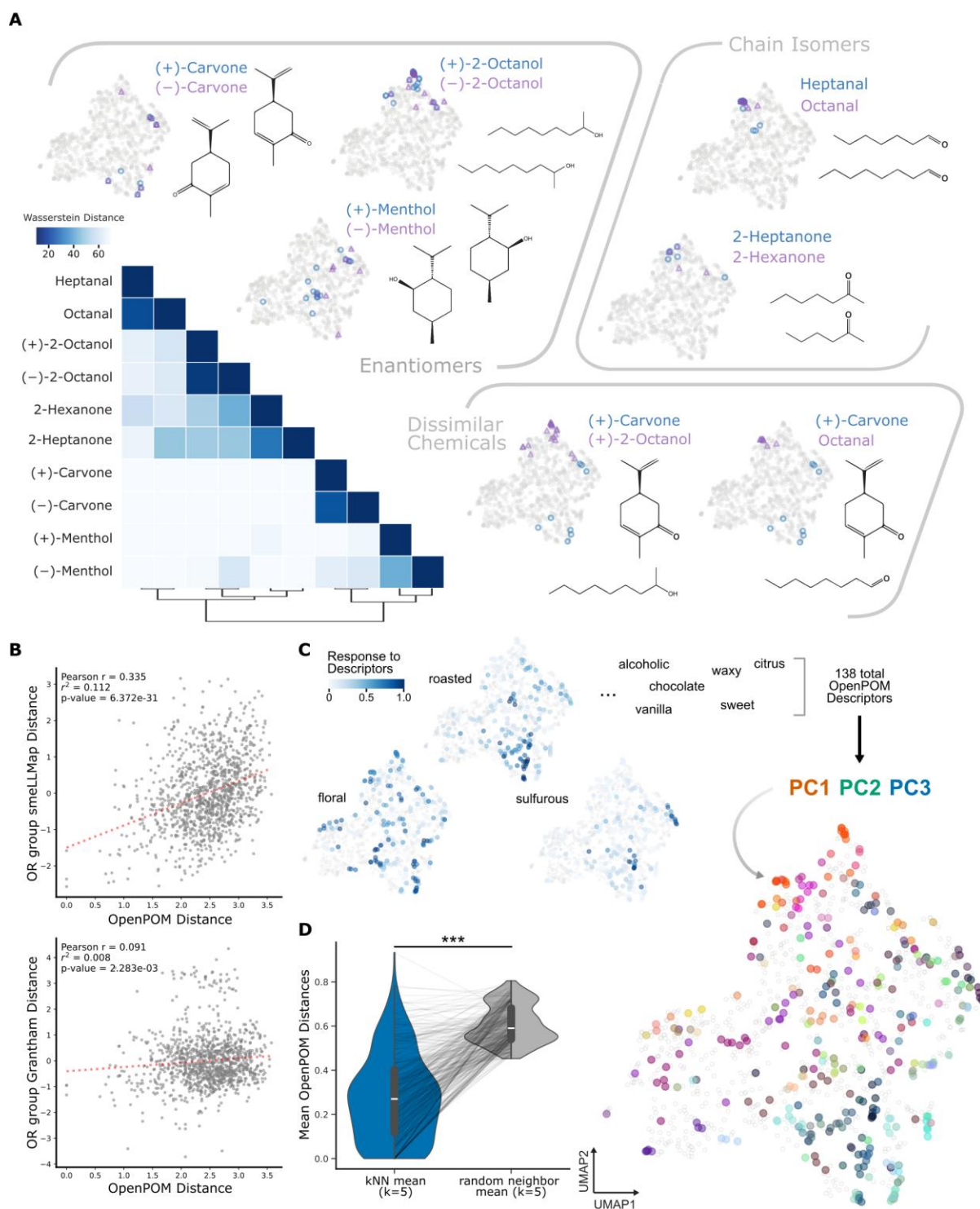


Figure 5. SmeLLMap captures chemical and perceptual similarity relationships among odorant receptors.

A. smeLLMap reflects chemical relationships between odorants. Left: Heatmap of Wasserstein distances between distributions of OR embeddings for each odorant, clustered by similarity. Right: Representative smeLLMap projections illustrating receptor distributions for selected odorant pairs, including enantiomers, structurally related compounds (heptanal vs. octanal; 2-heptanone vs. 2-hexanone), and chemically dissimilar odorants ((+)-carvone vs. octanal). **B.** Quantitative comparison between receptor embedding derived distances and perception-informed chemical space. Top: Pearson correlation between pairwise Wasserstein distances of responding OR groups in smeLLMap and corresponding Euclidean distances in OpenPOM space across all 48 odorants. Bottom: Equivalent analysis using Grantham sequence distances in place of embedding distances, showing reduced correspondence with OpenPOM-derived relationships. **C.** Mapping perceptual features onto receptor embedding space. Left: Selected OpenPOM descriptors (e.g., floral, roasted, sulfurous) projected onto smeLLMap, revealing spatially localized receptor tuning. Right: Integration of all 138 OpenPOM descriptors, reduced by principal component analysis and visualized as RGB color channels (PC1-PC3). **D.** Local neighborhood structure in embedding space reflects perceptual similarity. For each receptor, the mean OpenPOM Euclidean distance of its five nearest neighbors in smeLLMap is compared to a random expectation. Violin plots show distributions across receptors; line indicates matched random pairs. Paired t-test, $p = 6.46 \times 10^{-96}$.

