SCENT: ALIGNING MASS SPECTRA WITH MOLECULAR STRUCTURE FOR OLFACTORY PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting human olfactory perception from molecular structure has seen remarkable progress in recent years. However, existing approaches typically require access to the explicit chemical structure of an unknown substance at inference time, which is not directly observed. We address this gap by exploring electron ionization mass spectrometry (EI-MS) as a direct input modality for olfactory prediction, leveraging chemically informative fragmentation patterns acquired within seconds. We introduce Spectrum-to-Chemical Embedding alignmeNT (SCENT), a multi-modal contrastive learning framework that aligns EI-MS representations with pretrained chemical structure embeddings while requiring only mass spectra at inference, focusing here on single-compound odorants. On the multi-label odor descriptor prediction task, SCENT outperforms MS-only baselines and approaches the performance of structure-based models without access to explicit molecular structure at test time. On a prospective perceptual panel of structurally or perceptually distinct odorants, SCENT maintains its advantage over MS-only baselines. Its learned representations also provide more informative features for predicting continuous human perceptual ratings and generalize to independently acquired laboratory spectra. Together, these results suggest that cross-modal alignment can effectively transfer structure-derived chemical information into EI-MS representations for olfactory prediction.

1 INTRODUCTION

Human olfaction remains comparatively underexplored in machine learning research, and its computational modeling poses substantial open challenges. A longstanding premise in olfactory neuroscience holds that the perceptual quality of an odor is strongly shaped by the physicochemical structure of the odorant molecule (Rossiter, 1996; Suh et al., 2025). This relationship underlies the quantitative structure-odor relationship (QSOR) paradigm, which has received increasing attention from the machine learning community in recent years (Keller et al., 2017; Lee et al., 2023; Taleb et al., 2024). Existing structure-based approaches, however, assume that molecular identity or an explicit molecular representation is available at inference time. In practical sensing settings, obtaining this information typically requires a separate identification procedure so that they cannot act on a physical measurement directly. This creates a gap between current structure-based olfactory models and rapid sensing scenarios.

To address this gap, we ask whether odor-related properties can instead be predicted from a rapidly acquired physical measurement. Electronic nose and gas sensor array systems are widely used for this purpose, but typically produce non-specific signals with limited chemical interpretability (Yan et al., 2015; Sung et al., 2024). We instead consider electron ionization mass spectrometry with quadrupole detection (EI-MS; hereafter MS), which captures chemically informative molecular fragmentation patterns (Ji et al., 2020) while allowing single-compound spectra to be acquired within seconds. Using EI-MS for olfactory prediction introduces two key challenges. First, mass spectra encode molecular fragmentation patterns rather than the explicit molecular representations exploited by modern structure-based olfactory models, and molecular identity is not assumed to be known at inference time. Second, perceptual supervision is scarce and highly imbalanced: existing datasets contain only a few thousand annotated odorants across 138 descriptors, whereas hundreds of thousands of paired spectrum-structure examples are available for representation learning. Existing MS-based odor models, however, don't explicitly exploit this abundant structural information

