7 January 2026

ConfDENSE: A conformer aware electron density based machine learning paradigm for navigating the odorant landscape

Pinaki Saha[1], Sarabeshwar Balaji[2], Mrityunjay Sharma[3], Aryan Amit Barsainyan, Ritesh Kumar[3], Volker Steuber[1], Michael Schmuker[1]

1. University of Hertfordshire
2. IISER Bhopal
3. CSIR CSIO

Abstract

Olfaction arises from the interaction of odorants with olfactory receptors, a process shaped by molecular geometry, electron distribution, and conformational preference. We present ConfDENSE, a Set2Set enhanced PointNet model that learns directly from Hirshfeld promolecule electron-density point clouds, preserving full 3D electronic in- formation without downsampling.Despite using no receptor structural data, ConfDENSE accurately identifies bioac- tive conformers from ensemble inputs. For the only available human odorant receptor structures, the model's selected conformers achieve sub-angstrom RMSDs to crystallo- graphic ligand poses and frequently outperform conventional docking. Combining ConfDENSE with explainability analysis further reveals the substruc- tures most responsible for receptor engagement, aligning with experimental interaction patterns. This ligand-centric and interpretable framework naturally supports phar- macophore extraction and scaffold-based design, enabling identification of conserved binding motifs even when receptor structures are missing. ConfDENSE thus provides a compact, physics-aware approach to computational olfaction, linking electron density, conformational preference, and odorant recognition in a structurally agnostic manner.

Keywords

Olfaction, structure free learning, bioactive conformation prediction, QM aided AI, Electron density

ConfDENSE: A conformer aware electron density based machine learning paradigm for navigating the odorant landscape

Pinaki Saha,[*,†] Sarabeshwar Balaji,[‡] Mrityunjay Sharma,[¶,§,‖] Aryan Amit Barsainyan,[⊥] Ritesh Kumar,[*,¶,§] Volker Steuber,[†] and Michael Schmuker[#]

†*UH Biocomputation Group, University of Hertfordshire, United Kingdom*

‡*Indian Institute of Science Education and Research Bhopal (IISERB), Madhya Pradesh-462066,India*

¶*CSIR - Central Scientific Instruments Organisation, Sector 30-C, Chandigarh-160030, India*

§*Academy of Scientific and Innovative Research (AcSIR), Ghaziabad-201002, India*

‖*Department of Higher Education, Himachal Pradesh, Shimla-171001,India*

⊥*National Institute of Technology Karnataka Surathkal, Karnataka-575025*

#*Helmholtz-Gemeinschaft, Berlin, Germany*

E-mail: p.saha3@herts.ac.uk; riteshkr@csio.res.in

Abstract

Olfaction arises from the interaction of odorants with olfactory receptors, a process shaped by molecular geometry, electron distribution, and conformational preference. We present ConfDENSE, a Set2Set enhanced PointNet model that learns directly from Hirshfeld promolecule electron-density point clouds, preserving full 3D electronic information without downsampling.

Despite using no receptor structural data, ConfDENSE accurately identifies bioactive conformers from ensemble inputs. For the only available human odorant receptor structures, the model's selected conformers achieve sub-angstrom RMSDs to crystallographic ligand poses and frequently outperform conventional docking.

Combining ConfDENSE with explainability analysis further reveals the substructures most responsible for receptor engagement, aligning with experimental interaction patterns. This ligand-centric and interpretable framework naturally supports pharmacophore extraction and scaffold-based design, enabling identification of conserved binding motifs even when receptor structures are missing.

ConfDENSE thus provides a compact, physics-aware approach to computational olfaction, linking electron density, conformational preference, and odorant recognition in a structurally agnostic manner.

Introduction

In computational chemistry, electron density plays a crucial role. Density functional theory (DFT) allows researchers to derive essential properties of chemical systems from their electron density distribution,[1–3] but it can be computationally expensive.

An alternative solution for speeding up calculations is the use of molecular mechanics. Molecular mechanics (MM) treats atoms and bonds as classical Newtonian objects, and parameterized force fields are utilized to calculate the energy of molecular systems. While molecular mechanics based force field calculations are much faster than quantum chemical ones – by a large order of magnitude – they are often found to be less accurate. This is due to the simplistic Newtonian approximations which make MM force fields unable to capture quantum effects occurring at the molecular level.[4]

In recent years, machine learning has been widely used for studying chemical systems as this can result in rapid inference while maintaining accuracy. Machine learning (ML) is a powerful emerging technique for the construction of molecular transferrable and non-

transferrable potentials. In the field of computational chemistry, ML has currently and extensively been implemented to solve problems in a variety of chemical subjects such as, among others, the prediction of structures, reaction pathways and various physico-chemical properties.[5–8]

At present, research is ongoing to augment computational chemistry with machine learning for creating machine learning based force fields/potentials which are trained on computational chemistry data, an notable example being the Behler–Parrinello neural network (BPNN)[?]

Electron density is a spatial distribution that describes the probability of finding electrons at each point around a molecule or atom. A molecule's electron density determines its shape, reactivity, and interactions, and serves as the fundamental quantity from which most chemical properties can be derived. Presently, electron density has been also utilized in machine learning research: AIMLDM (Atoms in molecule localized and delocalized matrices) augmented GNNs models have been applied to predict olfaction labels of molecules. AIMLDM is based on Bader's partitioning scheme of electron density.[9] Another prominent electron density partitioning scheme is Hirshfeld's partitioning scheme, which has also been extensively used in machine learning based research.[10] Logan *et al.* have applied deep learning to predict formation energy and lattice structure parameters using Hirshfeld fingerprints as inputs for the CNNs.[11] Saha *et al* have utilized Hirshfeld promolecule electron density via a convolutional neural network to predict stable isomers of Boron nanoclusters.[12]

In this paper we have used promolecule densities to understand the realm of olfaction. The Hirshfeld partitioning scheme is a partitioning method where the molecular electron density is divided into its constituent atomic densities. This scheme assigns a promolecule density[10] to a molecule by superimposing the electron densities of its atomic constituents (precalculated at HF/3-211G level) and proportionally partitioning the electron density based on their contributions. This electron density is then partitioned by giving weights to the atomic constituents proportional to their contribution to the electron density. The

promolecule density ($\rho^o(\mathbf{r})$) can be written as a weighted summation of atomic densities $\rho_A^o(\mathbf{r})$:

$$\rho^o(\mathbf{r}) = \sum_A w_A \rho_A^o(\mathbf{r}) \tag{1}$$

where the weight is given by the following expression:

$$w_A = \rho_A^o(\mathbf{r}) / \rho^o(\mathbf{r}) \tag{2}$$

This allows us to create promolecule densities for molecules on the fly (**Figure** 1). Promolecule density is analogous to the trial density used in density functional theory (DFT) calculations. In DFT calculations, a trial electron density is iteratively input into the self-consistent field (SCF) equations until convergence is achieved with the ground-state electron density. All required molecular properties can, in principle, be derived from the ground-state electron density. This mapping from trial electron density to molecular properties can also be approximated using machine learning frameworks, particularly neural networks which are considered universal function approximators.