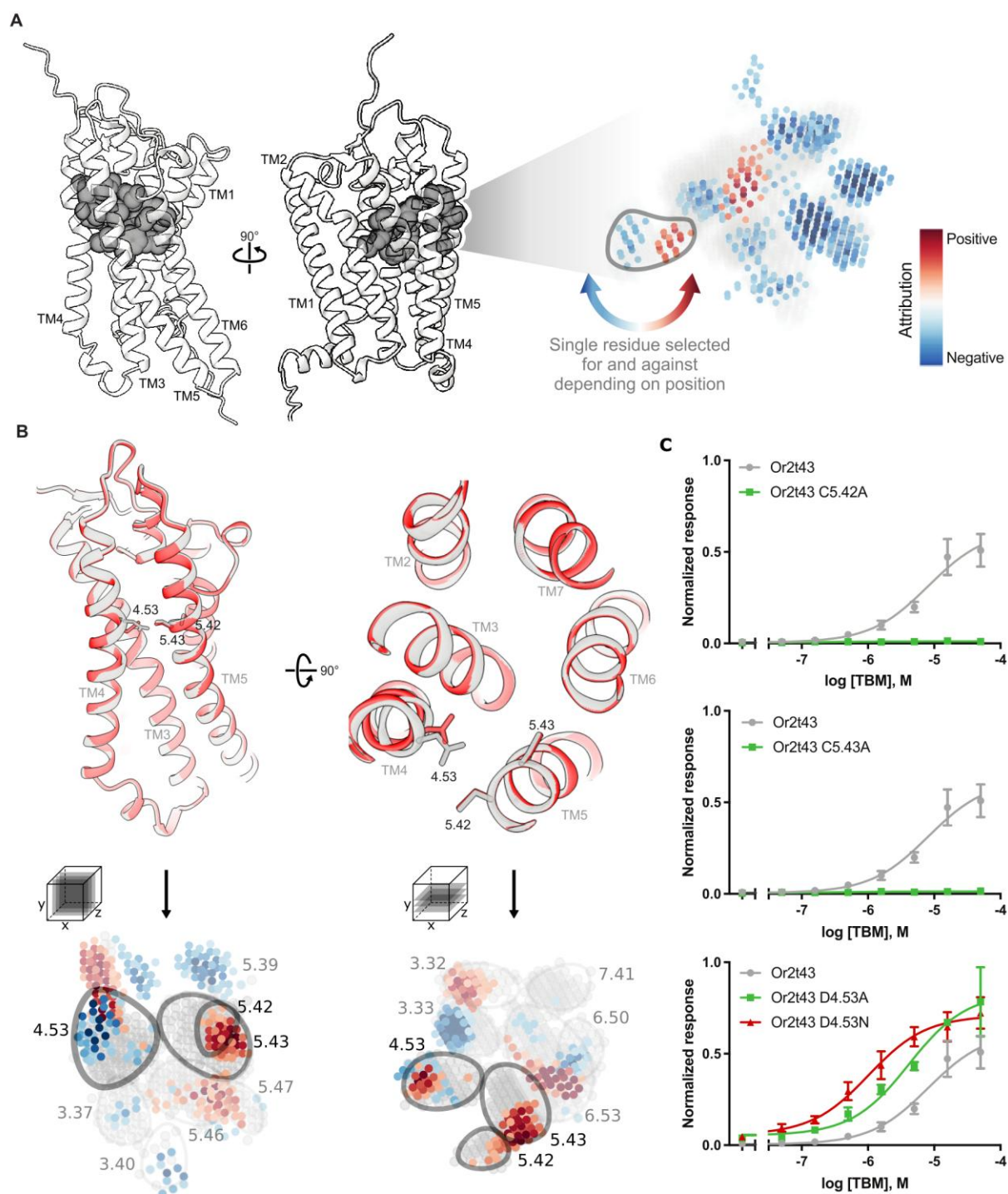


Figure 6. Voxel-level attribution identifies structural determinants of odor selectivity.

A. Structural and voxel-based representations of olfactory receptor binding cavities. Left, occupancy representation of amino acid residues defining the extracted canonical binding cavity mapped onto the OR structure, shown from two viewing angles. Right, corresponding three-dimensional voxel saliency map, highlighting spatial regions within the cavity that contribute most strongly to model predictions. **B.** Top: Predicted AlphaFold3 structures of native Or2t43 and Or2t43D4.43N mutant with side and top view orientation. Bottom: Stacks aggregated saliency maps for receptors responding to tert-butyl mercaptan (TBM). Left and middle, orthogonal slice views through the averaged voxel attribution map reveal conserved high-importance regions within the binding cavity. **C.** functional validation by mutagenesis of residues D4.53, C5.42, and C5.43 in Or2t43, assessed using a real-time cAMP luciferase assay. Data are shown as mean $\pm$ s.d. from $n = 3$ independent experiments.