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

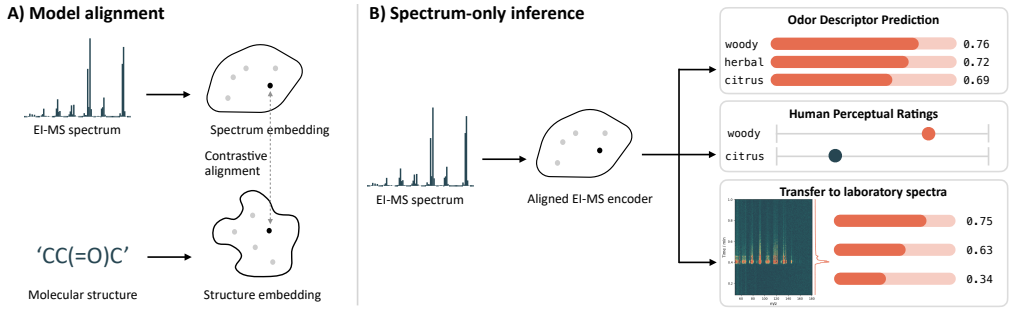


Figure 1: **Overview of SCENT**. During training, SCENT contrastively aligns EI-MS representations with pretrained molecular structure embeddings, using structure as privileged information. At inference, only EI-MS spectra are required. We evaluate the learned representations across three settings: multi-label odor descriptor prediction, continuous human perceptual rating regression, and transfer to independently acquired laboratory spectra.

during representation learning (Hasebe et al., 2022). The question this raises is whether structural information can be internalized by a spectral encoder during training, rather than recovered through an explicit identification step at inference. This motivates our central question: *can chemical structure serve as privileged supervision for learning MS representations that remain effective for odor prediction using only mass spectra at inference?*

To investigate this question, we introduce **SCENT** (**S**pectrum-to-**C**hemical **E**mbedding alig**N**men**T**)[1], a multimodal framework for chemically grounded MS representation learning (Fig. 1). SCENT uses contrastive learning to align MS representations with pretrained molecular structure embeddings over a large paired corpus. In doing so, it transfers structure-derived chemical information into the MS encoder while reserving scarce perceptual annotations for downstream learning. At inference time, SCENT operates on raw spectra alone. Importantly, SCENT is not designed to reconstruct molecular identity, nor does it assume that structural similarity directly determines perceptual similarity; rather, it tests whether cross-modal structural supervision can produce spectral representations that are more informative for olfactory prediction.

We evaluate SCENT through several complementary analyses. On the filtered GS-LF benchmark, SCENT consistently improves multi-label odor descriptor prediction over MS-only baselines and approaches the performance of structure-based representations, with an advantage retained on 178-odorant prospective panel (Section 4.1). On the DREAM dataset (Keller et al., 2017), frozen SCENT embeddings provide more informative features for predicting continuous human perceptual ratings than unaligned MS representations (Section 4.2). Finally, we evaluate transfer from library spectra to independently acquired laboratory measurements on 50 held-out single-compound odorants. SCENT retains its advantage over the MS-only baseline under this measurement-domain shift (Section 4.3). Together, these results show that cross-modal structure alignment improves EI-MS representations for both discrete and continuous olfactory prediction, with the benefit also preserved on independently acquired laboratory spectra, while requiring only mass spectra at inference.

2 Related Work

Computational olfaction.

The quantitative structure-odor relationship (QSOR) paradigm is a central framework in computational olfaction. The DREAM Olfaction Prediction Challenge (Keller et al., 2017), for example, characterizes odorants through hand-crafted cheminformatics features for odor prediction tasks. More recent approaches learn directly from molecular representations such as SMILES (Weininger, 1988) or molecular graphs using end-to-end models (Sanchez-Lengeling et al., 2019; Lee et al., 2023). Self-supervised molecular representations have also been shown to capture odor-relevant structure (Taleb et al., 2024), while other work incorporates complementary chemical information such as molecular electrostatics (Zhang et al., 2024). Despite these differences, such approaches assume that an explicit molecular representation is available at inference. On the sensing side, electronic noses and gas sensor arrays provide rapid measurements, and recent work has improved their

[1]The code and checkpoints are available at anonymous.4open.science/r/multimodal_emb_clean-7B26

