
Can Transformers Smell Like Humans?

Farzaneh Taleb

Department of Intelligent Systems
KTH Royal Institute of Technology
Stockholm, Sweden
farzantn@kth.se

Miguel Vasco

Department of Intelligent Systems
KTH Royal Institute of Technology
Stockholm, Sweden
miguelsv@kth.se

Antônio H. Ribeiro

Department of Information Technology
Uppsala University
Uppsala, Sweden
antonio.horta.ribeiro@it.uu.se

Mårten Björkman

Department of Intelligent Systems
KTH Royal Institute of Technology
Stockholm, Sweden
celle@kth.se

Danica Kragic

Department of Intelligent Systems
KTH Royal Institute of Technology
Stockholm, Sweden
dani@kth.se

Abstract

The human brain encodes stimuli from the environment into representations that form a sensory perception of the world. Despite recent advances in understanding visual and auditory perception, olfactory perception remains an under-explored topic in the machine learning community due to the lack of large-scale datasets annotated with labels of human olfactory perception. In this work, we ask the question of whether pre-trained transformer models of chemical structures encode representations that are aligned with human olfactory perception, i.e., *can transformers smell like humans?* We demonstrate that representations encoded from transformers pre-trained on general chemical structures are highly aligned with human olfactory perception. We use multiple datasets and different types of perceptual representations to show that the representations encoded by transformer models are able to predict: (i) labels associated with odorants provided by experts; (ii) continuous ratings provided by human participants with respect to pre-defined descriptors; and (iii) similarity ratings between odorants provided by human participants. Finally, we evaluate the extent to which this alignment is associated with physicochemical features of odorants known to be relevant for olfactory decoding.

1 Introduction

The human brain receives sensory input from the environment and encodes it into a high-dimensional representation space, forming a perception of the world [1]. Recent studies have significantly improved our understanding of the underlying mechanisms of visual, linguistic, and auditory perception [2–4]. Indeed, there is a significant level of alignment between human response (from neuron to behavior) and activations of deep neural networks when provided with the same stimuli [5–14].

Despite these recent advances, human olfactory perception remains an under-explored topic. There is no single organizing principle that determines the dimensions of odor space, making the characteriza-

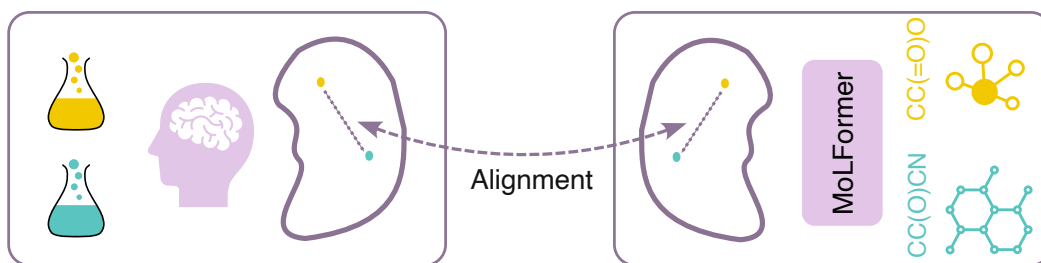


Figure 1: **Evaluating representational alignment** between human and pre-trained transformers. Human participants are stimulated with two odorant substances and asked to rate the perceptual similarity between them (Left). We encode representations of the same pair of odorants using MoLFormer and compute the similarity between pairs of representations (Right). Finally, we measure the alignment between the two systems.

tion of odor perception and its relation to chemical compounds an open and complex problem [15]. A lack of universally accepted methods to describe odorants either quantitatively or qualitatively makes this problem even more challenging. There are very few studies that have explored the mapping of chemical structures to olfactory perception [16–19]. In addition, processing chemical olfactory stimuli using deep neural networks has not been extensively investigated. Nevertheless, training the existing supervised models usually requires an extensive effort by experts to label data.

Transformer-based models [20, 21] are a breakthrough in machine learning, surpassing the need for extensive labeling by utilizing implicit supervision without the necessity for direct labels. These models have demonstrated impressive performance in various tasks such as image [22], video [23], and natural language processing [24]. More recently, they have also shown promising results in encoding chemical structures [25].

In this work, we ask the question of whether representations of odorant chemical structures extracted from transformers pre-trained on chemical structures align with human olfactory perception or, in other words, *can transformers smell like humans?* We employ MoLformer [25], a state-of-the-art transformer, which is pre-trained on chemical structures and we show that representations of odorants extracted from this model:

- can predict *labels assigned to odorants* by experts (Section 4.1);
- can predict *continuous perceptual ratings* provided by human participants (Section 4.2);
- present a high correlation index with *human perceptual similarity ratings* (Section 4.3);
- present a high correlation index with *physicochemical descriptors* known to be relevant for olfactory perception (Section 4.4).

Surprisingly these results hold for models that *were not explicitly trained for the purpose of predicting the human olfactory experience*. To the best of our knowledge, we provide the first empirical study on evaluating the alignment between odorant representations encoded by transformers and human olfactory perception.

2 Related Work

The availability of larger datasets, together with advances in predictive methods, has led to an increasing interest in the prediction of olfactory perception from molecular structures.

Olfactory perception prediction. Learning predictive models of olfactory from molecular structures has been addressed mostly by the neuroscience community. Several works used standard chemoinformatic representations of molecules to model olfactory perception [17, 19, 18]. Specifically, Snitz et al. [18] proposed a computational framework and algorithm based on structural features of molecules to predict perceptual similarities between odorant pairs. This algorithm leverages feature engineering to identify the most relevant subset of features among 1433 physicochemical descriptors to predict pair-wise odorant perceptual similarities.

Later, Ravia et al. [17] extended this model to also include the perceived intensity of molecular components. They employed 21 physicochemical descriptors discovered in previous works and proposed a weighting approach for multicomponent odorants (MC-odorants) based on their perceived intensity. They reported a higher correlation when employing the weighting approach compared to using the same model without it. However, the representation and generalization capabilities of these models are quite limited and unexplored.

Deep neural networks for odorants. Recently, Lee et al. [16] proposed a novel representation learning model of odorants, based on a message-passing graph neural network [26], which they refer to as Principal Odor Map (POM). To train this model, they curated and merged data from Leffingwell [27] and GoodScent [28] databases to compile a dataset of about 5000 molecules with 138 expert-labeled odor descriptors. This model outperforms the baselines in multiple odor prediction tasks and shows a relatively high alignment with human ratings in describing odorants. Nevertheless, training this model requires labeled data, relying on subjective evaluations of numerous odorants by experts. Besides being time-consuming and laborious, this process can introduce subjective biases into the model, a concern magnified by our incomplete understanding of the foundational factors of olfactory perception.

Large-scale molecular models. Large-scale pre-trained models, often known as foundation models [21], have been recently explored to perform diverse tasks by leveraging large amounts of unlabeled data. MoLFormer [25] model has been proposed in the context of chemical prediction tasks, able to extract rich representations from chemical structures. MoLFormer consists of a transformer-based architecture, with linear attention and relative positional encodings. This model is trained using a self-supervised approach, on multiple datasets (e.g., the PubChem [29] and ZINC [30] datasets) on a masked token prediction loss.

3 Method

In this section, we provide a detailed description of the datasets utilized in this study and outline the methodology for extracting both odorant (machine) and perceptual (human) representations. Additionally, we present the main model and baseline methods employed, along with the evaluation metrics used to assess their performance. Our experiments do not require significant computational resources: we mostly train linear models that do not involve GPU usage or models that can be trained on a single commercially available GPU under one hour. All computational code to reproduce our results is available at <https://github.com/Farzaneh-Taleb/transformer-olfactory-alignment>

3.1 Datasets

We use the publicly available version of the following datasets provided by Pyrfume repository [31].

Leffingwell-Goodscent (GS-LF) [27, 28]. We employ a curated and merged version of the Goodscents [28] and Leffingwell [27] datasets, provided by [32], following the procedure introduced by Lee et al. [16]. This dataset contains 4983 molecules with 138 expert-labeled descriptors (e.g. creamy, grassy), where each odorant may be linked to multiple descriptors.

Sagar [33]. This dataset contains the rating of 160 odorants by 3 human participants, with respect to 15+3 perceptual descriptors. In addition to 15 common descriptors among participants, there are 3 more descriptors that vary among them. We excluded these variable descriptors and focused solely on the common descriptors among the participants. The provided ratings were already normalized within the range of [-1, 1] and the mean ratings across all the subjects are computed for subsequent analysis.

Keller [34]. This dataset contains ratings of 480 structurally and perceptually diverse molecules by 55 human participants, evaluating 23 descriptors. Participants were instructed to adjust a slider to rate odorants according to individual descriptors, with the slider position subsequently translated into a scale ranging from 0 to 100. Ratings were then averaged across all participants for further analysis.

Ravia [17]. This dataset contains similarity ratings of 195 unique pairs of MC-odorants and mono-molecules by 94 participants. The similarity values were averaged across all the participants. In this work, we disregarded the factor of odorant intensity and averaged similarity ratings based on the unique pairs of odorants.