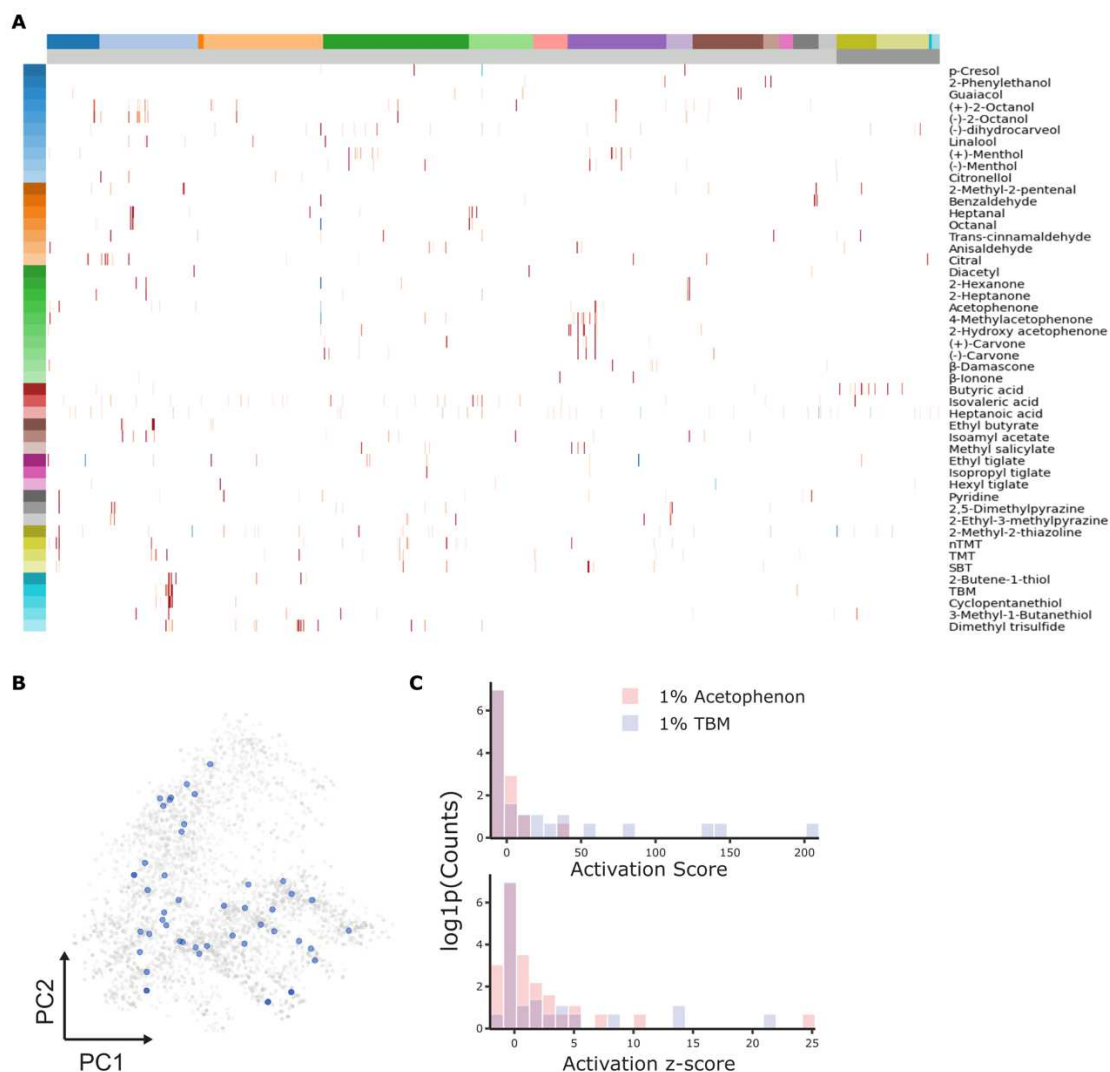


Figure S1. Global overview of pS6-IP activation profiles and chemical space coverage.

A. Heatmap of pS6-IP–derived activation z-scores across all tested odorant–receptor pairs. Rows correspond to the 48 odorants, and columns represent individual ORs, organized by class and family. **B.** Chemical space representation of odorants based on Morgan fingerprint principal component analysis (PCA). Odorants included in the pS6-IP panel are highlighted in blue relative to the full reference set.

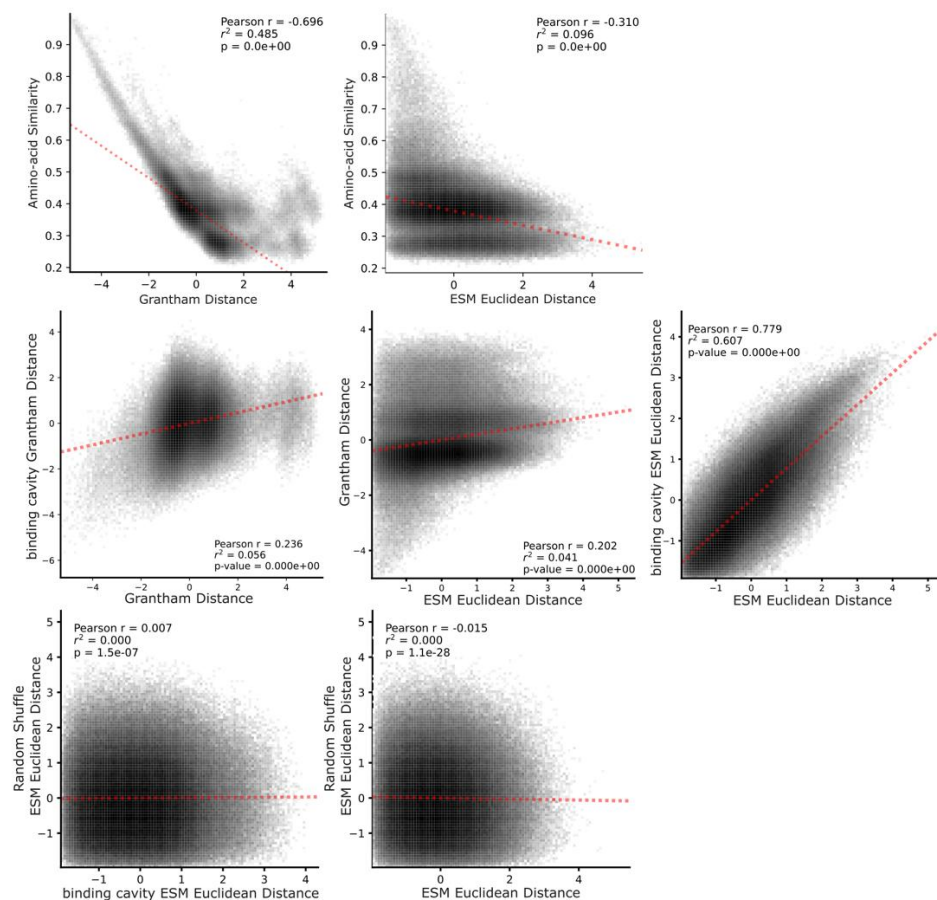


Figure S2. Comparison of sequence- and embedding-based similarity metrics across olfactory receptors.

Top: Pearson correlation analysis comparing amino acid similarity-based Grantham distances with ESM-derived Euclidean distances across OR pairs. Middle: Pairwise Pearson correlations among sequence- and structure-informed similarity measures, including full-length sequence Grantham distance vs. binding cavity Grantham distance, Grantham distance vs. ESM Euclidean distance, and binding cavity ESM Euclidean distance vs. full-sequence ESM Euclidean distance. Bottom: Control comparisons against randomized embeddings. Pearson

correlations between binding cavity ESM Euclidean distances and randomly shuffled ESM embeddings, and between full-sequence ESM Euclidean distances and shuffled controls, demonstrating preservation of biological signal in learned representations.