We have applied **PointNet** neural networks in our study.[13] PointNet is a neural network architecture designed specifically for processing point cloud data, which is a collection of points representing a 3D shape or object. PointNet is groundbreaking because it directly analyses raw point clouds without requiring their transformation into structured data formats like grids or meshes. PointNet effectively extracts meaningful features and patterns, making it valuable for tasks such as object recognition, classification, and segmentation. Its impact is particularly significant in computer vision and 3D scene understanding, where direct interaction with spatial data is essential. A notable strength of PointNet is its invariance to the number and ordering of input points. This is achieved through the use of a symmetric aggregation function that processes each point independently, enabling efficient and scalable handling of variable-sized point clouds. This network architecture also effectively captures

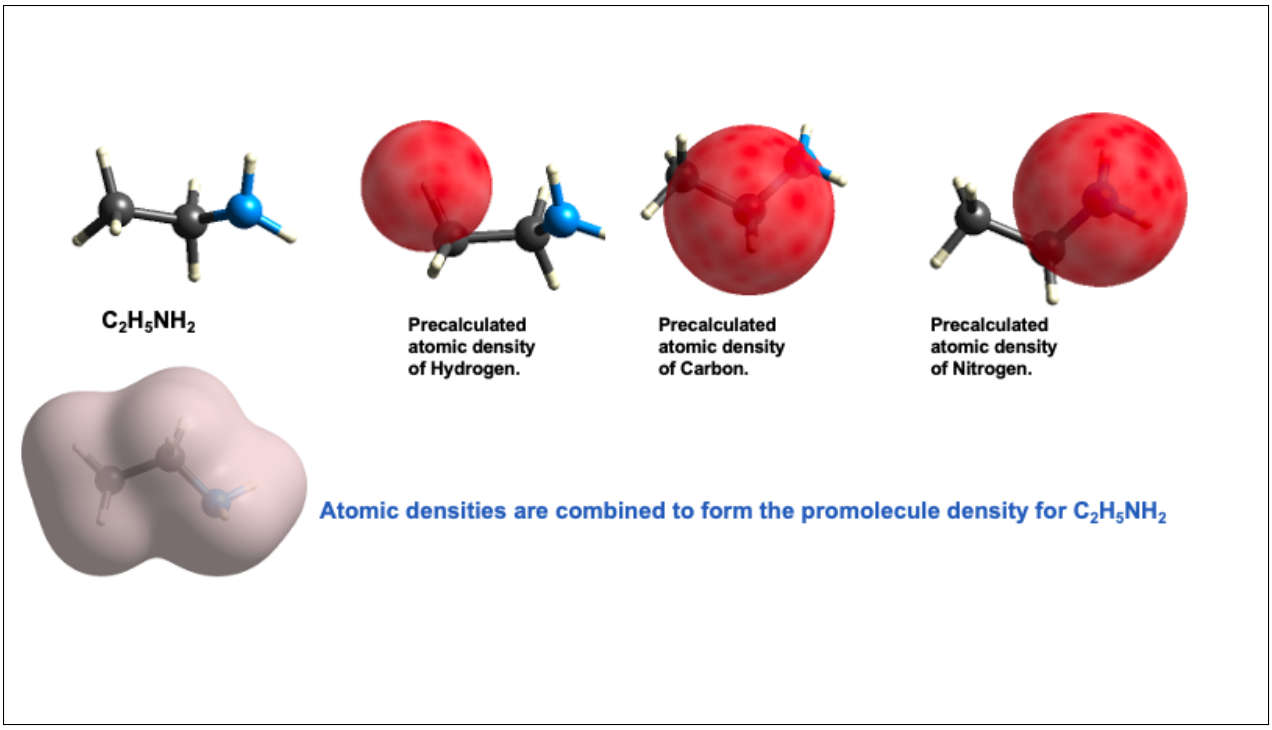


Figure 1: Creation of Hirshfeld based promolecule density of ethylamine from precalculated atomic densities at HF/3-211G level.

both global and local features of the input data, making PointNet architectures useful for the analysis electron densities, which are pointcloud vectors. PointNet architectures have recently been used for molecular systems, particularly in materials science research. Recently, Kim *et al.* have used PointNet to predict the expansion group/space group classification for a large dataset of X-ray crystal structures using their charge density.[14] Since PointNet was originally developed for dense LiDAR-derived 3D point clouds, the charge density data had to be downsampled by Kim *et al.* to ensure compatibility with the network's input requirements, reducing the number of points to result in a sparse representation of the charge density. Downsampling produces a sparser representation of the data, but inevitably comes at the cost of reduced information.

In this study, we employed a masking strategy to circumvent information loss typically associated with downsampling. This approach enables the full promolecule density of molecules regardless of their size to be used as input to the model.

We utilized this Pointnet model with masking to explore olfaction. Olfaction is a critical sense for many animals, essential for survival related behaviors such as foraging, mating, and detecting predators or prey. In humans, it plays a key role in hazard detection and hygiene. Despite its importance, the molecular mechanisms of olfaction remain poorly understood. Human olfactory perception involves 802 olfactory receptor (OR) genes, with 388 functional receptors and 414 pseudogenes.[15] Structural characterization of ORs has been challenging due to their genetic variability, low *in vitro* expression, and instability during isolation. Furthermore, odorant-OR binding is complex: individual ORs can detect multiple odorants, and a single odorant can activate several receptors. This many-to-many relationship complicates structure-based understanding of odor recognition.[16] Nonetheless, recent advances have emerged. Cryo-EM resolved the structure of human OR51E2, revealing size selectivity for short-chain carboxylic acids.[17] Alphafold and molecular dynamics simulations have also been employed to model various ORs in the human nasal epithelium.[18] Complementing structure-based approaches, ligand-based methods predict olfactory properties using molecular features alone.[19–21]

The complexity of olfaction also stems from the fact that, unlike conventional protein–ligand interactions occurring in solution, odorant binding to olfactory receptors takes place in the gas phase. As a result, odorant molecules exhibit significantly greater conformational freedom, making it challenging to accurately predict the binding pose of the active conformer computationally. Given that our training data incorporates a three dimensional property (electron density), we hypothesized that providing the correct conformer configuration will enhance the model's predictive accuracy. Since the active conformer of a molecule is not known *a priori*, a practical solution was to provide an ensemble of conformers for each molecule, allowing the PointNet model to process a more comprehensive representation. To effectively integrate information from multiple conformers of each molecule, we employed an order-invariant pooling mechanism called Set2Set,[22] designed to operate on unordered sets of representations. It leverages an LSTM-based controller that iteratively refines a con-

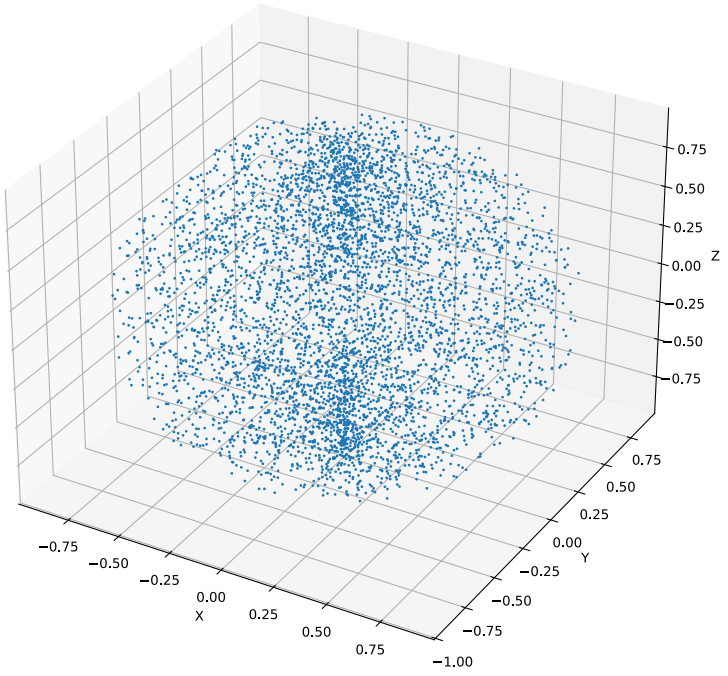


Figure 2: Illustration of a point cloud sampled from the electron density of a conformer of 'CCC(O)c1ccccc1'. The units of all axis are Å.

text vector, and enables the model to capture complex relationships. Unlike fixed pooling methods, Set2Set adaptively allows the model to focus on important conformers that exhibit key geometric or electronic features pertinent to the property being predicted. Herein, we present a Set2Set enhanced PointNet architecture designed to effectively process olfaction related data.

Methods

The dataset utilized here is identical to that used in our prior olfaction based ML study DenseSense.[21] The dataset is sourced from two websites which contains odorant and odor label data: Leffingwell and goodscents. In total we have 4983 molecules in our dataset. We utilized RDKIT to generate 100 conformers for each of the molecules. The conformers were then converted to cif format using the program VESTA.[23] This resulted in 498 K

($\sim 1/2million$) cif files**??**. These cif files were then used to generate promolecule density using the program Crystal explorer2.[24] This resulted in a dataset of approximately half a million promolecule electron densities. This dataset was then used by PointNet to predict olfaction labels.