Snitz [18]. This dataset includes similarity ratings from 139 participants and 359 unique pairs of odorants. In each trial, participants were presented with two distinct odorants and asked to rate the degree of similarity in their smells. These ratings were then averaged across all participants.

3.2 Odorant representations

Odorants can be described as a single molecule or a mixture of molecules, which we denote as multicomponent odorants (MC-odorants). In this section we describe the method to extract odorant representations from the main pre-trained model (MoLFormer) and our baseline models (DAM and Open-POM).

MoLFormer. We employ MoLFormer [25] to encode SMILES strings associated with a single molecule and extract a 768-dimensional vector from the last layer of the model. SMILES (simplified molecular-input line-entry system) is a string-based representation that encodes relevant chemical information such as the type of atoms, their bonds, and the substructures present in the molecule. For MC-odorants, we average the extracted representations across all available mono-molecule components within that MC-odorant. The odorant representation for each dataset is a matrix of $X_i \in \mathbb{R}^{n \times 768}$ where n is the number of unique odorants.

Open-POM. The principal odor map (POM) is a supervised-learning model, based on a message-passing graph neural network [26], which is trained on the GS-LF datasets to predict odorant perceptual labels. We employ a publicly available version of this model, which we denote by Open-POM [32]. We train Open-POM for 150 epochs, using 30 different train-test splits, and we extract representations from the penultimate layer of this model. The odorant representation for each dataset is a matrix of $X_i \in \mathbb{R}^{n \times 256}$ where n is the number of unique odorants. For MC-odorants, we average the representations extracted for each individual molecule within the mixture.

Distance Angle Model (DAM). Snitz et al. [18] proposed a distance angle model (DAM) that uses 21 physicochemical descriptors to predict the similarities between pairs of odorants. We extract these 21 descriptors for each odorant using AlvaDesc [35] and discard 6 of them due to NaN values produced by this software. We use the remaining subset of 15 physicochemical descriptors out of 21 to measure similarity between odorants or train a linear mapping from them to the perceptual representation space. The odorant representation for each dataset is a matrix of $X_i \in \mathbb{R}^{n \times 15}$ where n is the number of unique odorants. As suggested by Ravia et al. [17], we average the representations extracted for each individual molecule within the mixture to compute representations for MC-odorants.

3.3 Perceptual representations

Perceptual representations of odorants are provided by human participants when exposed to odorant stimuli. Perceptual olfactory data were collected in one of the following ways:

1. Experts label the odorants, where each odorant may be linked to multiple labels (e.g., [27, 28]). The perceptual representation is a matrix of $y_i \in \{0, 1\}^{n \times d}$ where n is the number of unique odorants and d is the number of classes.
2. Non-expert participants provided ratings with regards to a set of predefined descriptors (e.g., [33, 34].) In this case, the averaged perceptual representations over participants and replicas form a matrix of $y_i \in [a, b]^{n \times d}$, where n is the number of unique odorants, d is the number of descriptors, and a and b denote the minimum and maximum values participants can use to describe the odorants with respect to these descriptors.
3. Participants evaluated the perceived similarity between pairs of odorants (e.g., [18, 17]). In this case, the averaged perceptual representations over participants and replicas are a vector of $y_i \in [a, b]^{n \times 1}$ where n is the number of unique "pair of odorants" and a and b indicate the range of values participants can use to rate the odorants' similarity with respect to the descriptors.

3.4 Alignment between perceptual and odorants representations

We measure the similarity between two representation spaces directly when it is possible (Section 4.3), otherwise we train a linear model to predict perceptual representations for each odorant (Section 4.1, 4.2). We use nested 5-fold cross-validation to tune the hyper-parameters of the linear models

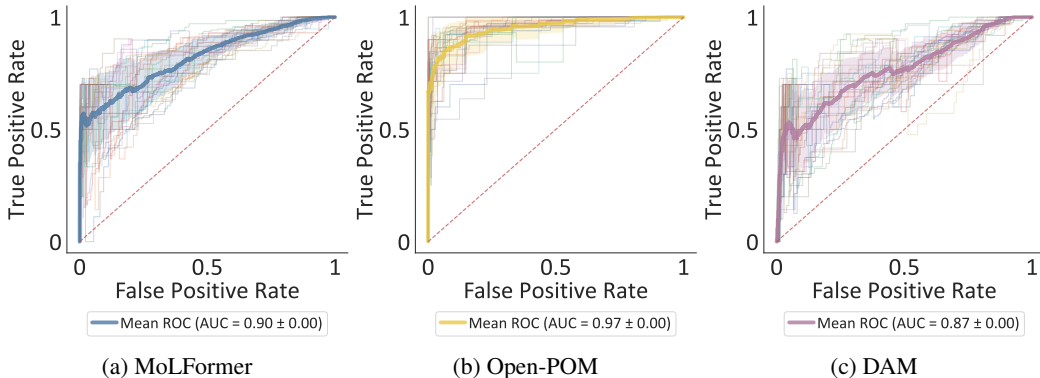


Figure 2: **ROC curve for linear classifiers** trained on GS-LF representations extracted from three different models. Each curve corresponds to a separate test split, with the thicker curve representing the average performance across all splits. We highlight that MoLFormer outperforms DAM, despite not being trained to predict perceptual labels but does not achieve the performance level of Open-POM, which demonstrates the highest performance. The chance level is shown with red dashed line.

and assess evaluation metrics on the test set that was held out during the training phase using an 80%-20% train-test split. This process is repeated 30 times using 30 different train-test splits.

3.5 Evaluation metrics

In this section, we introduce the main evaluation metrics to measure the alignment in this paper.

Micro-averaged ROC-AUC score. The micro-averaged ROC-AUC score was computed to assess the performance of each model for the multi-label classification task. The micro-averaged ROC-AUC score is computed by aggregating true positive, false positive, true negative, and false negative values across all classes.

Normalized Root Mean Squared Error (NRMSE). The root mean squared error (RMSE) is the difference between the observed values and predicted ones for the regression task. Here, we normalize it by the range of true observations – i.e., $\text{NRMSE} = \text{RMSE} / (\max(y) - \min(y))$.

Pearson Correlation Coefficient (CC). We report the Pearson correlation coefficient between predicted results and real values. It measures the linear correlation between two sets of data and is the ratio between the covariance of two variables and the product of their standard deviations.

4 Results

In this section, we evaluate whether the representations encoded by pre-trained models of chemical data can predict the human olfactory experience *despite not being explicitly trained for this purpose*. First, we focus on a subset of experiments aimed at predicting expert-assigned labels from odorants through linear mapping from representations to perceptual descriptors (Section 4.1). Subsequently, we aim to predict continuous scores provided by human participants (Section 4.2). Finally, we seek to predict the direct similarity scores from the representations extracted from odorants (Section 4.3). Additionally, we provide insights into the potential reasons underlying the observed alignments (Section 4.4).

4.1 Expert-assigned labels classification

To assess the performance of MoLFormer in predicting expert-assigned labels for odorants, we implemented a linear mapping from the representations extracted by MoLFormer to the odorant representations extracted from GS-LF dataset. First, the dimensionality of the extracted representations is reduced to 20 using PCA, followed by z-scoring of each feature. Then, we train individual logistic regression models for each descriptor. This process was repeated 30 times, each with a different train-test split, to quantify the uncertainty of the results.

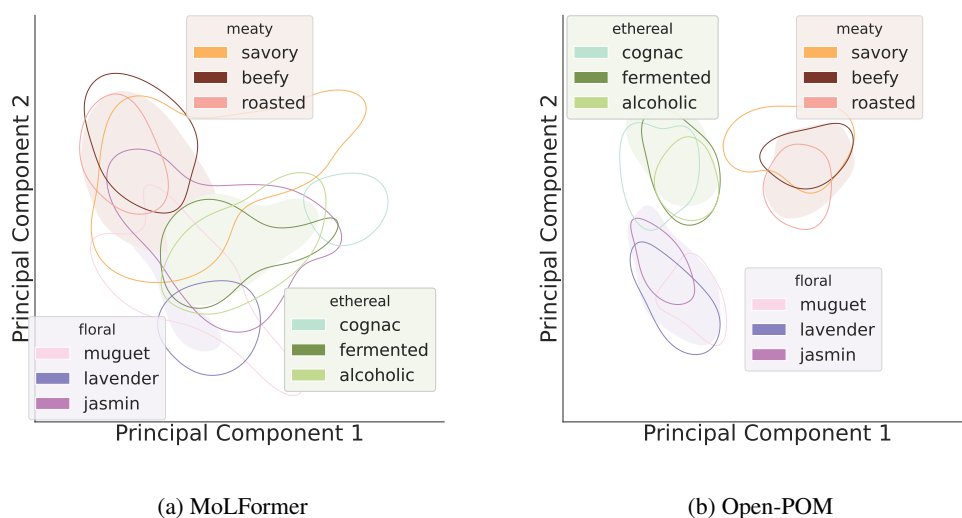


Figure 3: **Visualization of odorant representations** encoded by different models on the GS-LF dataset using the figure layout suggested by [16]. We plot the first and second principal components (PCs) of the representation spaces. Areas dense with molecules that have broad category labels (floral, meaty, or ethereal) are shaded, while areas dense with narrow category labels are outlined. MoLFormer captures the perceptual relationship between different odorants in its representation space, despite not being explicitly trained for this purpose.