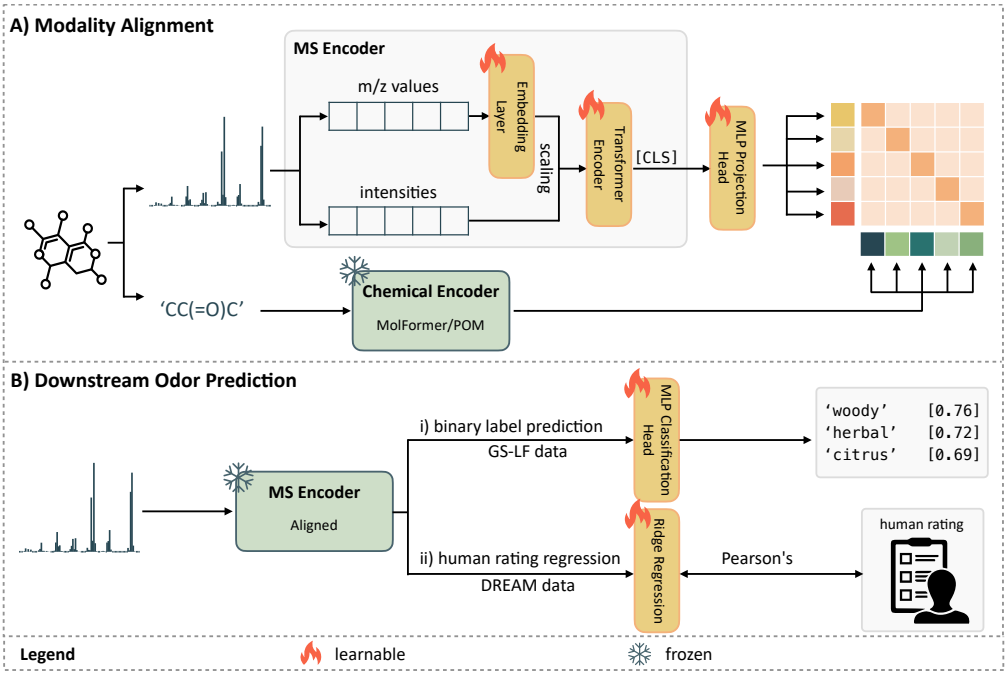


Figure 2: **The SCENT framework**: **(A)** During alignment, a learnable MS encoder is contrastively aligned with a frozen pretrained chemical encoder over paired spectrum–structure data. **(B)** The aligned MS encoder is frozen and evaluated using only mass spectra on two downstream tasks: (i) multi-label odor descriptor prediction on GS-LF with a trainable MLP head, and (ii) human perceptual rating regression on DREAM with ridge regression, evaluated using Pearson correlation. Flame and snowflake icons denote learnable and frozen components, respectively.

chemical interpretability by aligning sensor signals with GC–MS (Feng et al., 2025). Existing EI-MS approaches predict odor directly from spectra (Nozaki and Nakamoto, 2016; Ito and Nakamoto, 2020; Hasebe et al., 2022), but do not exploit molecular structure during training. SCENT bridges structure-based and sensing-based approaches by using molecular structure as privileged training information while requiring only mass spectra at inference.

Mass spectra representation and contrastive learning.

Prior work has shown that EI-MS fragmentation patterns contain learnable chemical information. Wei et al. (2019) demonstrated that neural networks can predict EI-MS spectra from molecular structure, while Liu et al. (2024) introduced EIMS2Vec, which learns spectral representations from peak co-occurrence statistics over large EI-MS libraries. Although EIMS2Vec captures recurring fragmentation patterns, its representations are learned from spectral statistics alone and are not explicitly aligned with molecular representations. Cross-modal alignment between mass spectra and molecular structure has primarily been studied for tandem MS, with molecular identification or retrieval as the downstream objective (Goldman et al., 2023; Chen et al., 2024; Kalia et al., 2025; Zhang et al., 2026). SCENT instead focuses on direct EI-MS and uses structure alignment for olfactory representation learning, with molecular structure available only during training and mass spectra alone used at inference. Our contribution is therefore not a new contrastive objective, but *the formulation and evaluation of MS-structure alignment as a representation-learning strategy for direct MS-based olfactory prediction.*

3 Method

SCENT is a multimodal contrastive learning framework that uses molecular structure as privileged information to learn chemically informed EI-MS representations. As shown in Figure 2, the method comprises two stages: (1) contrastive alignment of EI-MS and molecular structure representations, and (2) downstream olfactory prediction using the aligned MS representation. Molecular structure is used only during the alignment stage; downstream models and inference operate on mass spectra alone. Implementation details and hyperparameters are provided in Appendix C.