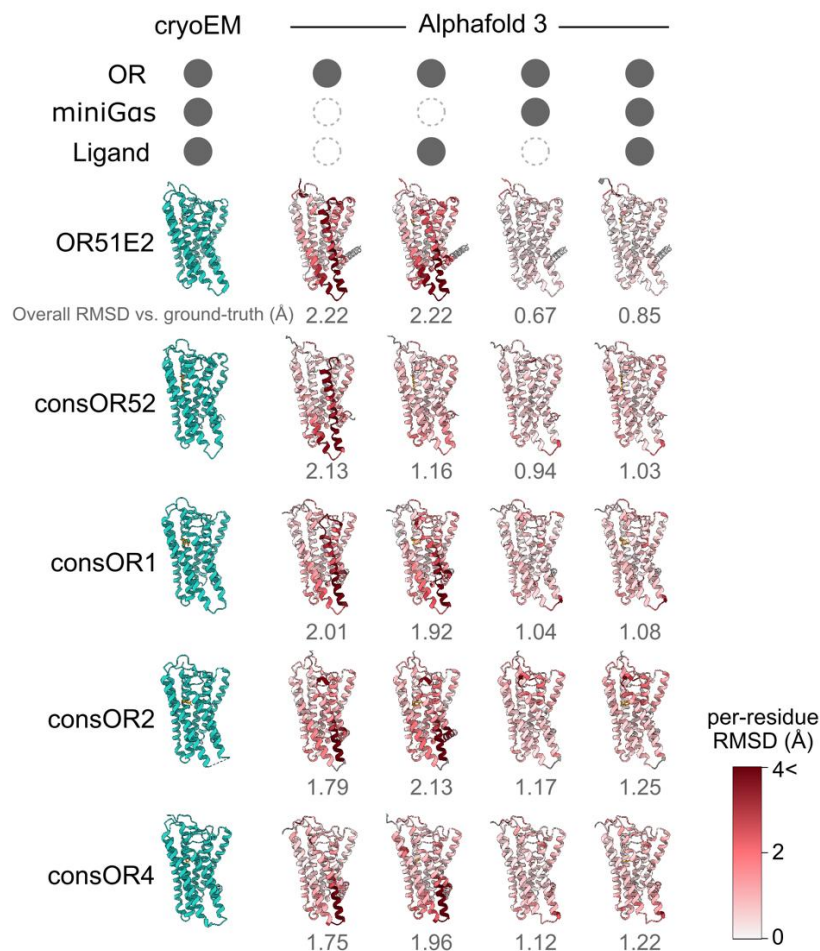


Figure S3. Structural agreement between cryo-EM olfactory receptor structures and AlphaFold3 predictions.

Comparison of experimentally determined cryo-EM structures with AlphaFold3-
 5 predicted models across multiple modeling conditions, including receptor alone (OR), receptor
 with ligand (OR + ligand), receptor with mini G protein (OR + miniGas), and receptor with both
 ligand and miniGas protein (OR + ligand + miniGas). For each receptor, per-residue root-mean-

square deviation (RMSD) relative to the cryo-EM structure is shown to assess local structural agreement, alongside overall backbone RMSD values summarized below each model.