We apply the same procedure without dimensionality reduction to DAM model representations. For the Open-POM model, which is already trained end-to-end and supervised on the same dataset, we directly extracted the predictions for the test set without retraining the model. As shown in Figure 2 the MoLFormer model achieves high ROC-AUC scores in odorant classification, outperforming the DAM model, which is trained using 15 physicochemical descriptors. However, the performance of MoLFormer is lower than that of Open-POM, which is trained end-to-end with supervision on the same dataset.

An additional experiment is conducted to understand the degree of perceptual details captured in the odorant representation space of MoLFormer by comparing odorant representations encoded by this model with the representations encoded by Open-POM. In Figure 3 we depict the first two principal components of the representations. We highlight the similarity between the representations encoded by both Open-POM and MoLFormer and observe that the latter is able to capture the perceptual relationship between different odorants despite not using any perceptual labels during training (unlike the supervised Open-POM Model).

4.2 Continues perceptual rating prediction

To evaluate the capabilities of the MoLFormer model to predict continuous rating scores with respect to pre-defined descriptors, provided by human participants, we train separate linear regression models with regularization applied using the Lasso penalty for each descriptor. Once again, the dimensionality of the extracted representations is reduced to 20 using PCA (for MoLFormer and Open-POM), followed by z-scoring of each feature. This procedure is repeated using 30 different train-test splits.

The results of these experiments are shown in Table 1 and Figure 4. Table 1 shows the average Pearson correlation coefficient and NRMSE across all descriptors, while Figure 4 presents the results for each individual descriptor. As shown in Table 1, overall, none of the models exhibit a high correlation. Nevertheless, MoLFormer slightly underperforms compared to Open-POM in both datasets. However, it performs better than DAM for the Keller dataset but worse than DAM for the Sagar dataset, where DAM even outperforms Open-POM.

Table 1: **Performance of the models to predict continuous ratings averaged across all perceptual descriptors.** We compute the average Pearson correlation coefficient (CC) and normalized root mean squared error (NRMSE) across all descriptors. MoLFormer shows slightly worse performance than Open-POM but better than DAM for the Keller dataset and worse than DAM for the Sagar dataset, where DAM outperforms Open-POM.

| | Keller | | | Sagar | | |
|------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | MoLFormer | Open-POM | DAM | MoLFormer | Open-POM | DAM |
| CC (\uparrow) | 0.20 \pm 0.00 | 0.22 \pm 0.01 | 0.17 \pm 0.00 | 0.25 \pm 0.01 | 0.29 \pm 0.01 | 0.35 \pm 0.01 |
| NRMSE (\downarrow) | 0.15 \pm 0.00 | 0.15 \pm 0.00 | 0.15 \pm 0.00 | 0.19 \pm 0.00 | 0.18 \pm 0.00 | 0.17 \pm 0.00 |

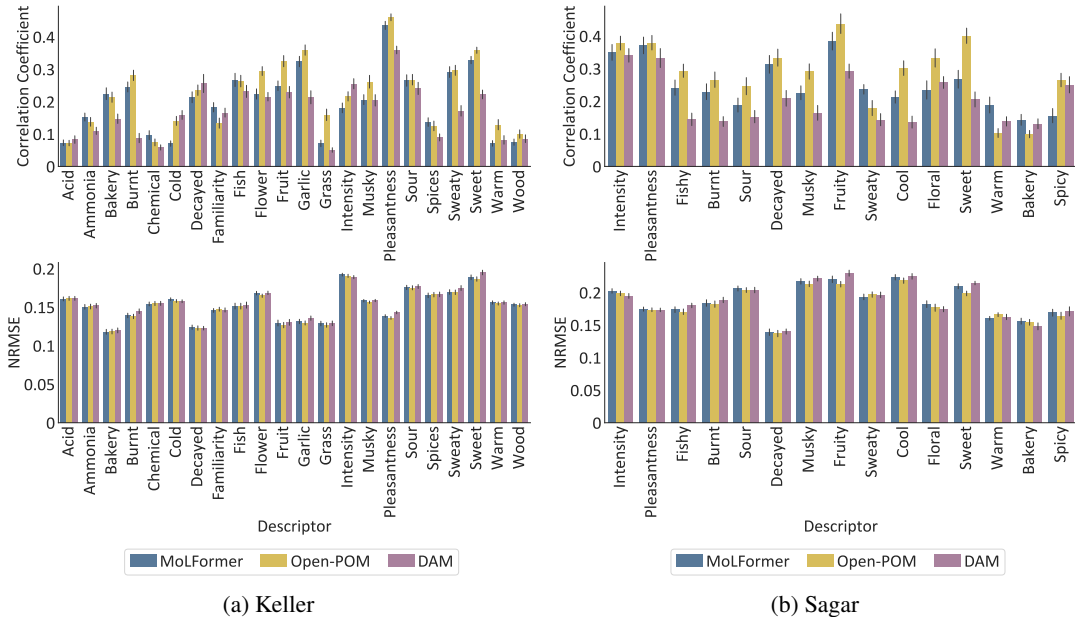


Figure 4: **Performance of the models to predict continuous ratings per descriptor.** We computed Correlation and NRMSE between predicted and actual ratings per perceptual descriptor. Despite not being trained to predict human olfactory labels, the MoLFormer model performs on par with the Open-POM and DAM models.

According to Figure 4 MoLFormer model performs on par with the Open-POM and DAM models, which are trained with supervision in predicting the rating for each descriptor. In summary, although, on average, MoLFormer performs slightly worse than Open-POM, it still demonstrates a similar degree of alignment, especially despite the absence of supervision in its training process.

4.3 Representational similarity analysis

In order to evaluate the direct alignment between the odorant similarities encoded by MoLFormer and those obtained from human participants, we separately encode each odorant by MoLFormer (and the baseline models) and compute the cosine similarity between the extracted representations. Subsequently, we compute the Pearson correlation between the similarity scores computed by the models and those provided by human participants in the Ravia and Snitz datasets. The results are presented in Figure 5a.

These results show that the MoLFormer is able to extract representations that encode information related to the human olfactory perception, despite not having access to that information during model training. We highlight a significant high correlation between perceptual and odorant representation for the Snitz ($r = 0.64, p < 0.0001$) and Ravia datasets ($r = 0.66, p < 0.0001$).

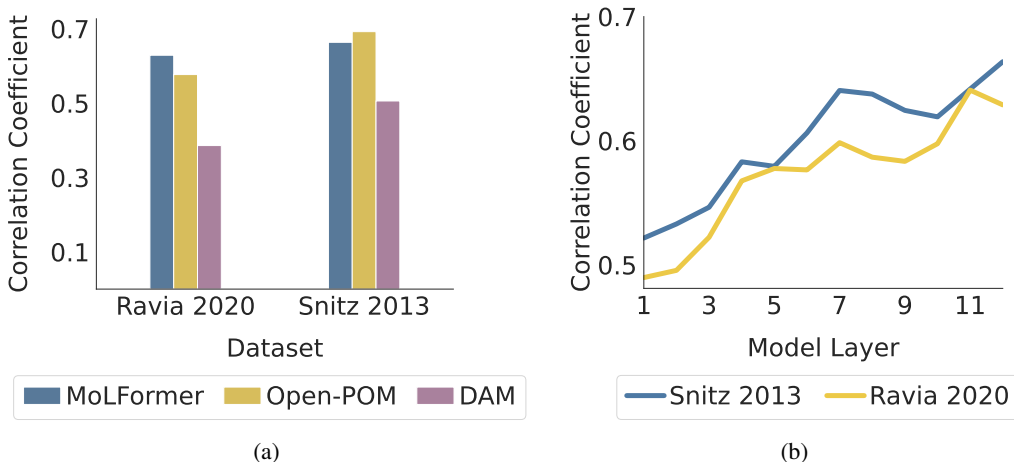


Figure 5: **Representational similarity analysis** for Snitz and Ravia datasets: a) Correlation coefficients between similarity scores provided by human participants and computed using representations encoded by the different models ; b) Correlation coefficients considering odorant representations extracted from different layers of the MoLFormer model.

The comparison with the baseline models indicates that it performs on par with the Open-POM model and significantly outperforms the DAM model. These results suggest that, despite being trained with some form of supervision, these models may struggle to effectively extract similarities between odorants. Additionally, the findings demonstrate that MoLFormer is more proficient at identifying similarities between pairs of odorants than mapping them to a set of predefined descriptors. This superior performance may be due to the model’s ability to capture a measure of similarity, as perceived by humans, rather than introducing subjective language bias associated with pre-defined descriptors.

Finally, we aim to evaluate whether the depth of the layer in the MoLFormer model, from which we extract the odorant representations, affects the representational alignment. To assess this, we repeat the described procedure in this section for each layer separately. As shown in Figure 5b, representational alignment improves with increasing layer depth, indicating that deeper layers of the transformer are more aligned with high-level perceptual representations.