PointNet and 3D CNNs[25]are two major architectures used for processing point cloud datasets. However, 3D CNNs require the voxelization of point cloud data, which can lead to information loss and significantly higher computational cost compared to the more efficient and direct approach used by PointNet.

The PointNet architecture for point clouds is analogous to the conventional Convolutional Neural Network (CNN) used for image data. PointNet employs a series of 1D convolutional (conv1D) operations to transform data from the cartesian space of the point cloud into a learned feature space. The resulting feature vectors are then aggregated using global max pooling to produce a fixed-size embedding, which is passed through a feedforward neural network for classification or regression tasks.

In the original PointNet implementation, all point clouds were uniformly downsampled to 4,096 points prior to being inputted into the network. While this ensured consistent input size, it also introduced potential information loss. To circumvent this issue of information loss, we employed a zero-padding strategy in place of downsampling. Specifically, we padded each molecular point cloud with dummy points whose Cartesian coordinates were set to zero, such that all point clouds in the dataset achieved a uniform size. These zero-padded point clouds were then passed to the PointNet architecture. To ensure that the padded points did not influence the learned representation, we applied a masking mechanism prior to the global max pooling operation, explicitly excluding feature vectors corresponding to the zero-coordinate (padded) points.

This masking strategy enabled us to preserve the full informational content of the molecular densities while accommodating inputs of varying sizes.

We utilized the original PointNet architecture[13] without the feature transform for clas-

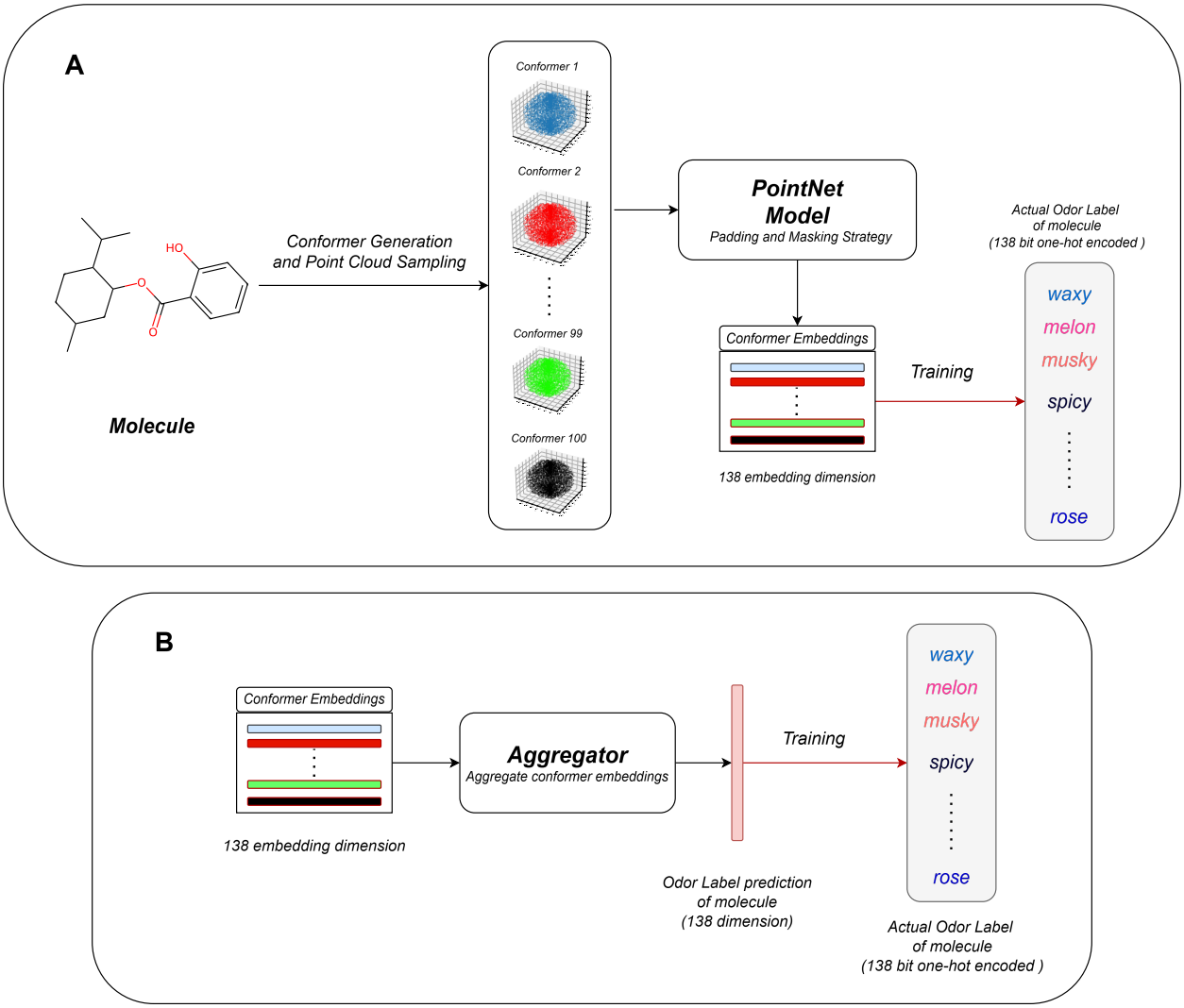


Figure 3: **A.** 100 conformers of a given molecule are generated and point clouds are sampled from it. A *PointNet* model[13] is trained to predict the odor labels of the molecule using the conformer point clouds, effectively learning a mapping from all conformer point clouds to multi-labeled target label for the corresponding molecule. **B.** An *Aggregator* is trained independently using the conformer embeddings to predict the ground truth odor labels of the molecule. To validate our model, we perform molecular analysis on the conformer whose embedding best represents the molecule's true odor label. Such embeddings are identified by: a) Selecting the conformer that has highest similarity to the prediction of the *Aggregator*. b) Selecting the conformer that has highest similarity to the true label embedding of the molecule.

sification. The PointNet model was trained for 10 epochs. The embeddings of conformers obtained from PointNet is then fed to Set2Set aggregation mechanism to predict the odor label of the molecule. The PointNet model was trained for 10 epochs.

The Set2Set mechanism was trained for 300 epochs. The task of odor prediction is treated as a multi-label classification problem, as individual odorant molecules can be associated with multiple odor descriptors (labels) simultaneously. A significant challenge in this task is the inherent imbalance in the odor labels within the dataset. To effectively handle this issue, the Imbalance ratio per label (IRLbl)[22] loss was used to train both the PointNet architecture and the Set2Set aggregation mechanism3.

Kabsch alignment[26] was performed using inhouse python script. The molecular docking was performed using webina web server.[27] Webina runs the Autodock Vina algorithm on web server.[28] The default parameters were used to perform docking.

Results and Discussions

Our main objective was to benchmark the performance of our ConfDENSE model against the current state-of-the-art ligand-based machine learning (ML) approach in olfaction. Lee /textitet al recently demonstrated that an ensemble of 50 message passing GNNs achieved state-of-the-art performance in odor prediction, with an AUROC of 0.894. In contrast, their baseline model—random forest using Morgan fingerprints—reached an AUROC of 0.85. Since their GNN models were not publicly available, we had replicated them in-house. We recreated this framework using the same 138 odor labels from Lee *et al.*[19] The odor labels used are given below:

alcoholic, aldehydic, alliaceous, almond, amber, animal, anisic, apple, apricot, aromatic, balsamic, banana, beefy, bergamot, berry, bitter, black currant, brandy, burnt, buttery, cabbage, camphoreous, caramellic, cedar, celery, chamomile, cheesy, cherry, chocolate, cinnamon, citrus, clean, clove, cocoa, coconut, coffee, cognac, cooked, cool-

ing, cortex, coumarinic, creamy, cucumber, dairy, dry, earthy, ethereal, fatty, fermented, fishy, floral, fresh, fruit skin, fruity, garlic, gassy, geranium, grape, grapefruit, grassy, green, hawthorn, hay, hazelnut, herbal, honey, hyacinth, jasmine, juicy, ketonic, lactonic, lavender, leafy, leathery, lemon, lily, malty, meaty, medicinal, melon, metallic, milky, mint, muguet, mushroom, musk, musty, natural, nutty, odorless, oily, onion, orange, orangeflower, orris, ozone, peach, pear, phenolic, pine, pineapple, plum, popcorn, potato, powdery, pungent, radish, raspberry, ripe, roasted, rose, rummy, sandalwood, savory, sharp, smoky, soapy, solvent, sour, spicy, strawberry, sulfurous, sweaty, sweet, tea, terpenic, tobacco, tomato, tropical, vanilla, vegetable, vetiver, violet, warm, waxy, weedy, winey, woody.