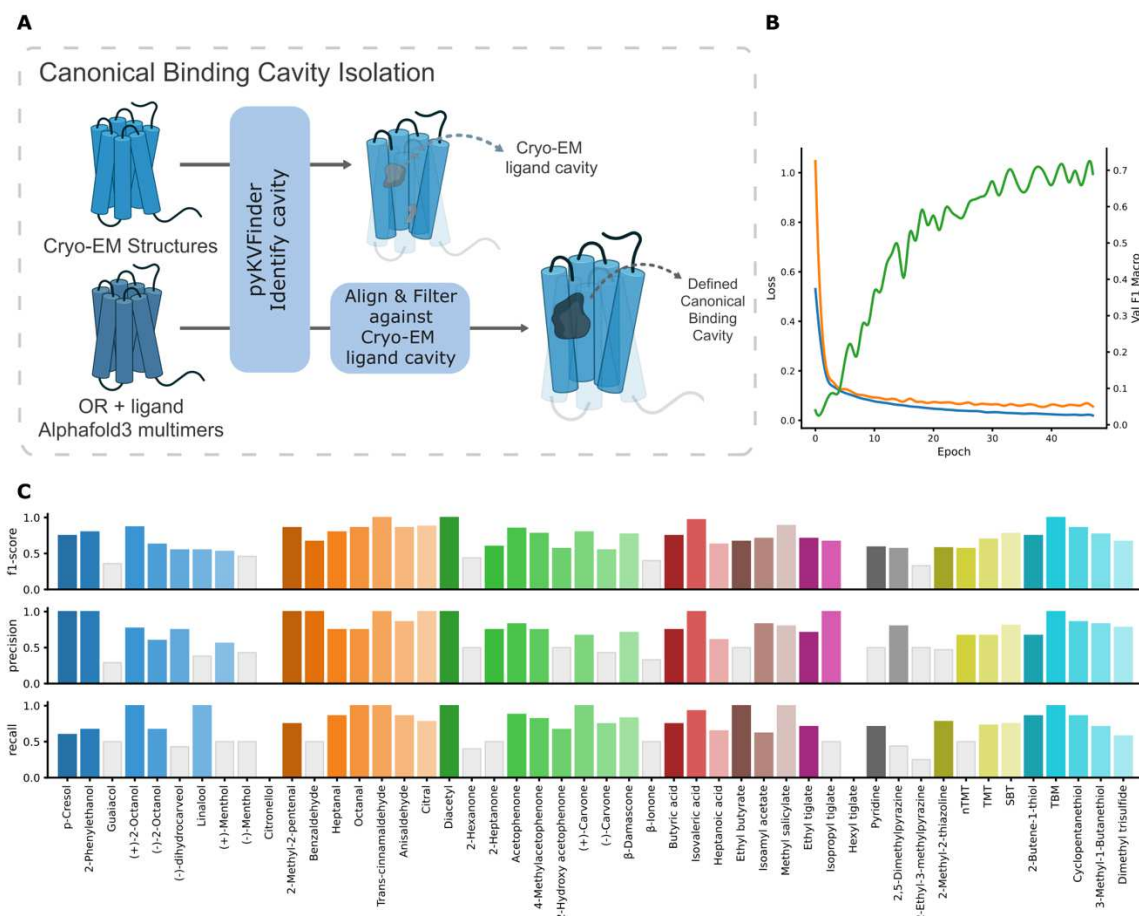
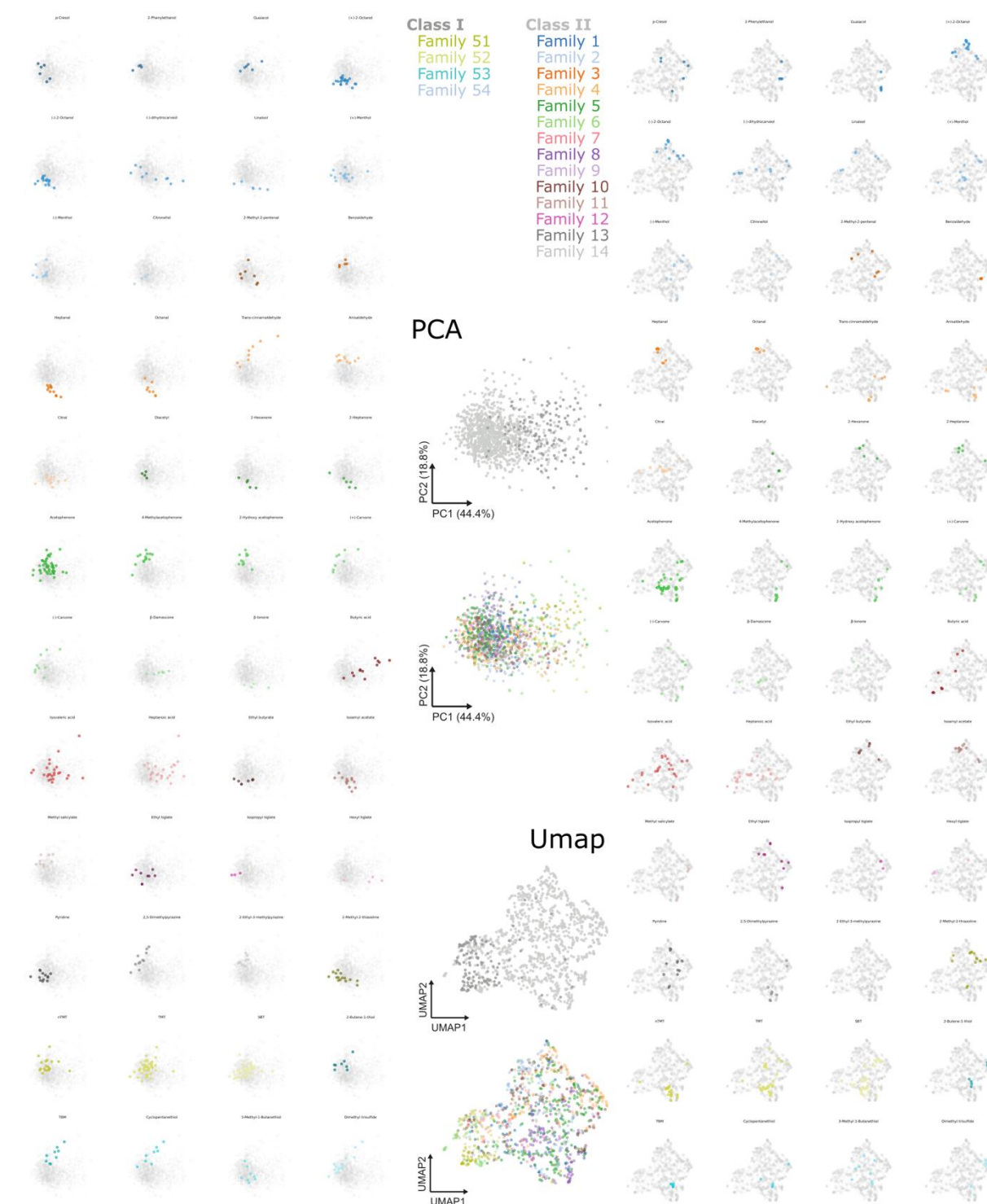


Figure S4. Binding cavity definition and model performance evaluation.

A. Schematic of the pipeline for canonical binding cavity identification. Cryo-EM structures and corresponding AlphaFold3-predicted models (OR + ligand + miniGas) were analyzed using pyKVFinder to detect putative cavities. The cryo-EM derived cavity was used as a reference to align and superimpose predicted cavities. The resulting consensus region was defined as the canonical binding cavity used for downstream filtering and voxelization (see Methods for details). **B.** Training dynamics of the final CNN model. Loss curves for training (blue) and validation (orange) are shown across epochs, alongside validation macro F1-score

(green), demonstrating stable optimization and concordant performance improvement. **C.**

Precision, recall, and F1-score for each of the 48 odorants. Odorants with performance values below 0.5 are highlighted in gray.

Figure

S5. Global organization of CNN-derived receptor embeddings across dimensionality reduction methods.

Comprehensive visualization of receptor embeddings using principal component analysis (PCA) and UMAP (smeLLMap). Left: PCA projections showing OR distributions for each of the 48 odorants, with receptors responding to each ligand highlighted. Right: Corresponding UMAP (smeLLMap) projections illustrating odor-specific clustering in the learned embedding space. Middle: Distribution of OR classes and families mapped onto both PCA and UMAP representations, demonstrating that higher-level class separation is preserved while finer family-level organization is less pronounced relative to functional clustering by odor response.