4.4 Decoding relevant physicochemical features from pre-trained representations

To evaluate whether MoLformer effectively extracts features from chemical structures relevant to olfactory perception, we evaluate the alignment of MoLFormer with physicochemical descriptors that are used in the DAM model. To do so, we train 15 linear regression models, each one to predict a single physicochemical descriptor from the extracted representations of the MoLFormer. We subsequently evaluate the correlation between the predicted and true values. As shown in Figure 6, MoLformer demonstrates a high degree of alignment in predicting these values. Out of the 15 physicochemical descriptors, MoLformer successfully predicts the values for 13 descriptors as well as or better than the Open-POM model.

Next, we evaluate whether this alignment changes across the layers of MoLformer. Therefore, we repeat the same procedure for each layer separately. As illustrated in Figure 7, the alignment with the identified chemical features decreases with increasing layer depth. However, as demonstrated in Figure 5b, the alignment with perception improves. These results collectively are consistent with well-known principles in vision models, where the lower layers typically capture low-level, localized features like edges and textures, while deeper layers gradually shift toward higher-level, abstract representations, such as shapes and objects[36]. Nonetheless, additional investigation is required to fully reveal and comprehend this potential hierarchical structure.

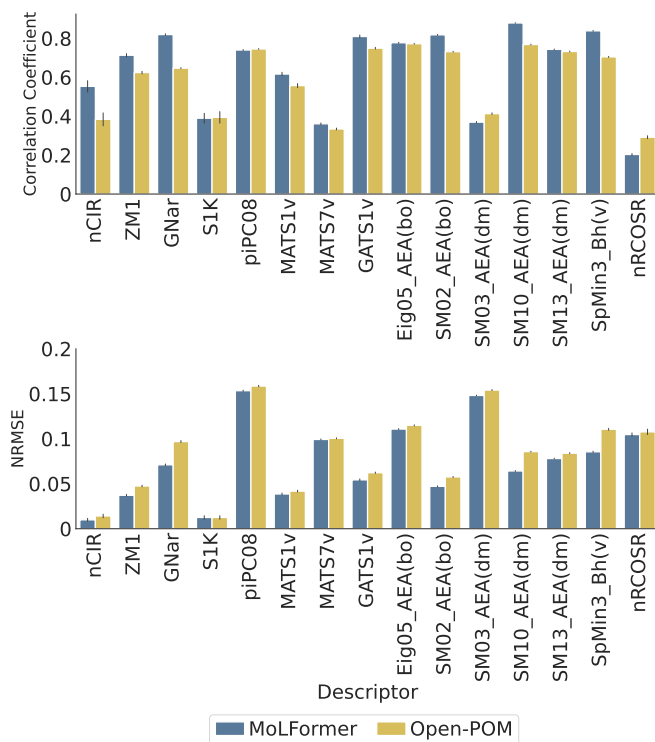


Figure 6: **Performance of the models to predict relevant physicochemical descriptors.** We computed Correlation and NRMSE between the predicted and actual values of descriptors. MolFormer is able to predict 13 out of 15 physicochemical descriptors related to smell as well as or better than the Open-POM model, demonstrating high alignment with physicochemical descriptors.

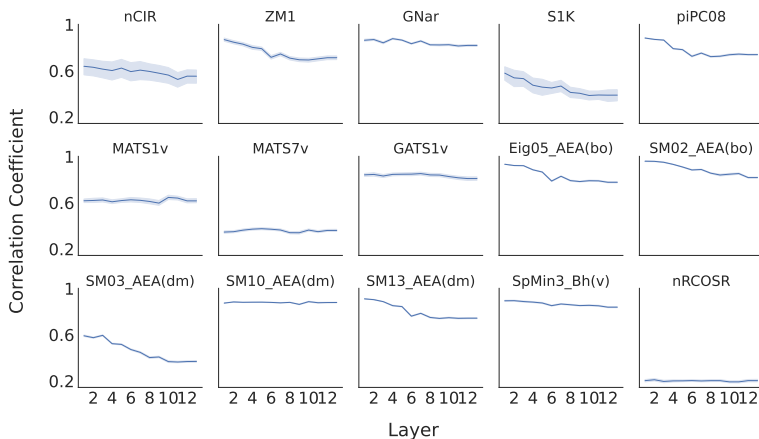


Figure 7: **Correlation between the actual and predicted value of physicochemical descriptors diminishes as the layer depth increases.**

5 Discussion

In this study, we investigated the alignment between odorant representations encoded by the MolFormer, a self-supervised transformer model pre-trained on chemical structures, and human olfactory perception. We evaluated the alignment between these representations by analyzing the similarity between them or finding a linear mapping between the representations. Additionally, we offered insights into the potential reasons behind the observed alignments by exploring relevant chemical features extracted by the model.

Perceptual prediction from pretrained models. We demonstrate for the first time that representations extracted from pre-trained large models, solely trained on chemical structures, align closely with the perceptual representations of odorants. This finding suggests that odorant perception can be accurately predicted from chemical structures. Furthermore, we show that this model can predict a subset of physicochemical descriptors known to be relevant to olfactory perception. Together, these results offer valuable predictions for chemists and neuroscientists to explore in future research.

Evaluating alignment across multiple datasets. To evaluate alignment from various perspectives, we designed three different experiments. First, we leveraged a dataset with expert-provided labels for odorants, assessing the model’s ability to independently predict multi-target binary labels for each odorant. This task did not involve variability from human participants or continuous odorant ratings. MoLFormer exhibited relatively high performance in predicting these binarized labels. Second, we used datasets containing average continuous ratings from human participants, which inherently present more challenges due to variability among non-expert participants’ ratings. Our evaluation revealed that while all models performed poorly on this task, MoLFormer performed comparably to supervised models. Lastly, we evaluated direct similarity scores between odorants from two datasets, examining the alignment between human-provided similarity scores and those computed from the representations encoded by models. MoLFormer showed a high alignment, highlighting its ability to predict similarity between odorants rather than relying on human-made descriptors. This suggests that pre-defined descriptors for describing odorants may need to be more carefully chosen, and models trained with these descriptors might not accurately reflect the true similarity between odorants.

Reduction in alignment with physicochemical descriptors across layers of the models. We conducted a complementary analysis to identify potential reasons underlying the observed perceptual alignment. Our focus was on the subset of features previously identified as significant for decoding olfactory perception from chemical structures. Our findings indicate that MoLFormer representations exhibit a high degree of alignment with these features. While most features show strong alignment, a few demonstrate less alignment (such as nRCOSR). These results collectively suggest that while these features are important, their significance varies. Additionally, our analysis of the predictability of these features across the different layers of the model shows that as we go through the layers, we observe a decrease in alignment with physicochemical descriptors despite an increase in alignment with perception. This observation aligns well with established principles in vision models, where lower layers have been shown to capture low-level, local features such as edges and textures, while progressively transitioning to align with higher-level, abstract representations, such as shapes and objects, in deeper layers [36]. However, further exploration is needed to fully uncover and understand this potential hierarchy.

Limitations. Our work is perhaps best understood in the context of its limitations. We do not directly take into consideration the intensity or concentration of each individual molecule within a mixture during the encoding of odorants. Incorporating these intensity factors in future work could potentially improve the alignment. Additionally, our research was constrained by the available datasets, which typically lack sufficient variations in different odorants, particularly for continuous rating regression tasks. Furthermore, we only considered the average rating scores and did not evaluate the alignment on a per-subject basis.

Future Work. We aim to leverage these findings to develop improved models of olfactory perception. Specifically, we plan to utilize unsupervised models trained exclusively on chemical structures to identify which chemical features are crucial for predicting perception, thereby avoiding the introduction of biases from human subjective perception. Additionally, we intend to investigate the mechanisms underlying olfactory perceptions decoded from chemical features. The observed alignment trends across different layers of the model may provide key insights into this process. Finally, evaluating representational alignment between the extracted representations from transformers trained on chemical structures and fMRI data from the brain can provide deeper insights into the underlying mechanisms of olfactory perception.

Acknowledgement

This work was supported by the Knut and Alice Wallenberg Foundation, Swedish Research Council, ERC AdV BIRD 88480, and ERC Syn D2Smell 101118977. The authors would like to thank Johan Lundström for helpful discussions.

References

- [1] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.
- [2] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.
- [3] Giorgio Ganis, William L Thompson, and Stephen M Kosslyn. Brain areas underlying visual mental imagery and visual perception: an fmri study. *Cognitive Brain Research*, 20(2):226–241, 2004.
- [4] Angela D Friederici. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268, 2012.
- [5] Subba Reddy Oota, Manish Gupta, Raju S Bapi, Gael Jobard, Frédéric Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *arXiv preprint arXiv:2307.10246*, 2023.
- [6] Jerry Tang, Meng Du, Vy A Vo, Vasudev Lal, and Alexander G Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv preprint arXiv:2305.12248*, 2023.
- [7] Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.
- [8] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.
- [9] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.
- [10] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.
- [11] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327, 2022.
- [12] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.
- [13] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.
- [14] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.
- [15] Mario Pannunzi and Thomas Nowotny. Odor stimuli: not just chemical identity. *Frontiers in physiology*, 10:1428, 2019.
- [16] Brian K Lee, Emily J Mayhew, Benjamin Sanchez-Lengeling, Jennifer N Wei, Wesley W Qian, Kelsie Little, Matthew Andres, Britney B Nguyen, Theresa Moloy, Jane K Parker, et al. A principal odor map unifies diverse tasks in human olfactory perception. *BioRxiv*, pages 2022–09, 2022.