Lee et al did not report the performance of individual MPGNN models. Through our replication, we determined that a single GNN model achieved a validation AUROC of 0.877. Burn et al have also reported validation AUROC of 0.853 but on a relatively smaller dataset ($\sim 3.5K$ molecules).[20] We aimed to surpass the validation AUROC score of 0.877 using our PointNet-based ConfDENSE model. The PointNet model is trained to predict the labels for different conformers of a given molecule.

We utilized Optuna,[29] a Bayesian Optimization framework to tune the hyperparameters of the Set2Set based mechanism. Table 1 outlines the search space ranges and the corresponding optimal values selected. A total of 50 independent trials, each involving 100 training epochs were conducted to thoroughly explore the hyperparameter space. Early stopping was employed to avoid overfitting and to stop training once the validation loss failed to improve over a predefined number of consecutive epochs. This was done by using a patience score which maintained the epochs after which the early stopping was triggered. This strategy not only reduces computational overhead by avoiding unnecessary training but also enhances model robustness. The dropout rate helps prevent overfitting by randomly deactivating a portion of neurons during training whereas the hidden dimension size determine the capacity of the model's internal representations. The learning rate controls the step size during

gradient-based optimization and the batch size, which affects the stability and speed of the learning process.

Table 1: Best Hyperparameter Values Chosen After Bayesian Optimization with Optuna

hyperparameter	**search space**	**chosen value**
Batch size	{16, 32, 64, 128}	16
Hidden dimension	{32, 64}	32
Set2Set_steps	{ 2, 3 }	3
Feedforward dimension	{16, 32, 64}	64
Dropout	[0.1, 0.5]	0.14
Learning rate	[1e-5, 1e-3] (log scale)	7.69e-04
StepLR step_size	{10, 20, 30, 50}	50
StepLR gamma	[0.3, 0.6]	0.547

Table 2: Comparison of ROC-AUC Scores between various model architectures

Model comparison	**Valid ROC-AUC**	**Test ROC-AUC**
ConfDENSE	0.7607	0.7402
DenseSense (individual DMPNN+LDM)	0.87	0.86
openPOM(individual MPGNN)	0.86	0.87

Burns *et al.* created QuantumScents ML model for olfaction.[20] Quantum scents utilized directed message passing neural networks (DMPNN) with features like atomic numbers, bond types, and Hirshfeld charges. QuantumScents reported AUROC of 0.874 and 0.865 (validation and test set respectively) but at 400 epochs. At 40 epochs the AUROC for validation is shown to be approximately around 0.65 which is less than the results reported by us or Lee *et al.*[19] at 40 epochs. Also a fair comparison cannot be warranted between their and our models as their dataset is smaller ($\sim$ 3500 molecules) compared to our dataset of $\sim$ 5000 molecules.

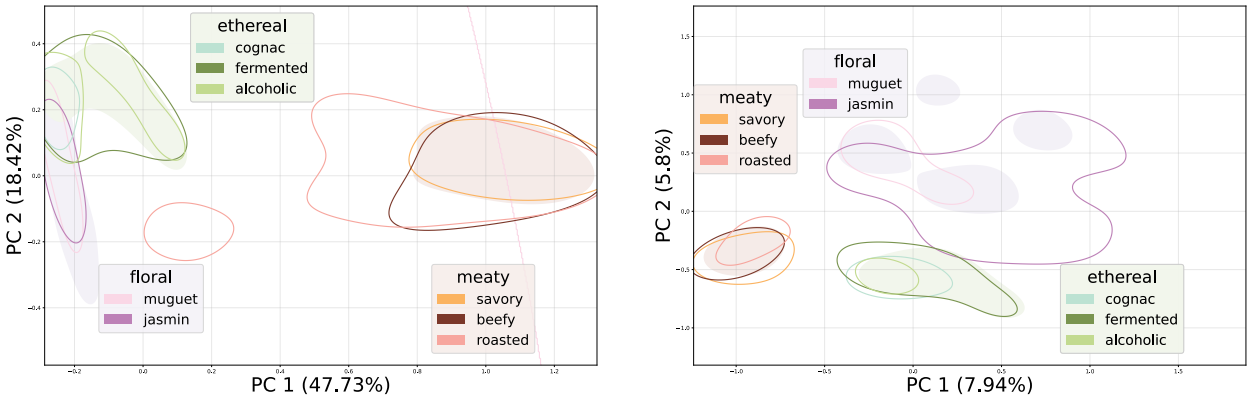
ConfDENSE has a respectable AUC of 0.76 against this challenging olfaction dataset. This is also remarkable due to the fact that no atom or bond information was provided for training the model, only pointcloud representation of electron density.

Principal odour map (POM) is the state of the art model for olfaction based ML. We compare our model with openPOM[21] (open version of POM), openPOM indeed has higher AUC than the ConfDENSE, we wanted to know for which labels openPOM was better than ConfDENSE. We analyzed AUC per label for both ConfDENSE and openPOM and found openPOM to be better than ConfDENSE for most of the cases but for few labels ConfDENSE was better than openPOM: **black currant, malty, cooked, weedy, cinnamon, ripe, grapefruit, metallic, solvent, burnt, tomato, cabbage, medicinal, smoky, beefy, onion, garlic, musk**.

Molecules in our dataset which are associated with these labels might exhibit subtle conformational preferences which GNNS cannot capture. Traditional GNNs, which operate primarily on 2D molecular topology and coarse structural descriptors, inherently struggle to capture these nuanced features because graph representations cannot fully encode the spatial distribution of electrons or the shape of the molecular electrostatic envelope.

The performance on these chemically demanding odour classes therefore highlights ConfDENSE's ability to extract structural cues that are invisible to topology based models (GNNs like openPOM) and to recognize odor categories where true molecular perception is governed by electron distribution rather than by connectivity alone. This establishes ConfDENSE as a complementary and highly valuable framework.

We also looked at the Kernel density estimation (KDE) distribution for certain odour labels for ConfDENSE model. Fig. 2a shows the embedding KDE derived from promolecule electron-density representations, where odor categories form compact and well-defined clusters. Floral classes such as floral, muguet, and jasmin also remain relatively localized, indicating that the underlying density based embeddings capture meaningful physicochemical structure. In contrast, the label KDE (Fig 2b) is markedly more diffuse, especially for the floral descriptors, which spreads widely across the plot. This dispersion reflects the broad structural diversity and perceptual ambiguity of floral odors, causing the final label space to be less tightly organized than the embedding itself.



(a) (b)

Figure 4: Comparison of KDE distributions: (**a**) label-space density and (**b**) embedding-space density.

Conformer analysis

As previously noted in the manuscript, the ConfDENSE model can infer which conformer contributes most significantly to the final prediction. Its permutation-invariant Set2Set layer aggregates information across all conformers, allowing the model to focus on the most relevant structural and electronic features. We hypothesized that this methodology could further aid in identifying the specific conformation of an odorant molecule that binds to olfactory receptors, since the perception of smell and the associated odor label emerges from the interaction between the odorant and one or more receptor proteins.

The conformer thus selected by our methodology would be thus a conformer that is representative of the actual odor profile of a given molecule (Bioactive confirmation of the molecule). We have identified the bioactive conformer using two strategies. In the first case, we selected the conformer whose predicted odor label has the highest similarity (cosine similarity) with respect to the actual odor labels of the molecule. In the second case, we selected the conformer whose predicted odor label has the highest similarity (cosine similarity) with respect to the prediction of the Set2Set aggregation of the conformer embeddings of the molecule.