3.1 MODALITY ALIGNMENT

Mass spectrum encoder. An EI-MS represents a molecule as a set of peaks defined by mass-to-charge (m/z) values and their corresponding intensities (Appendix D for details). Inspired by EIMS2Vec (Liu et al., 2024), we embed peak co-occurrences and weight the resulting embeddings by power-scaled intensities ($p = 0.5$). Unlike EIMS2Vec, which uses sum pooling, we process the weighted peak embeddings with a two-layer Transformer encoder with four self-attention heads, which we empirically found to improve performance. We extract the final MS representation by applying a projection MLP to the `[CLS]` token (Devlin et al., 2019).

Chemical structure encoder. We use two frozen structure encoders as alignment targets: Open-POM (Barsainyan et al., 2023; Lee et al., 2023), an odor-supervised graph model, and MolFormer (Ross et al., 2022), a self-supervised molecular language model. Open-POM is retrained after excluding the fixed GS-LF test molecules, whereas MolFormer uses the publicly available pretrained checkpoint. We refer to the resulting models as SCENT (Open-POM) and SCENT (MolFormer).

Training data. Starting from the NIST 2023 EI-MS library, we retain compounds with molecular weights between 50 and 300 Da and retrieve corresponding SMILES from PubChem. After canonicalization and removal of entries without valid structures, 148,212 spectrum-structure pairs remain for alignment. GS-LF held-out test molecules are excluded from both Open-POM retraining and alignment data, leaving 147,913 pairs. Further preprocessing details are provided in Appendix A.

Training objective. SCENT aligns MS and structure representations using a symmetric contrastive objective inspired by CLIP (Radford et al., 2021). For a mini-batch of N paired spectra and structures, we ℓ_2-normalize their projected embeddings u_i and v_i and compute $s_{ij} = \frac{u_i^\top v_j}{\tau}$, where τ is the temperature. We minimize the average of the spectrum-to-structure and structure-to-spectrum cross-entropy losses, treating matched spectrum-structure pairs as positives and all other pairs in the batch as negatives. The temperature is selected on the alignment validation set via grid search, and validation loss is used for selecting checkpoint.

3.2 DOWNSTREAM ODOR PREDICTION

We evaluate the aligned MS representations on two complementary tasks: multi-label odor descriptor prediction and continuous human perceptual rating regression.

3.2.1 MULTI-LABEL ODOR DESCRIPTOR PREDICTION

Dataset and protocol. Following Lee et al. (2023), we evaluate multi-label odor descriptor prediction on the GS-LF dataset (GoodScent, 2026; Leffingwell and Associates, 2001), which contains 138 odor descriptors. After retaining single-compound odorants with valid NIST EI-MS spectra and deduplicating molecular identities, 2,588 samples remain. We construct a fixed 10% test set using iterative stratified splitting to preserve label co-occurrence statistics. On the remaining 90%, we perform 5-fold iterative-stratified cross-validation for model selection and training. The label distribution for each fold is shown in Appendix H. All test molecules are excluded from upstream Open-POM retraining and SCENT alignment. We train a two-layer MLP on the frozen `[CLS]` embeddings using weighted binary cross-entropy, with positive-class weights accounting for label imbalance.

Prospective perceptual panel. We evaluate model prediction beyond the GS-LF odor-training set using the prospective perceptual panel of Lee et al. (2023), which contains 400 novel odorants absent from GS-LF, rated by a trained sensory panel using a 55-descriptor subset of the 138 GS-LF labels. Chemical structures are publicly available for 209 molecules, of which 208 have corresponding panel ratings. Ratings are averaged across panelists and replicates to obtain a 55-dimensional panel-mean perceptual profile for each molecule. Of these, 178 have EI-MS spectra and are used for evaluation.