- [17] Aharon Ravia, Kobi Snitz, Danielle Honigstein, Maya Finkel, Rotem Zirler, Ofer Perl, Lavi Secundo, Christophe Laudamiel, David Harel, and Noam Sobel. A measure of smell enables the creation of olfactory metamers. *Nature*, 588(7836):118–123, 2020.
- [18] Kobi Snitz, Adi Yablonka, Tali Weiss, Idan Frumin, Rehan M Khan, and Noam Sobel. Predicting odor perceptual similarity from odor structure. *PLoS computational biology*, 9(9):e1003184, 2013.
- [19] Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.
- [20] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [23] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [25] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.
- [26] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [27] Leffingwell and Associates. Database of perfumery materials and performance, 2001.
- [28] GoodScent. The good scents company.
- [29] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.
- [30] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.
- [31] Jason B Castro, Travis J Gould, Robert Pellegrino, Zhiwei Liang, Liyah A Coleman, Famesh Patel, Derek S Wallace, Tanushri Bhatnagar, Joel D Mainland, and Richard C Gerkin. Pyrfume: A window to the world’s olfactory data. *bioRxiv*, pages 2022–09, 2022.
- [32] Aryan Amit Barsainyan, Ritesh Kumar, Pinaki Saha, and Michael Schmuker. Openpom, 2023.
- [33] Vivek Sagar, Laura K Shanahan, Christina M Zelano, Jay A Gottfried, and Thorsten Kahnt. High-precision mapping reveals the structure of odor coding in the human brain. *Nature neuroscience*, pages 1–8, 2023.
- [34] Andreas Keller and Leslie B Vosshall. Olfactory perception of chemically diverse molecules. *BMC neuroscience*, 17:1–17, 2016.

- [35] Andrea Mauri. alvadesc: A tool to calculate and analyze molecular descriptors and fingerprints. *Ecotoxicological QSARs*, pages 801–820, 2020.
- [36] MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, volume 1311, 2014.

A Noise Ceiling

In order to evaluate the quality of data and the upper limit of the models, we computed the noise ceiling (Equation 1, 2) for the Sagar and Keller datasets as these are the only ones that have multiple evaluators for each odorant, and those are publicly available. The results show that we have the average noise ceilings of 0.28 ± 0.1 for the Keller dataset and 0.7 ± 0.05 for the Sagar dataset (Table S.1, S.2). The results show that the data of the Sagar dataset is less noisy, and there is still room for the models to increase the alignment. However, the Keller dataset alignment results are relatively close to the noise ceiling value.

$$r_j = \text{corr}(\text{Responses from Participant } j, \text{Mean Response across participants}) \quad (1)$$

$$\text{Noise Ceiling} = \frac{1}{N} \sum_{j=1}^N r_j \quad (2)$$

Table S.1: Noise ceiling per descriptor for Sagar dataset

| | | | | | | | | |
|----------------------|--------|--------------|------|---------|--------|--------|--------|-----------|
| Descriptor | Bakery | Burnt | Cool | Decayed | Fishy | Floral | Fruity | Intensity |
| Noise Ceiling | 0.68 | 0.70 | 0.68 | 0.72 | 0.75 | 0.73 | 0.79 | 0.75 |
| Descriptor | Musky | Pleasantness | Sour | Spicy | Sweaty | Sweet | Warm | |
| Noise Ceiling | 0.71 | 0.74 | 0.66 | 0.66 | 0.62 | 0.72 | 0.61 | |

Table S.2: Noise ceiling per descriptor for Keller dataset

| | | | | | | | |
|----------------------|-------------|--------------|--------|--------|----------|-------|-----------|
| Descriptor | Acid | Ammonia | Bakery | Burnt | Chemical | Cold | Decayed |
| Noise Ceiling | 0.21 | 0.21 | 0.32 | 0.27 | 0.27 | 0.17 | 0.29 |
| Descriptor | Familiarity | Fish | Flower | Fruit | Garlic | Grass | Intensity |
| Noise Ceiling | 0.33 | 0.21 | 0.26 | 0.37 | 0.31 | 0.25 | 0.53 |
| Descriptor | Musky | Pleasantness | Sour | Spices | Sweaty | Sweet | Warm |
| Noise Ceiling | 0.22 | 0.52 | 0.23 | 0.24 | 0.24 | 0.41 | 0.17 |

B Representational Similarity Matrix (RSM)

In order to better visualize how the models and humans represent different odors, we visualized representational similarity matrices for humans participants, Open-Pom, and MoLFormer across pairs of odorants for Ravia dataset in Figure S.1. The white cells show the pair of odorants for which no similarity score is available. MC-odorants corresponding to each mixture are provided in Table S.3

C t-SNE Visualizations

We also reduce representations of GS-LF dataset extracted from MoLFormer and Open-POM datasets using t-SNE (Figure S.2).

D Fine-tuned MolFormer

We fine-tuned MoLFormer using GS-LF dataset which is a large and inclusive dataset of odorants. Then we extracted representations for all the datasets and tasks. Figure S.3 shows ROC-AUC curve for GS-LF dataset. Figure S.4 demonstrates the results for the continuous rating prediction tasks for Keller and Sagar datasets, and Figure S.5 shows the results for RSA for Ravia and Snitz datasets.

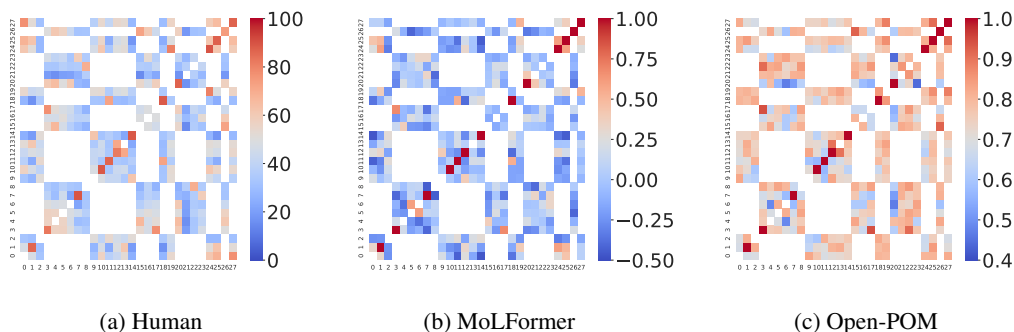


Figure S.1: RSM for different pairs of odorants for Ravi a dataset.

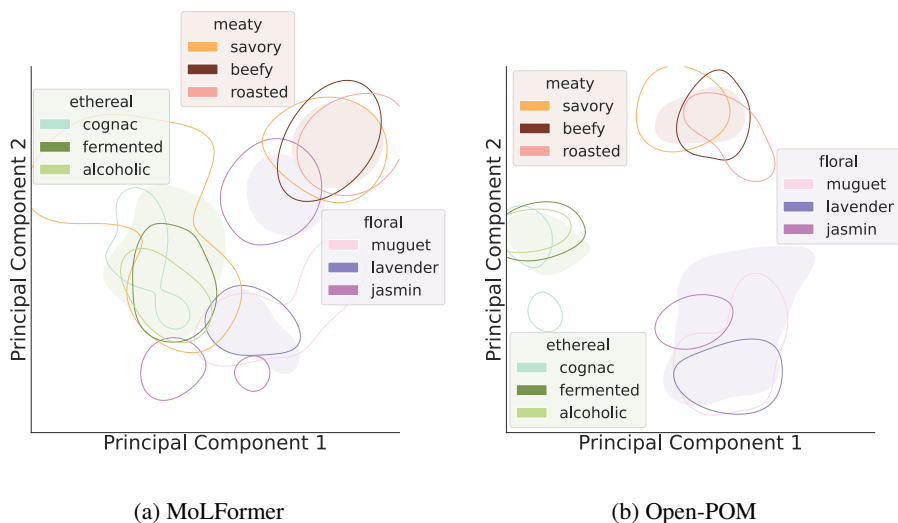


Figure S.2: **t-SNE Visualization of odorant representations encoded by different models** on the GS-LF dataset using the figure layout suggested by [16]. We reduced dimensionality using t-SNE. Areas dense with molecules that have broad category labels (floral, meaty, or ethereal) are shaded, while areas dense with narrow category labels are outlined. MoLFormer captures the perceptual relationship between different odorants in its representation space, despite not being explicitly trained for this purpose.

E Decoding chemical features

In this section, we present the results for predicting physicochemical descriptors from odorants for each dataset separately (Figure S.6-S.10). We observe that MoLFormer can better predict physicochemical descriptors in most cases.

F Decoding chemical features across layers of MoLFormer

We also show how the alignment between physicochemical descriptors changes across layers of the model for each dataset separately. As it is shown, alignments decrease across layers. This might be because the first layers extract more low-level information while deeper layers, extract more high-level features.