We started the conformer analysis with the only available crystal structure of human

olfactory receptor available in the PDB: Human olfactory receptor OR51E2 bound to propionate (PDB code: **8F76**).[17] Additionally, we have incorporated the crystal structure of human olfactory receptors derived through consensus modeling[30].[31] A consensus structure is generated by integrating conserved features from multiple homologous structures to produce a representative, robust structure. Crystallization of human olfactory receptors remains notoriously challenging. Consensus-based modeling provides receptor structures that are more amenable to isolation, while retaining the canonical architecture of the ligand-binding pocket: An evolutionarily conserved feature essential for odourant recognition. The consensus human ORs crystal structures used are: **8HTI** (Human consensus olfactory receptor OR52c in complex with octanoic acid)[30] and **8UYQ** (Human consensus olfactory receptor consOR4 bound to 2-methylthiazoline).[31] The three aforementioned molecules are present in the training dataset, where each conformer yields a predicted output label. For evaluation, we select the conformer that exhibits the minimum cosine distance between the predicted and true labels. In cases if a molecule appears in the test dataset without an associated true label, the predicted label generated by the Set2Set mechanism is used for similarity calculations. Both sets of scores are reported in this paper.

Table 3: Conformers of each of the ligands bound to human olfactory receptor present in the PDB database are shown in the below table. The conformer ID and their cosine similarity has been listed for each conformer.

Molecule	True Label Proximity	Predicted Label Proximity
Propionic acid	32 (0.3)	90 (0.8)
Octanoic acid	85 (0.3)	85 (0.7)
2-methylthiazoline	55 (0.4)	4 (0.8)

For each molecule thus, two conformers were identified that maximize the cosine similarity with respect to both the true label and the Set2Set-aggregated prediction (Table 3).

These conformers were subsequently aligned with the ligand present in the crystal structure using the Kabsch algorithm. The Kabsch method performs rigid-body alignment by applying optimal rotation without altering the internal conformation of the molecule, thereby ensuring that the alignment preserves its geometry. Kabsch algorithm does this rotation by

finding optimal rotation matrix using Singular value decomposition (SVD). The root-mean-square deviation (RMSD) was then calculated to assess whether the predicted conformer reproduces the conformation of the reference ligand. For propionate, we obtained excellent RMSD values of 0.07 Å(True label) & 0.08 Å(Predicted label) (Table 4). The conformers are approximately identical to the bioactive conformation of propionate in the crystal structure: **8F76**.

In the case of octanoic acid, both the true-label and Set2Set aggregation approach yielded identical conformers, and the resulting RMSD value of 1.35 Å is highly satisfactory for a long, flexible linear molecule. Similarly, methyl thiazoline showed excellent RMSD values under both evaluation methods: 0.21 Å (True label) and 0.45 Å (Predicted label) Although this compound contains a heterocyclic ring, its complete unsaturation confers a certain degree of conformational flexibility. In molecular docking RMSD less than 2 Å is generally considered indicative of a good pose, the RMSD values reported for the aforementioned cases fall well within this threshold.

Table 4: Kabsch alignment score (rmsd) for the conformers of propionate, octanoate and 2-methylthiazoline

Molecule	True Label Proximity	Predicted Label Proximity
Propionic acid	0.07 Å	0.08 Å
Octanoic acid	1.35 Å	1.35 Å
2-methylthiazoline	0.21 Å	0.45 Å

For comparison, results obtained using conventional molecular docking are also presented. The ligands of **8F76**, **8HTI** and **8UYQ**: propionate, octanoate and 2-methylthiazoline were redocked to their respective apo proteins. The redocking was performed by webina server, smiles of the aforementioned ligands were given to webina for conversion to respective pdbqt format required for the docking. The default parameters of the webina was used for the docking. We chose the lowest energy docked conformer for each of the molecule, following which we performed Kabsch alignment of the docked ligand with the actual crystallographic ligand. In all the three cases RMSE obtained from (Table 5) the docked ligand were higher than

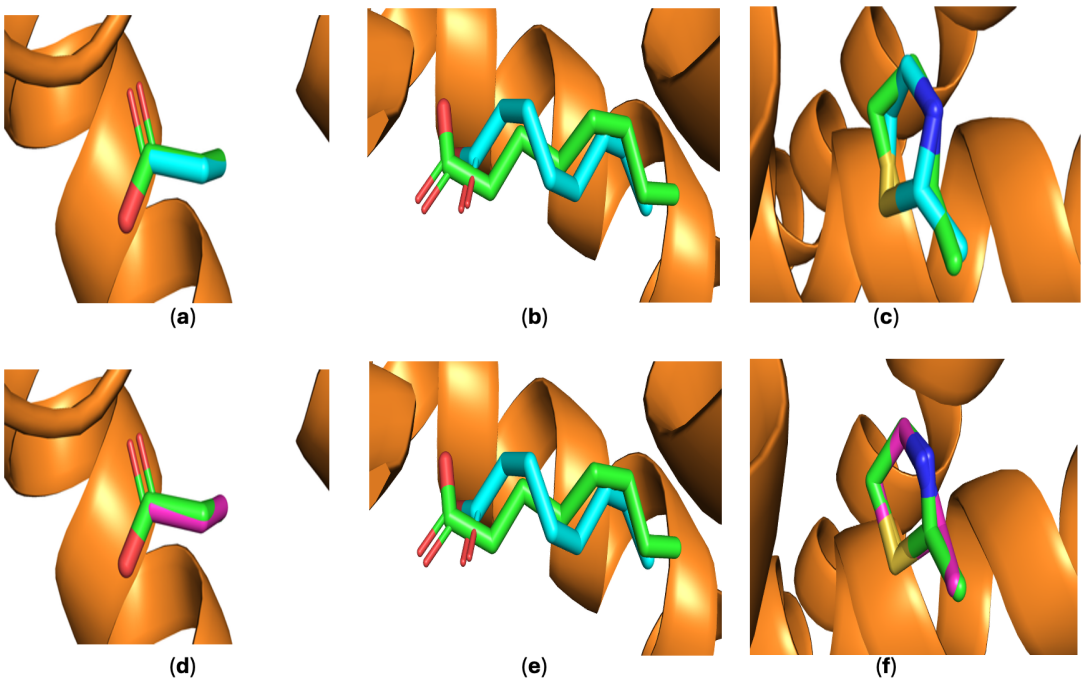


Figure 5: The first three panels (a-c) represent the alignment of conformers (**blue**) with the actual crystallographic ligand (**green**) where the conformers are selected via the true label proximity method while the conformers (**purple**) shown from panels (d-f) are selected via the predicted label proximity method. In case of propionate (a,d) both the conformers are nearly identical to the active conformation (**8F76** protein [**copper**]), similar thing is seen for 2-methylthiazoline (e,f) (**8UYQ** protein [**copper**]). In case of ocatanoic acid (b,e) the conformer obtained via true label and predicted label proximity is same. The conformer is structurally very similar to the active conformation (**8HTI** protein [**copper**]).

the ConfDENSE selected conformers. The conformers selected by ConfDENSE algorithm showed higher structural similarity than docking algorithm (Fig. 6).

In case of the linear molecules: propionate and octanoate, even after the Kabsch alignment the docking pose is not similar to the actual ligand. This is due to the fact, docking algorithm tend to struggle with molecule with high conformational flexibility, this is quite apparent for the octanoic molecule[32].[33] In case of 2-methylthiazoline where the conformational flexibility is limited the docking algorithm does a better job in reproducing the active conformation of the molecule. We could only evaluate three crystal structures, due to paucity of crystal structures of human olfactory receptors. Hopefully in the future there are more crystal structures available to conduct indepth analysis for our ConfDENSE model.

Table 5: Kabsch alignment score (rmsd) for the docked conformers of propionate, octanoate and 2-methylthiazoline.