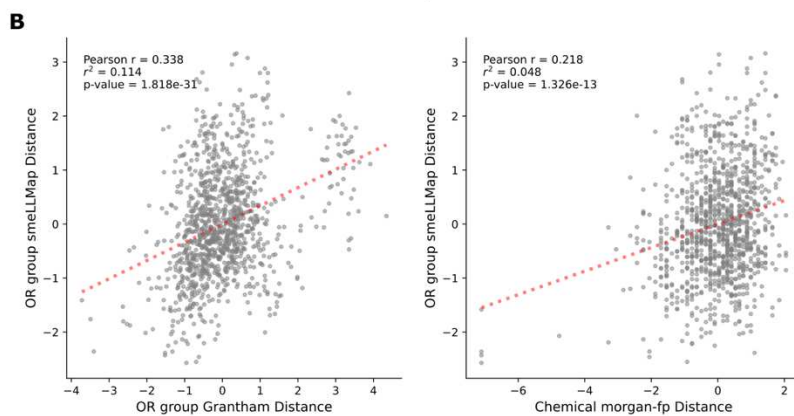
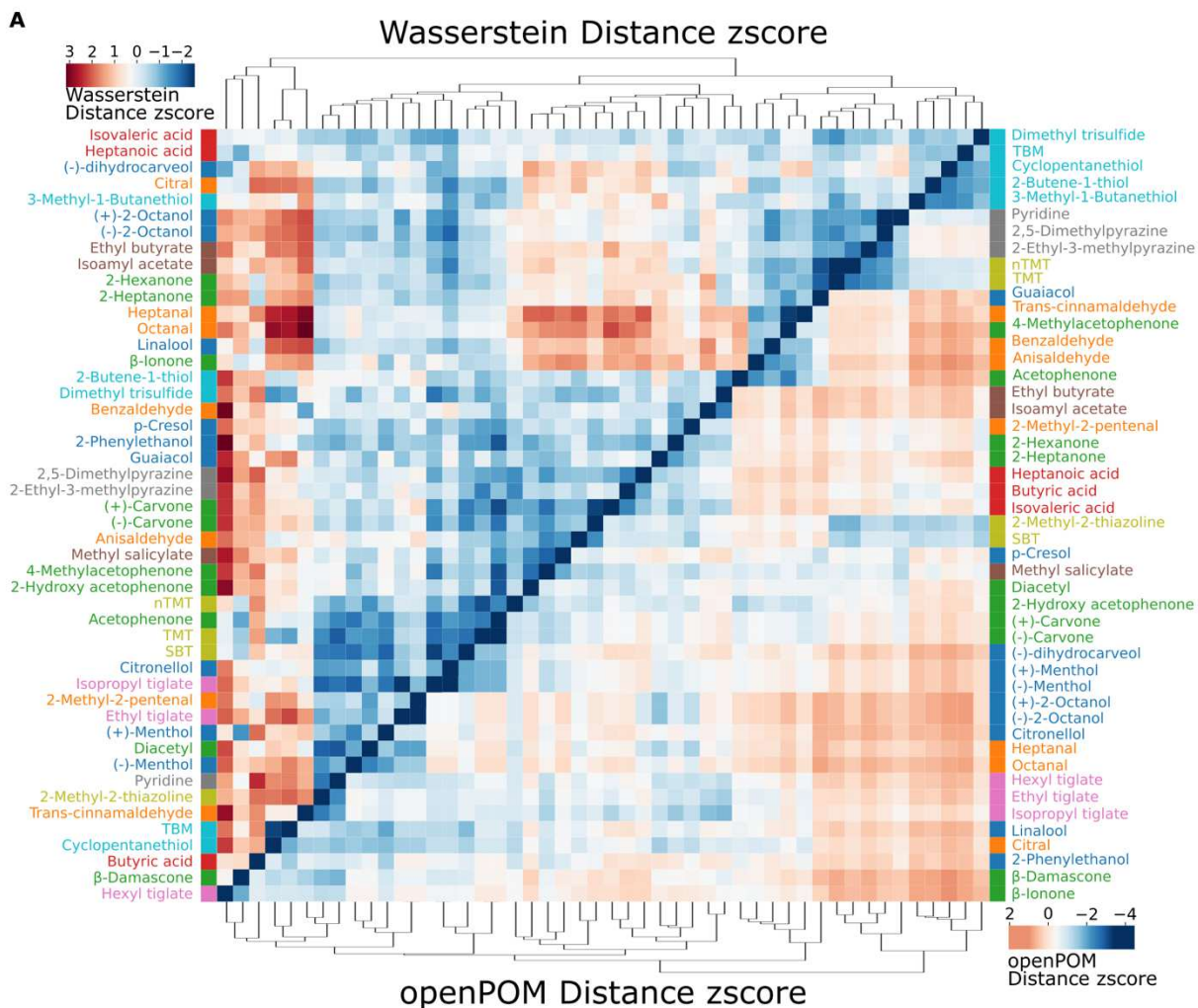


Figure S6. Comparison of embedding-derived receptor distances with sequence- and chemistry-based similarity metrics.

A. Dual-clustered heatmaps comparing normalized distance metrics across odorants. Top: z-scored Wasserstein distances between distributions of OR embeddings (smeLLMap) for each odorant pair. Bottom: corresponding z-scored Euclidean distances in OpenPOM space, representing perception-informed chemical similarity. **B.** Correlation analyses comparing receptor embedding distances with alternative similarity measures. Left: Pearson correlation between OR group distances in smeLLMap (Wasserstein) and sequence-based Grantham distances. Right: Pearson correlation between OR group distances in smeLLMap and chemical similarity derived from Morgan fingerprint representations