Table S.3: MC-odorants corresponding of indices in RSM presented in Figure S.1

| Index | MC-Odorant |
|-------|---|
| 0 | 126;520296;7122;6050;5273467;5364231 |
| 1 | 1550470;778574;11980;61771;6998;444972;14104;325;23642 |
| 2 | 2214;556940;8180;8077;325;11086 |
| 3 | 240;2758;8130;8129;7710;7059;4133;8918;957;6654 |
| 4 | 240;637511;7731;2758;12178;62336;8635 |
| 5 | 31276;62433;8129;12178;7519;18827;10722 |
| 6 | 31276;8148;7762;18827;7714 |
| 7 | 326;26331;1140;11002 |
| 8 | 5281168;637511;7685;12178;4133;7991;6054;7770;7714 |
| 9 | 5363233;10925;5365049;6050;5273467;31219;7765;23642 |
| 10 | 5363233;89440;126;11980;61293 |
| 11 | 556940;7601;11086;61670 |
| 12 | 565690;8180;5365049;6560;8077;31219;6998;7765;6997;18554 |
| 13 | 62351;1550470;7657;6997;6560;5273467;18554;2214 |
| 14 | 62351;565690;10925;7593 |
| 15 | 62433;8797;2758;3314;8635;61138;11002;6054;10722 |
| 16 | 6544;62433;7519;7685;3314 |
| 17 | 6544;93009;8130;8103;7710;7059;8918;7714 |
| 18 | 7194;520296;61670;637776;23642 |
| 19 | 7194;89440;6560;17121;126;637776;9012 |
| 20 | 7410;240;93009;8635 |
| 21 | 7410;326;2758;62444;7770;1140 |
| 22 | 7410;5281168;8797;7519;8129;7710;6654;8030 |
| 23 | 7519;8148;31252;8103;7710;11002 |
| 24 | 7601;778574;61331;8180;17121;24834;7593 |
| 25 | 7657;61331;61771;61293;24834;31219;444972;5367698;14104;5364231 |
| 26 | 8797;7731;7966;3314;62336;7059;7991;61138;6054;11002 |
| 27 | 89440;7657;7122;61293;7593;5367698;5364231;14104;9012 |

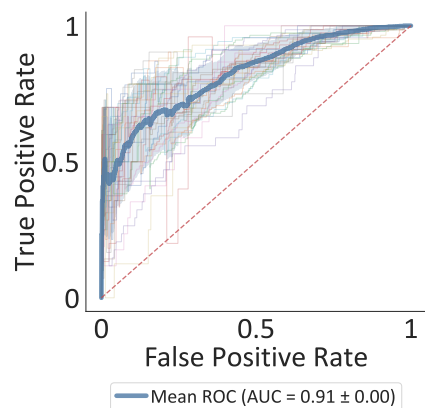


Figure S.3: **ROC curve for the linear classifier** trained on GS-LF representations extracted from the fine-tuned MoLFormer. Each curve corresponds to a separate test split, with the thicker curve representing the average performance across all splits.

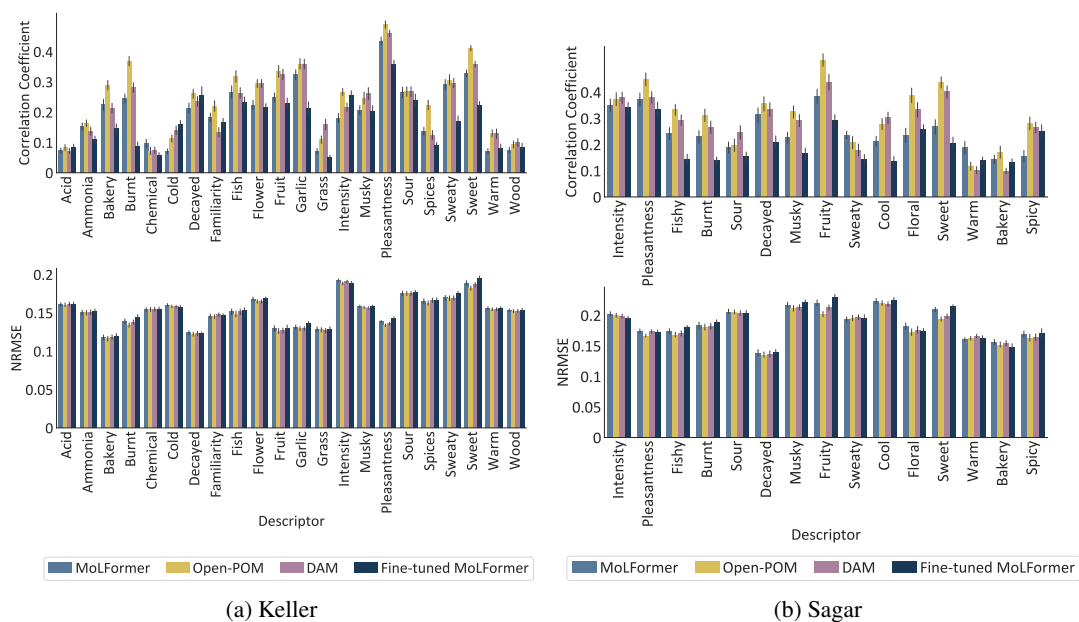


Figure S.4: **Performance of the models to predict continuous ratings per descriptor.** We computed Correlation and NRMSE between predicted and actual ratings per perceptual descriptor. Fine-tuned MoLFormer shows slightly worse performance in predicting continues ratings.

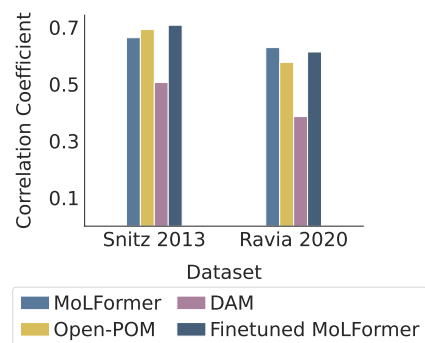


Figure S.5: **Representational similarity analysis** for Snitz and Ravia datasets. Fine-tuned MoLFormer performs on par with MoLFormer.

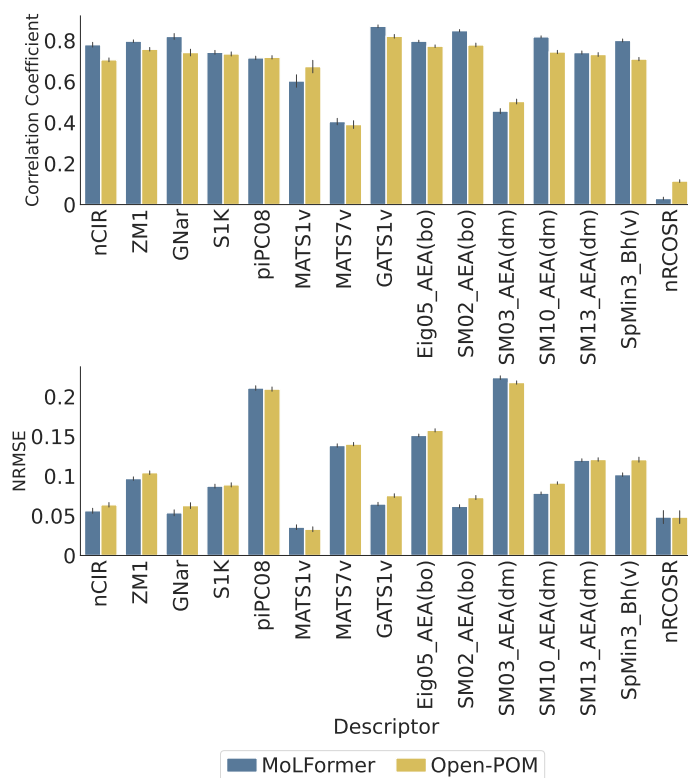


Figure S.6: **Performance of the models to predict relevant physicochemical descriptors** for Keller dataset. MolFormer performs slightly better than Open-POM in predicting descriptors.

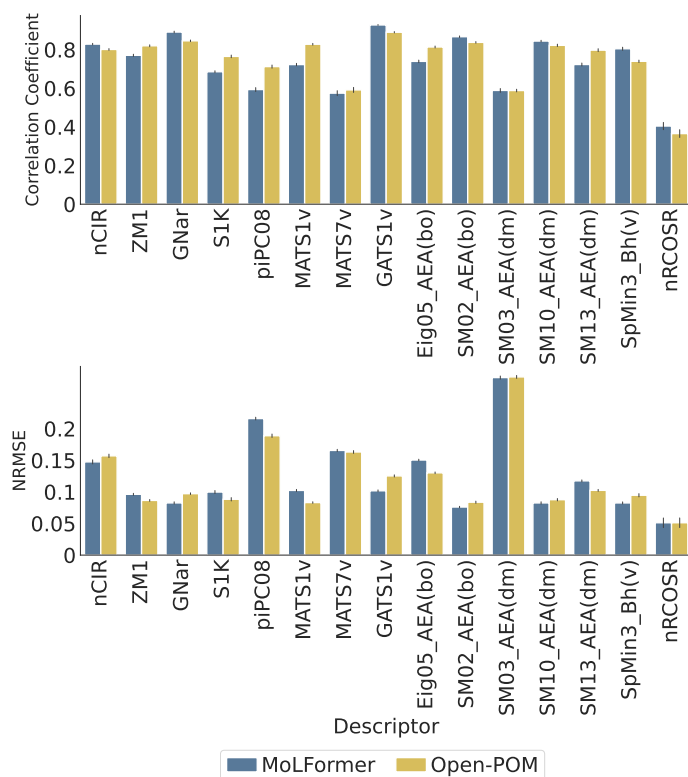


Figure S.7: **Performance of the models to predict relevant physicochemical descriptors** for Sagar dataset. MolFormer performs slightly better than Open-POM in predicting descriptors.