Molecule	RMSD
Propionic acid	0.87 Å
Octanoic acid	1.8 Å
2-methylthiazoline	1.24 Å

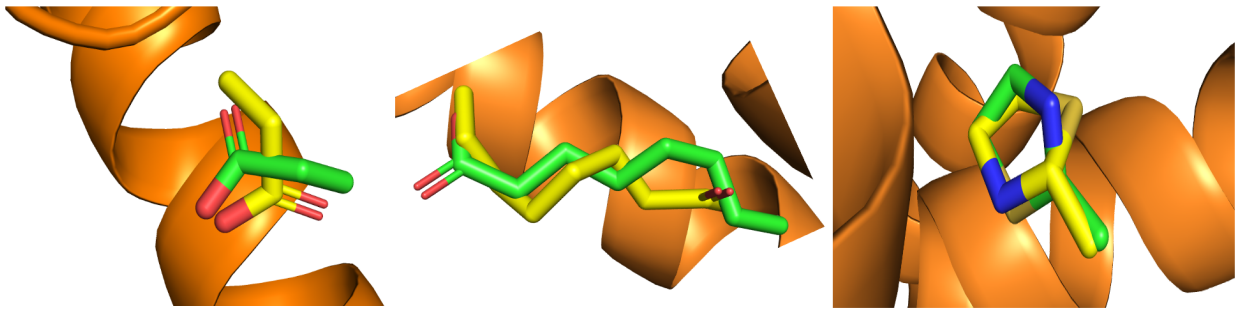


Figure 6: The figure depicts alignment of docked ligands a.) Propionate (**8F76**); b.) Octanoate (**8HTI**); c.) 2-methylthiazoline(**8UYQ**). Green ligand represent the actual crystallographic structure, while the yellow ligand represents the docked ligand.

The results obtained by ConfDENSE model is quite interesting given the fact that no prior of protein structure is provided to the model. We hypothesized the performance of model is due to the fact that promolecule electron density provides a realistic representation of a molecule's surface the 3-dimensional shape and anisotropic electron clouds that a receptor encounters during binding. Unlike atom centered descriptors, the electron density isosurface captures curvature and steric contours in continuous space. These features define how well a ligand can complement the steric and electrostatic landscape of a receptor pocket. The bioactive conformer is typically the one whose electron-density surface aligns most closely with the pocket's shape.

ConfDENSE leverages this principle by evaluating each conformer through its electron-density point cloud and selecting the one whose surface most plausibly matches a protein binding pocket, even without knowing the receptor structure. By learning these surface compatibility cues directly from electron density, ConfDENSE reliably identifies the conformer that best reflects the true binding-competent state, closely matching experimentally

observed bioactive geometries.

Explainability Analysis

We predicted the bioactive conformations of ligands without requiring protein structural information. We also explored whether we could identify the specific ligand regions involved in binding, which could guide the design of novel odorants by preserving chemotypes that interact with olfactory sites.

For interpretability, we applied the **Integrated Gradients (IG)** method,[34] which computes the path-integral of gradients from a baseline input to the actual input to quantify how each feature contributes to a model's prediction. We performed this explainability analysis on our GNN model, **openPOM**,[21] as it incorporates atom and bond level features, making the resulting interpretations more intuitive and chemically meaningful.

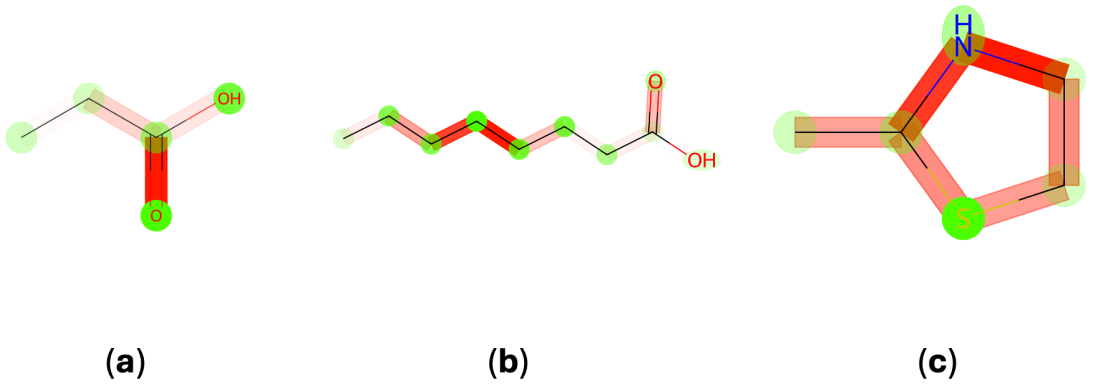


Figure 7: Explainability analysis of propionate, octanoate and 2-methylthiaoline, highlighting the atoms and bonds responsible for binding with olfactory receptors

For propionate and octanoate (Fig. 7), the model strongly highlights the carboxylate

group in propionate, but only weakly in octanoate. Interestingly, for octanoate, the carbon chain region (C3–C5) shows prominent attribution. We also examined the crystal structures of the three aforementioned ligands bound to their respective olfactory receptors (**8F76**, **8HTI** and **8UYQ**).

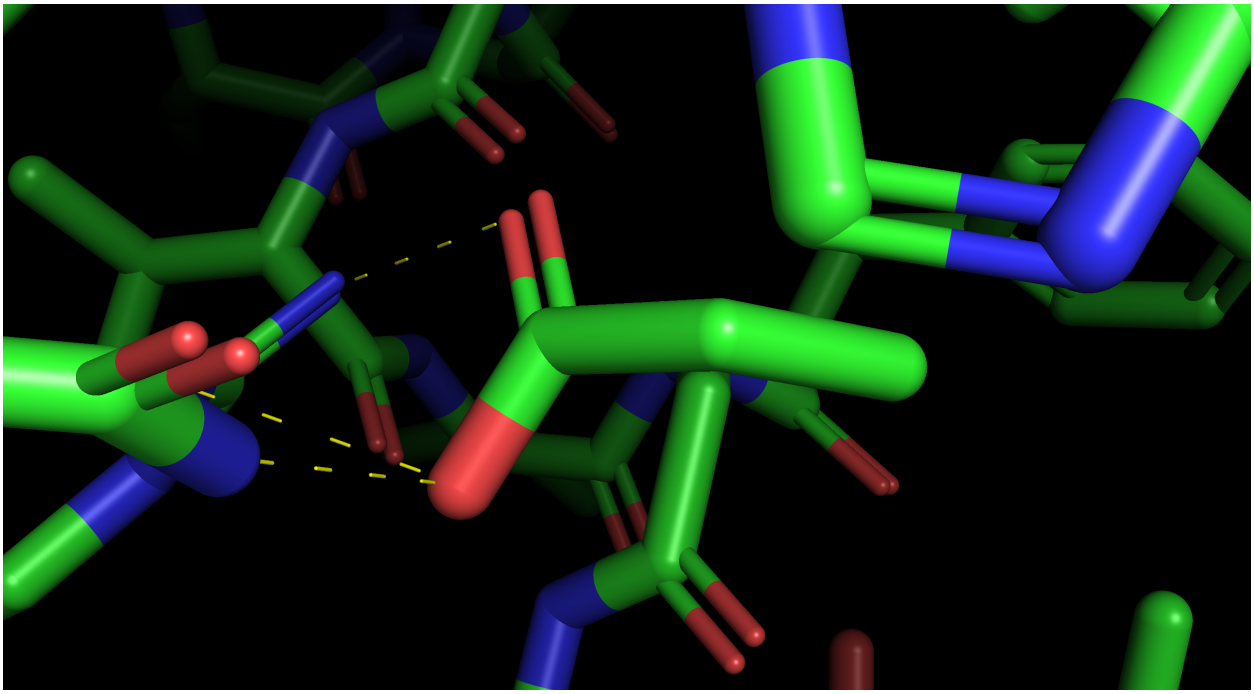


Figure 8: The interaction of ligand with residues present in olfactory receptors are shown for propionate **8F76**

We used the PyMOL visualizer[35] to examine the interactions between the ligands and the amino-acid residues of the olfactory receptors. In the propionate crystal structure (**8F76**), the carboxylate group forms contacts with arginine (R262) and glutamine (Q181), consistent with our explainability analysis, which also highlights the carboxylate moiety as a key interacting region (Fig. 8).