	aucroc	prob_corr	f1-score	precision	recall
p-Cresol	0.89	0.73255	0.75	1	0.6
2-Phenylethanol	0.9422442	0.80710	0.8	1	0.67
Guaiacol	0.9751243	0.49389	0.36	0.29	0.5
(+)-2-Octanol	1	0.91053	0.87	0.77	1
(-)-2-Octanol	0.9620181	0.67341	0.63	0.6	0.67
(-)-dihydrocarveol	0.9639249	0.69319	0.55	0.75	0.43
Linalool	1	0.72351	0.55	0.38	1
(+)-Menthol	0.8558974	0.62406	0.53	0.56	0.5
(-)-Menthol	0.9480737	0.51978	0.46	0.43	0.5
Citronellol	0.9639303	0.37660	0	0	0
2-Methyl-2-pentenal	0.9726368	0.84515	0.86	1	0.75
Benzaldehyde	0.9827586	0.68817	0.67	1	0.5
Heptanal	0.9891774	0.73052	0.8	0.75	0.86
Octanal	1	0.90210	0.86	0.75	1
Trans-cinnamaldehyde	1	0.98596	1	1	1
Anisaldehyde	0.9963925	0.86971	0.86	0.86	0.86
Citral	0.9948979	0.86625	0.88	1	0.78
Diacetyl	1	0.89681	1	1	1
2-Hexanone	0.962	0.52969	0.44	0.5	0.4
2-Heptanone	0.9949748	0.75804	0.6	0.75	0.5
Acetophenone	0.9501409	0.81457	0.85	0.83	0.88
4-Methylacetophenone	0.9582942	0.83288	0.78	0.75	0.82
2-Hydroxy	0.9690117	0.62587	0.57	0.5	0.67
(+)-Carvone	1	0.90482	0.8	0.67	1
(-)-Carvone	0.9912935	0.70908	0.55	0.43	0.75
CE<=Damascone	0.9120603	0.86859	0.77	0.71	0.83
CE<=Ionone	0.9458128	0.49842	0.4	0.33	0.5
Butyric acid	0.9987562	0.88191	0.75	0.75	0.75
Isovaleric acid	0.9971929	0.91817	0.97	1	0.93
Heptanoic acid	0.8989361	0.66110	0.63	0.61	0.65
Ethyl butyrate	0.9926108	0.57559	0.67	0.5	1
Isoamyl acetate	0.9574873	0.74228	0.71	0.83	0.62
Methyl salicylate	1	0.90854	0.89	0.8	1
Ethyl tiglate	0.9855699	0.71354	0.71	0.71	0.71
Isopropyl tiglate	1	0.79596	0.67	1	0.5
Hexyl tiglate	0.9141791	0.05535	0	0	0
Pyridine	0.8593073	0.67125	0.59	0.5	0.71
2,5-Dimethylpyrazine	0.9739229	0.67666	0.57	0.8	0.44
2-Ethyl-3-	0.9315920	0.38868	0.33	0.5	0.25
2-Methyl-2-thiazoline	0.9387755	0.63984	0.58	0.47	0.78
nTMT	0.7512437	0.64863	0.57	0.67	0.5

TMT	0.9696969	0.76199	0.7	0.67	0.73
SBT	0.9461259	0.75225	0.78	0.81	0.75
2-Butene-1-thiol	0.8607503	0.78151	0.75	0.67	0.86
TBM	1	0.99962	1	1	1
Cyclopentanethiol	0.998557	0.92029	0.86	0.86	0.86
3-Methyl-1-Butanethiol	0.9812409	0.80131	0.77	0.83	0.71
Dimethyl trisulfide	0.9827288	0.68576	0.67	0.78	0.58

Table S1. Trained model's validation odor prediction metric

Summary of validation performance for each odorant, including area under the receiver operating characteristic curve (AUROC), probability–response correlation (prob_corr), F1-score, precision, and recall.