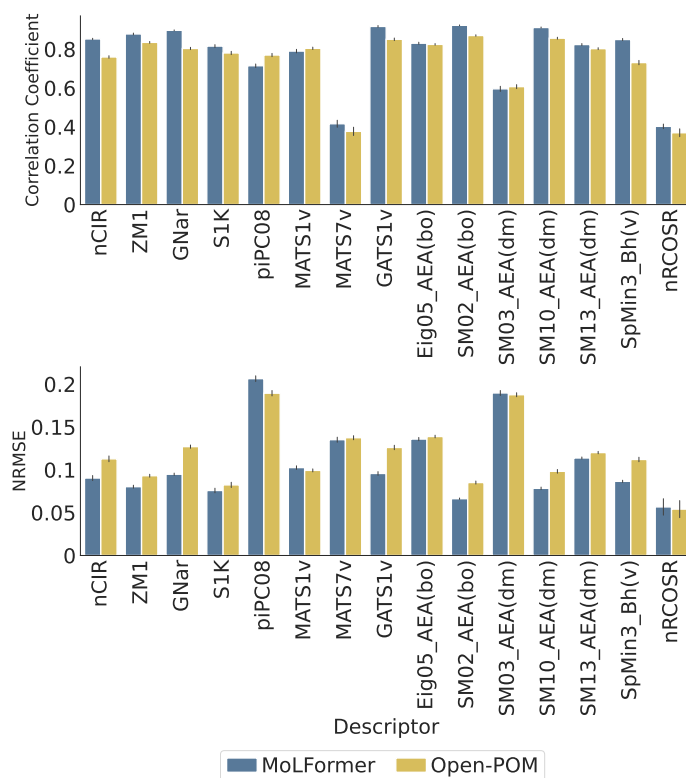


Figure S.8: **Performance of the models to predict relevant physicochemical descriptors** for Ravia dataset. MolFormer performs slightly better than Open-POM in predicting descriptors.

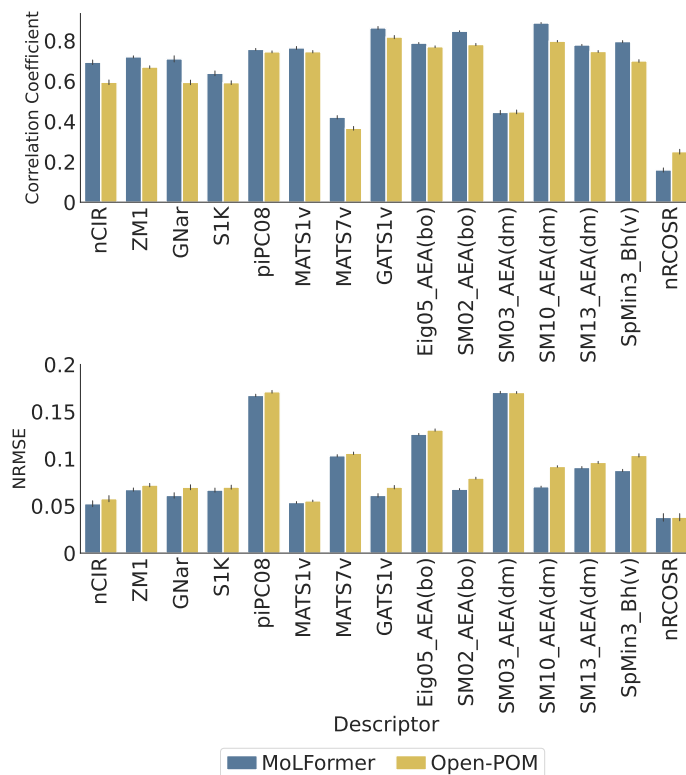


Figure S.9: **Performance of the models to predict relevant physicochemical descriptors** for Snitz dataset. MolFormer performs slightly better than Open-POM in predicting descriptors.

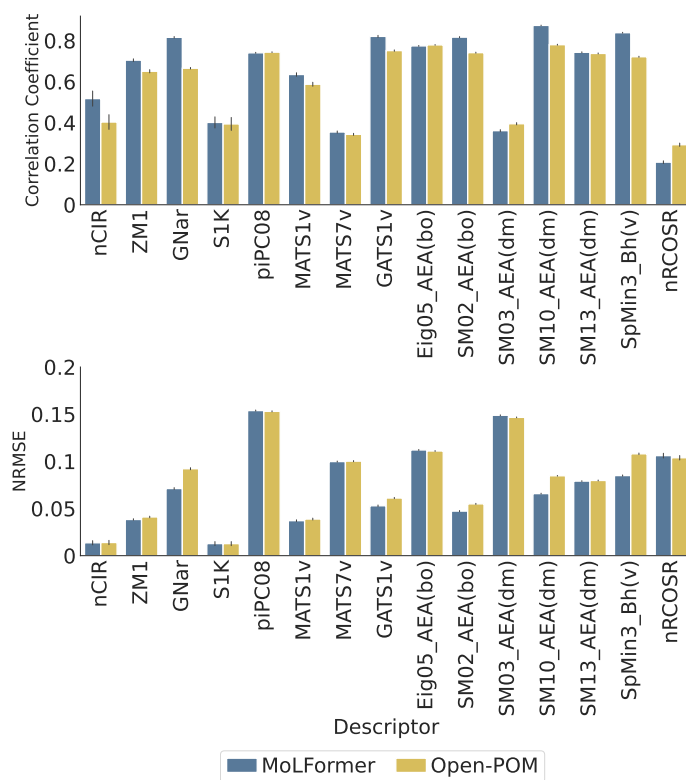


Figure S.10: **Performance of the models to predict relevant physicochemical descriptors** for GS-LF dataset. MolFormer performs slightly better than Open-POM in predicting descriptors.

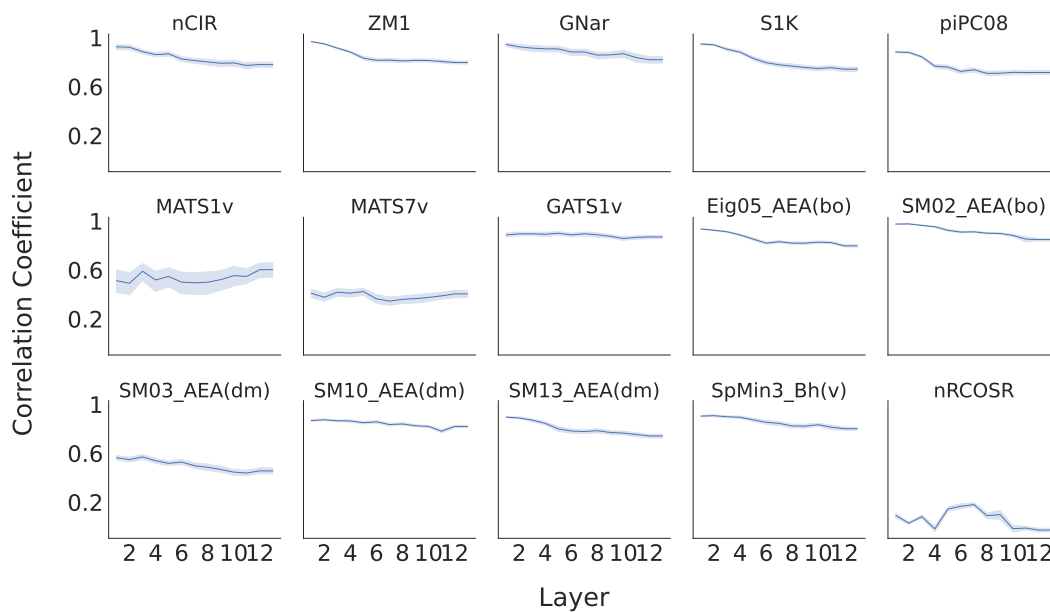


Figure S.11: **Correlation between the actual and predicted value** of physicochemical descriptors in Keller dataset diminishes as the layer depth increases .