In the octanoate explainability analysis, the carboxylate group is only weakly highlighted, whereas the C3–C5 atoms show strong attribution. The crystal structure of the octanoate–OR complex (**8HTI**) supports this: the C3 and C4 atoms participate in hydrophobic interactions with two phenylalanine residues, F161 and F261, respectively (Fig. 9), while the carboxylate group forms a hydrogen bond with arginine (R265). These struc-

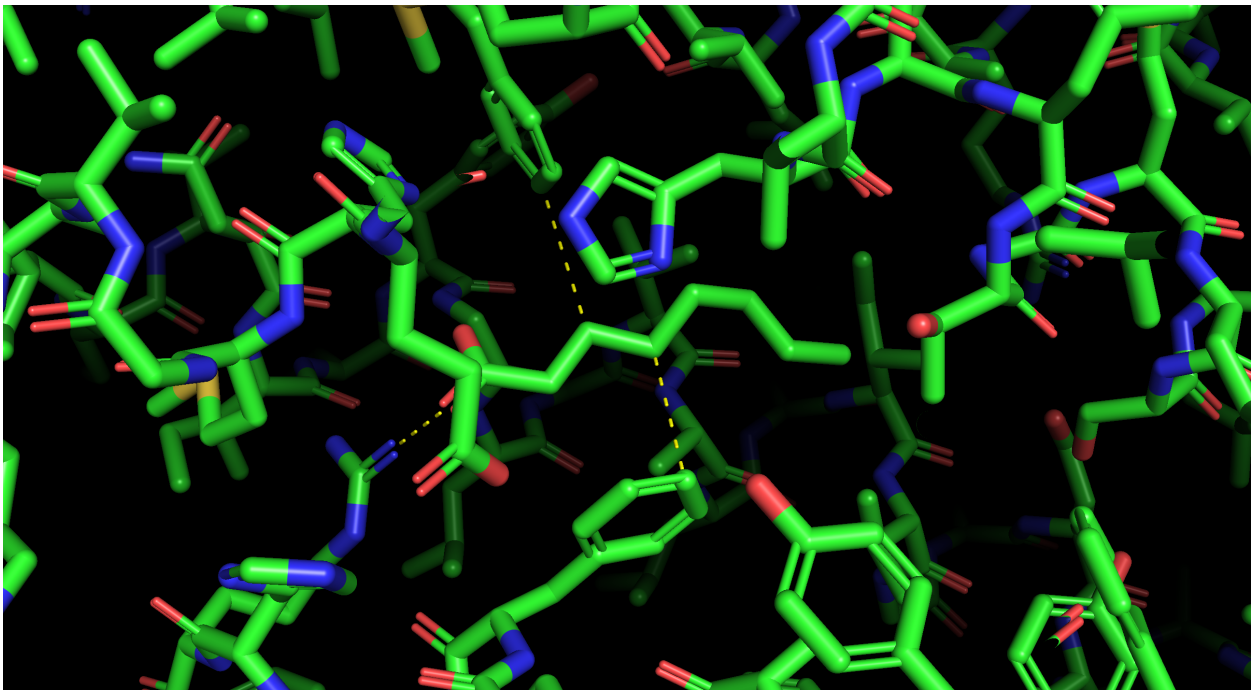


Figure 9: Aryl-alkyl interactions in the octaonate-OR crystal complex **8HTI**

tural observations are consistent with the explainability results.

In the explainability analysis for 2-methylthiazoline, the nitrogen atom of the heterocyclic ring is prominently highlighted. Examination of the crystal structure of 2-methylthiazoline bound to its olfactory receptor (**8UYQ**) confirms this: the heterocyclic nitrogen forms a hydrogen bond with the tyrosine residue Y257 (Fig.10).

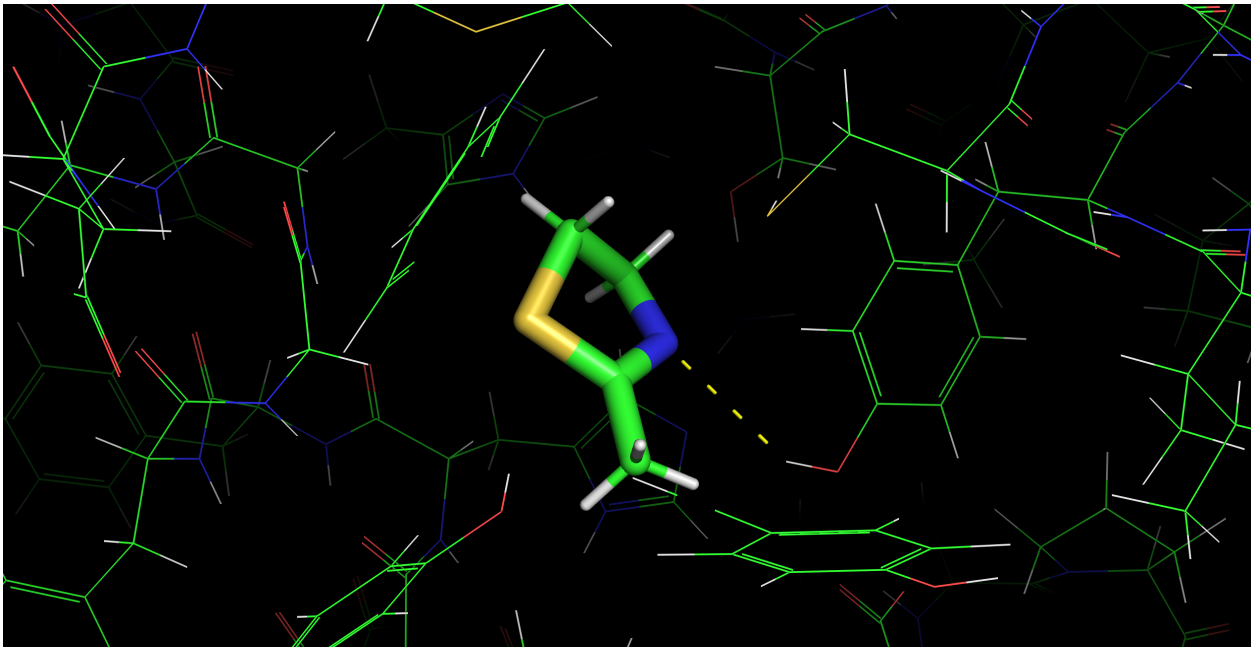


Figure 10: Hydrogen bonding between 2-methylthiazoline and Tyr 257 residue in **8UYQ** crystral structure

Across propionate, octanoate, and 2-methylthiazoline, the explainability analysis consistently identified the ligand features most critical for receptor binding, and these predictions were strongly supported by the corresponding crystal structures. Crucially, these results show that ligand based machine learning models capable of predicting bioactive conformations without requiring receptor structures can still uncover chemically meaningful interaction determinants when combined with explainability tools. This is a significant advance, as it enables the identification of key binding driving substructures even in the absence of experimental or computational receptor models.

Moreover, this combined framework: bioactive conformation prediction plus explainability offers substantial advantages for pharmacophore modeling and scaffold based design. By revealing which functional groups or molecular regions consistently contribute to binding across chemical series, the approach can help define robust pharmacophoric features, guide scaffold hopping, and support the rational design of new odorants or ligands that preserve essential interaction motifs. In settings where receptor structures are unavailable or difficult to resolve, this capability represents a powerful and practical route for accelerating ligand

discovery and optimization.

Conclusion

In this work, we introduced ConfDENSE, a Set2Set-coupled PointNet architecture that leverages Hirshfeld promolecule electron densities and conformer ensembles to explore the olfactory chemical space. By avoiding conventional downsampling and instead adopting a masking-based zero-padding strategy, our method preserves the complete electron-density information for each molecule while enabling PointNet to operate efficiently on variable sized point clouds. The addition of a Set2Set aggregation mechanism allows the model to identify and prioritize conformers that most strongly influence odor predictions, thereby aligning the machine-learning framework more closely with the biophysical reality of odorant–receptor interactions.