	odor	cid	n_O R	obs_gran tham_m ean	obs_gran tham_m edian	null_gra ntham_ mean	null_gra ntham_ median	p_granth am_per m_mean	p_granth am_per m_medi an	obs_eucl id_mean	obs_eucl id_medi an	null_euc lid_mea n	null_euc lid_medi an	p_euclid _perm_ mean	p_euclid _perm_ median
0	Dimethyl trisulfide	19310	16	0.014	0.251	0.005	0.037	0.502	0.826	-0.443	-0.915	0.012	-0.163	0.056	0
1	3-Methyl-1- Butanethiol	10925	8	-0.600	-0.642	-0.005	0.022	0.036	0.032	-0.581	-0.556	-0.014	-0.151	0.06	0.132
2	Cyclopentanet hiol	15510	9	-1.254	-0.716	0.007	0.032	0	0.008	-0.283	-0.444	0.040	-0.108	0.198	0.178
3	TBM	6387	10	-1.371	-0.836	-0.012	0.027	0	0.002	-0.703	-0.789	0.020	-0.141	0.018	0.014
4	2-Butene-1- thiol	643345 1	9	-0.128	-0.056	-0.022	0.014	0.374	0.424	-0.764	-0.769	0.024	-0.126	0.016	0.028
5	SBT	162148	49	-0.331	-0.297	-0.004	0.022	0	0	-0.752	-0.866	0.007	-0.158	0	0
6	TMT	263626	45	-0.358	-0.284	-0.003	0.013	0.002	0.014	-0.661	-0.813	0.014	-0.158	0	0
7	nTMT	61653	19	-0.610	-0.602	-0.006	-0.002	0	0	-0.573	-0.627	0.009	-0.156	0.014	0.024
8	2-Methyl-2- thiazoline	16867	15	0.131	0.200	0.006	0.027	0.702	0.778	-0.725	-0.907	-0.014	-0.184	0.004	0
9	2-Ethyl-3- methylpyrazin e	27457	9	-0.191	0.155	-0.003	0.001	0.252	0.684	-0.953	-0.898	0.009	-0.123	0.006	0.016
10	2,5- Dimethylpyra zine	31252	11	-1.021	-0.758	0.005	0.043	0	0.004	-0.655	-0.714	0.000	-0.187	0.016	0.022
11	Pyridine	1049	13	-0.169	-0.076	0.003	0.020	0.228	0.322	-0.944	-1.028	0.019	-0.138	0	0
12	Hexyl tiglate	637523	4	1.175	1.071	0.003	0.045	0.992	0.974	0.064	-0.064	0.003	-0.099	0.584	0.522
13	Isopropyl tiglate	536774 5	3	-0.800	-0.571	0.019	0.070	0.114	0.196	-1.421	-1.464			0.012	0.01
14	Ethyl tiglate	528116 3	10	-0.234	0.106	-0.007	0.006	0.248	0.602	-0.572	-0.549	-0.033	-0.165	0.08	0.152
15	Methyl salicylate	4133	9	-0.503	-0.477	-0.013	0.000	0.036	0.042	-1.018	-1.061	0.007	-0.146	0.002	0.002
16	Isomyl acetate	31276	15	-0.612	-0.717	-0.003	0.029	0	0	-0.930	-0.982	0.011	-0.135	0	0
17	Ethyl butyrate	7762	9	-1.008	-0.419	0.005	0.006	0	0.086	-0.926	-0.575	-0.003	-0.148	0.004	0.09
18	Hepanoic acid	8094	29	-0.059	-0.013	0.001	0.021	0.35	0.41	-0.170	-0.324	0.015	-0.158	0.172	0.186
19	Isovaleric acid	10430	25	0.069	0.072	-0.004	0.009	0.666	0.638	-0.142	-0.440	-0.001	-0.164	0.26	0.078
20	Butyric acid	264	13	-0.843	-0.516	0.000	0.039	0	0.014	0.032	-0.185	0.010	-0.157	0.52	0.468
21	(E)-Ionone	638014	4	-0.278	-0.356	0.047	0.110	0.264	0.226	-0.435	-0.693	0.002	-0.091	0.256	0.122
22	(E)- Damascone	32052	6	-0.271	-0.116	-0.017	0.043	0.262	0.366	-0.863	-0.865	-0.010	-0.149	0.032	0.036
23	(-)-Carvone	439570	9	-0.910	-0.622	0.013	0.044	0.002	0.022	-0.259	-0.320	0.011	-0.137	0.256	0.312
24	(+)-Carvone	16724	8	-1.013	-0.816	-0.017	0.008	0.004	0.016	-0.397	-0.601	-0.021	-0.171	0.19	0.12
25	2-Hydroxy acetophenone	68490	11	-1.247	-0.796	-0.004	0.040	0	0.006	-1.050	-1.164	-0.016	-0.159	0	0
26	4- Methylacetop henone	8500	14	-1.515	-1.152	-0.004	0.031	0	0	-0.967	-1.086	0.033	-0.124	0	0
27	Acetophenone	7410	50	-0.409	-0.358	-0.006	0.019	0	0	-0.662	-0.706	-0.011	-0.174	0	0
28	2-Heptanone	8051	6	-1.072	-1.058	-0.006	0.015	0.006	0.006	-0.735	-1.090	0.010	-0.168	0.048	0.004
29	2-Hexanone	11583	5	-1.506	-1.657	0.013	0.061	0	0	-0.971	-0.958	0.060	-0.077	0.012	0.02
30	Diacetyl	650	4	-0.280	-0.176	-0.013	0.025	0.334	0.382	-1.494	-1.543	-0.002	-0.071	0	0
31	Citral	638011	12	-0.847	-0.837	-0.007	0.021	0	0	-0.914	-1.021	-0.010	-0.172	0	0
32	Anisaldehyde	31244	11	-0.731	-0.282	-0.017	0.007	0.01	0.144	-1.019	-1.069	0.035	-0.117	0.002	0
33	Trans- cinnamaldehy de	637511	8	0.138	0.211	0.012	0.012	0.67	0.696	0.126	-0.104	0.001	-0.127	0.642	0.528
34	Octanal	454	8	-1.601	-1.352	0.000	0.024	0	0	-1.175	-1.100	-0.016	-0.165	0	0.002
35	Heptanal	8130	11	-0.968	-0.877	0.009	0.024	0	0.002	-1.289	-1.330	0.001	-0.167	0	0
36	Benzaldehyde	240	5	0.230	-0.043	0.008	0.023	0.698	0.46	-1.492	-1.387	0.022	-0.123	0	0
37	2-Methyl-2- pentenal	531975 4	7	-0.433	-0.089	-0.001	-0.001	0.1	0.39	-0.462	-0.436	0.006	-0.133	0.148	0.25
38	Citronellol	8842	5	-0.415	0.131	-0.009	0.012	0.172	0.6	-1.160	-1.309	-0.001	-0.135	0.006	0
39	(-)-Menthol	16666	9	-0.581	-0.668	-0.021	0.006	0.04	0.024	-0.924	-0.945	-0.023	-0.171	0.006	0.012
40	(+)-Menthol	165675	15	-0.935	-0.844	-0.005	0.017	0	0	-0.659	-0.730	0.008	-0.175	0.002	0.006
41	Linalool	6549	6	-0.929	-0.825	-0.006	0.014	0.012	0.038	-0.453	-0.353	0.026	-0.099	0.146	0.28
42	(-)- dihydrocarveol	443163	10	-0.201	-0.126	0.016	0.048	0.232	0.316	0.527	0.217	-0.002	-0.170	0.92	0.834
43	(-)-2-Octanol	80080	17	-0.822	-0.855	-0.004	0.028	0	0	-1.124	-1.114	0.006	-0.165	0	0
44	(+)-2-Octanol	272388 8	22	-0.280	-0.305	0.000	0.032	0.068	0.064	-0.974	-1.079	0.013	-0.157	0	0
45	Guaiacol	460	6	-0.864	-0.334	-0.014	0.021	0.014	0.186	-1.024	-1.223	-0.011	-0.142	0.004	0.002
46	2- Phenylethanol	6054	4	-1.095	-0.968	-0.057	0.007	0.02	0.042	-1.258	-1.341	0.020	-0.026	0.002	0
47	p-Cresol	2879	6	-0.806	-0.787	-0.024	0.024	0.026	0.048	-0.414	-0.385	-0.001	-0.111	0.206	0.276

Table S2. Pairwise receptor similarity metrics for odor-specific responding ORs.

Summary of observed and permuted pairwise distances among ORs responding to each odorant, computed using Grantham sequence distances and smeLLMap embedding Euclidean distances. For each odorant, mean and median distances are reported alongside corresponding null distributions generated by random permutation, with associated permutation-based *p-values* indicating statistical significance.

5

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Naturesupplementarymaterials.docx](#)