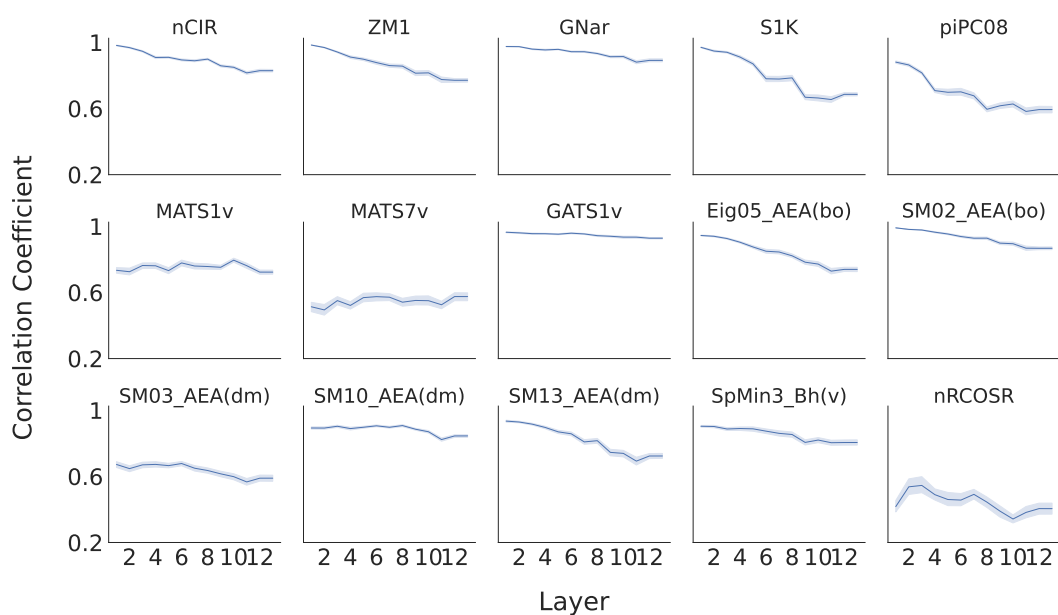


Figure S.12: **Correlation between the actual and predicted value** of physicochemical descriptors in Sagar dataset diminishes as the layer depth increases .

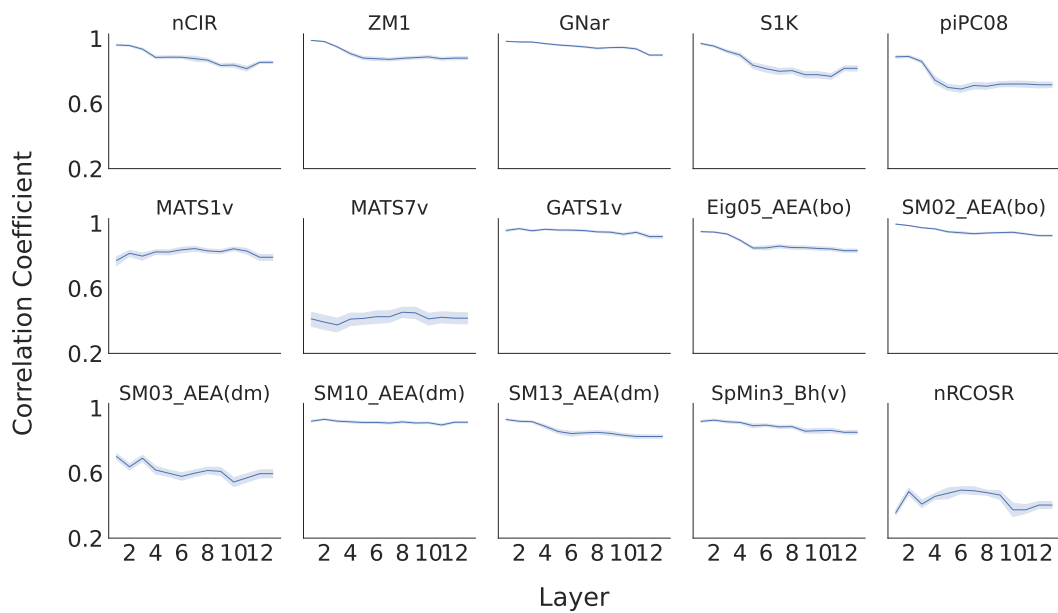


Figure S.13: **Correlation between the actual and predicted value** of physicochemical descriptors in Ravia dataset diminishes as the layer depth increases .

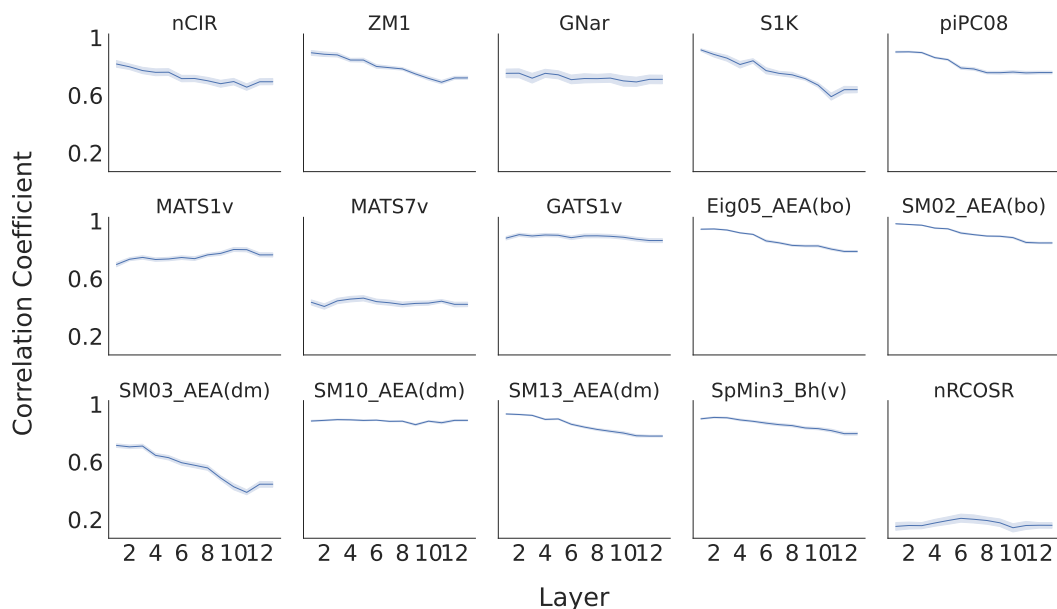


Figure S.14: **Correlation between the actual and predicted value** of physicochemical descriptors in Snitz dataset diminishes as the layer depth increases .

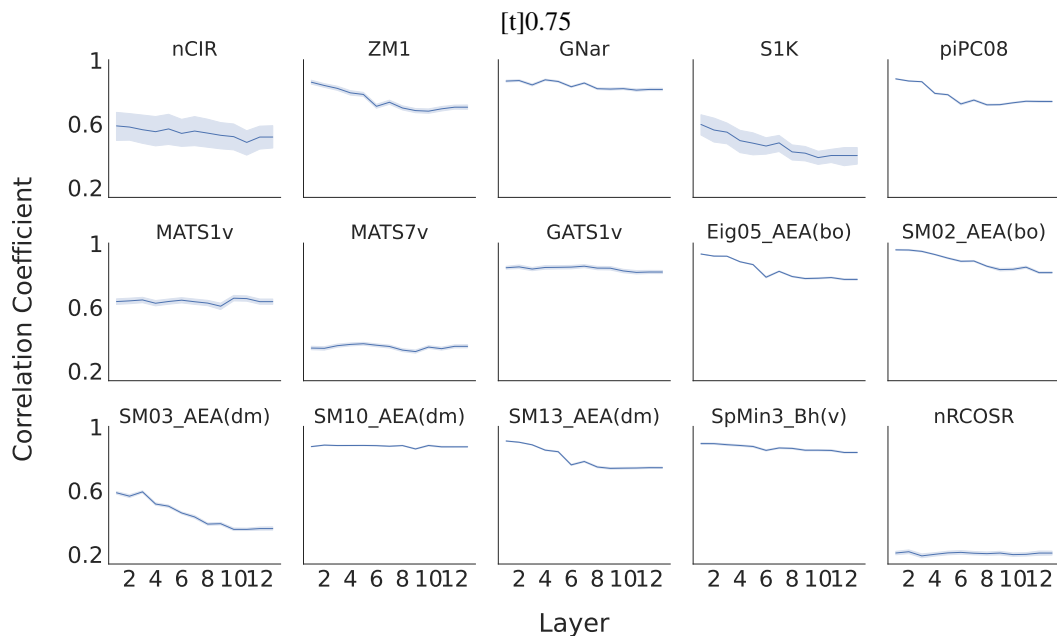


Figure S.15: **Correlation between the actual and predicted value** of physicochemical descriptors in GS-LF dataset diminishes as the layer depth increases .

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of the current work in the penultimate paragraph of the discussion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The parts of main experiment is uploaded and the whole code is available through the anonymized link provided in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars suitably and correctly defined.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We describe sufficient information on the computer resources needed to produce the experiments in the supplementary material section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics [https://neurips.cc/public/EthicsGuidelines?](https://neurips.cc/public/EthicsGuidelines)

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original owners of assets (e.g., code, data, models), used in the paper, are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: the paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing. Although we have used datasets that involve human subjects, we did not collect those data and used the datasets that are already approved and their details are publicly available in the original work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Although we have used datasets that involve human subjects, we did not collect those data and used the datasets that are already approved and publicly available.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.