While the ConfDENSE model does not yet surpass state of the art MPGNN-based approaches such as DenseSense or openPOM in overall AUROC, it introduces a fundamentally different paradigm: one that explicitly incorporates 3D electron density and conformational diversity into olfaction modeling. Importantly, our conformer analysis reveals that the model is remarkably successful at identifying bioactive conformers. For the only experimentally validated human odorant receptors available: OR51E2 and consORs, the conformers selected via Set2Set aggregation or through true-label proximity exhibit exceptionally low RMSD values relative to crystal-bound ligands and outperformed standard molecular docking. These results strongly indicate that ConfDENSE captures structural cues relevant to receptor recognition, despite never being trained on receptor structures.

Moreover, combining ConfDENSE's ligand-based bioactive-conformation prediction with explainability analysis provides a powerful avenue for **pharmacophore** and **scaffold-based modeling**. By highlighting functional groups and substructures most responsible for receptor engagement, this framework can reveal conserved binding motifs even when receptor

structures are unavailable. This enables rational scaffold hopping, the construction of more precise pharmacophore hypotheses, and systematic exploration of odorant chemical space grounded in biophysically meaningful interaction patterns. In domains such as olfaction, where structural data remain limited, this ligand-centric interpretability offers a practical and impactful route for guiding molecular design.

This work demonstrates that integrating electron density representations with geometric deep learning can provide physiochemically grounded insights into olfactory perception and molecular recognition. As more human olfactory receptor structures become available, ConfDENSE offers a promising framework for linking molecular electron distribution, conformational preferences, and receptor binding. Overall, ConfDENSE establishes a new direction for olfaction research, laying the groundwork for more interpretable and biophysically faithful machine-learning models in chemical sensing.

Data and Software Availability

All data and code supporting this study are openly available in public repositories. The machine learning codes and analysis scripts for ConfDENSE are available at `https://github.com/CSIO-FPIL/PointNET-ODOR`.

References

(1) Saha, P.; Rahane, A.; Kumar, V.; Sukumar, N. Analysis of the electron density features of small boron clusters and the effects of doping with C, P, Al, Si, and Zn: Magic B7P and B8Si clusters. *Physica Scripta* **2016**, *91*, 053005.

(2) Saha, P.; Rahane, A. B.; Kumar, V. Electronic origin of the stability of transition-metal-doped B14 drum-shaped boron clusters and their assembly into a nanotube. *The Journal of Physical Chemistry C* **2017**, *121*, 10728–10742.

(3) Saha, P. Application of Electron Density Based Analysis in the Study of Nanoclusters and Biomolecular Interactions. *Shiv Nadar University Thesis* **2018**,

(4) Vanommeslaeghe, K.; Guvench, O.; others Molecular mechanics. *Current pharmaceutical design* **2014**, *20*, 3281–3292.

(5) Ward, L.; Wolverton, C. Atomistic calculations and materials informatics: A review. *Current Opinion in Solid State and Materials Science* **2017**, *21*, 167–176.

(6) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angewandte Chemie International Edition in English* **1993**, *32*, 503–527.

(7) Baskin, I. I.; Winkler, D.; Tetko, I. V. A renaissance of neural networks in drug discovery. *Expert opinion on drug discovery* **2016**, *11*, 785–795.

(8) Ekins, S. The next era: deep learning in pharmaceutical research. *Pharmaceutical research* **2016**, *33*, 2594–2603.

(9) Bader, R. F. Atoms in molecules. *Accounts of chemical research* **1985**, *18*, 9–15.

(10) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, *44*, 129–138.

(11) Williams, L.; Mukherjee, A.; Dasgupta, A.; Rajan, K. Monitoring the role of site chemistry on the formation energy of perovskites via deep learning analysis of Hirshfeld surfaces. *Journal of Materials Chemistry C* **2021**, *9*, 11153–11162.

(12) Saha, P.; Nguyen, M. T. Electron density mapping of boron clusters via convolutional neural networks to augment structure prediction algorithms. *RSC advances* **2023**, *13*, 30743–30752.

(13) Qi, C. R.; Su, H.; Mo, K.; Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; pp 652–660.

(14) Kim, S.; Lee, B. D.; Cho, M. Y.; Pyo, M.; Lee, Y.-K.; Park, W. B.; Sohn, K.-S. Deep learning for symmetry classification using sparse 3D electron density data for inorganic compounds. *npj Computational Materials* **2024**, *10*, 211.

(15) Niimura, Y.; Matsui, A.; Touhara, K. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome research* **2014**, *24*, 1485–1496.

(16) Sell, C. S. On the unpredictability of odor. *Angewandte Chemie International Edition* **2006**, *45*, 6254–6261.

(17) Billesbølle, C. B.; de March, C. A.; van der Velden, W. J.; Ma, N.; Tewari, J.; Del Torrent, C. L.; Li, L.; Faust, B.; Vaidehi, N.; Matsunami, H.; others Structural basis of odorant recognition by a human odorant receptor. *Nature* **2023**, *615*, 742–749.

(18) Berwal, B.; Saha, P.; Kumar, R. A Fully In Silico Protocol to Understand Olfactory Receptor–Odorant Interactions. *ACS Omega* **2025**,

(19) Lee, B. K.; Mayhew, E. J.; Sanchez-Lengeling, B.; Wei, J. N.; Qian, W. W.; Little, K. A.; Andres, M.; Nguyen, B. B.; Moloy, T.; Yasonik, J.; others A principal odor map unifies diverse tasks in olfactory perception. *Science* **2023**, *381*, 999–1006.

(20) Burns, J. W.; Rogers, D. M. QuantumScents: Quantum-mechanical properties for 3.5 k olfactory molecules. *Journal of Chemical Information and Modeling* **2023**, *63*, 7330–7337.

(21) Saha, P.; Sharma, M.; Balaji, S.; Barsainyan, A. A.; Kumar, R.; Steuber, V.; Schmuker, M. DENSE SENSE: A novel approach utilizing an electron density augmented machine learning paradigm to understand a complex odour landscape. *Digital Discovery* **2025**,

(22) Tarekegn, A. N.; Giacobini, M.; Michalak, K. A review of methods for imbalanced multi-label classification. *Pattern Recognition* **2021**, *118*, 107965.

(23) Momma, K.; Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Applied Crystallography* **2011**, *44*, 1272–1276.

(24) McKinnon, J. J.; Jayatilaka, D.; Spackman, M. A. Towards quantitative analysis of intermolecular interactions with Hirshfeld surfaces. *Chemical Communications* **2007**, 3814–3816.

(25) Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision. 2015; pp 4489–4497.

(26) Lawrence, J.; Bernal, J.; Witzgall, C. A purely algebraic justification of the Kabsch-Umeyama algorithm. *Journal of research of the National Institute of Standards and Technology* **2019**, *124*, 1.

(27) Kochnev, Y.; Hellemann, E.; Cassidy, K. C.; Durrant, J. D. Webina: an open-source library and web app that runs AutoDock Vina entirely in the web browser. *Bioinformatics* **2020**, *36*, 4513–4515.

(28) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010**, *31*, 455–461.

(29) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019; pp 2623–2631.

(30) Choi, C.; Bae, J.; Kim, S.; Lee, S.; Kang, H.; Kim, J.; Bang, I.; Kim, K.; Huh, W.-

K.; Seok, C.; others Understanding the molecular mechanisms of odorant binding and activation of the human OR52 family. *Nature Communications* **2023**, *14*, 8105.

(31) de March, C. A.; Ma, N.; Billesbølle, C. B.; Tewari, J.; Llinas del Torrent, C.; van der Velden, W. J.; Ojiro, I.; Takayama, I.; Faust, B.; Li, L.; others Engineered odorant receptors illuminate the basis of odour discrimination. *Nature* **2024**, *635*, 499–508.

(32) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: a multistep strategy approach. *Proteins: Structure, Function, and Bioinformatics* **1999**, *36*, 1–19.

(33) Huang, S.-Y. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Briefings in bioinformatics* **2018**, *19*, 982–994.

(34) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. International conference on machine learning. 2017; pp 3319–3328.

(35) DeLano, W. L.; others Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein crystallogr* **2002**, *40*, 82–92.