Metrics. We report the following evaluation metrics:

- **ROC-AUC Score**: Macro-AUC is the unweighted mean of per-descriptor AUCs, and Micro-AUC is used for early stopping. Classes without positive test samples are excluded.

Table 1: **Model performance on the filtered GS-LF dataset**. Results are reported as mean ± std over 5 folds training set on same test set. ↑ denotes improvement over EIMS2Vec. Bold indicates the best result per metric.

Model	Input	Micro-AUC↑	Macro-AUC↑	Adj.P@k↑
Open-POM	SMILES	89.94 ± 0.05	$\mathbf{86.70 \pm 0.15}$	$\mathbf{52.22 \pm 0.28}$
MolFormer	SMILES	89.29 ± 0.20	84.14 ± 0.44	50.23 ± 0.16
EIMS2Vec (Liu et al., 2024)	MS	87.98 ± 0.15	82.29 ± 0.61	46.49 ± 0.53
AE (Debnath and Nakamoto, 2022)	MS	86.31 ± 0.27	76.83 ± 0.49	40.97 ± 0.99
SCENT (Open-POM)	MS	89.80 ± 0.09 ↑1.82	85.45 ± 0.30 ↑3.16	50.60 ± 0.53 ↑4.11
SCENT (MolFormer)	MS	$\mathbf{89.99 \pm 0.18}$ ↑2.01	85.71 ± 0.26 ↑3.42	50.62 ± 0.75 ↑4.13

- **Precision@k**: We adopt an adjusted Precision@k metric to evaluate the proportion of true positive labels among the top-k predictions produced by the model. Since some samples may contain fewer than k positive labels, we define it as $\frac{\text{hit@k}}{\min(k, n_i^+)}$, where n_i^+ denotes the number of ground-truth positive labels for the i-th sample, and $hit@k$ is the number of correctly predicted labels in the top-k predictions. In this work we use $k = 5$ since the average positive label number of dataset is around 5.

3.2.2 Human perceptual rating regression

Dataset and protocol. We use the DREAM olfaction dataset (Keller et al., 2017), containing ratings from 49 non-expert participants across 21 perceptual attributes for 476 molecules. After retaining molecules with valid library EI-MS spectra, 385 remain. Because normalized relative-intensity EI-MS does not encode absolute concentration, we retain the highest available concentration for each molecule, average ratings across participants, and normalize each score to $[0, 1]$. We perform 100 repeated 80/20 train–test splits and fit ridge regression models on frozen MS embeddings to predict the 21 perceptual attributes.

Metric. Following prior work (Taleb et al., 2024), we use the Pearson correlation coefficient (r) between predicted and human-rated perceptual scores to evaluate how well the aligned MS representations approximate human perceptual judgments.

4 Results

We report results for odor descriptor prediction, including ablation and data-efficiency analyses (Section 4.1), human perceptual rating regression (Section 4.2), and evaluation on independently acquired laboratory spectra (Section 4.3). Additional results are provided in Appendices E, F, and G.

4.1 Odor descriptor prediction

Main results. Table 1 reports performance on the filtered GS-LF benchmark. Both SCENT variants consistently outperform the unaligned MS baselines and approach the performance of the corresponding structure-based models, despite using only MS at inference. These results indicate that cross-modal alignment transfers useful structure-derived information to the MS representation and substantially narrows the gap between MS- and structure-based odor prediction.

We also compare SCENT with a two-stage pipeline that first retrieves molecular identity from MS and then applies Open-POM to the retrieved structure; SCENT consistently outperforms this retrieval-based approach (Appendix E.2). SCENT improves over EIMS2Vec across descriptor-frequency groups on training set (Figure 3; Appendix E.3) as well as different levels of chemical similarity with training set (Appendix E.4). We further analyze the learned representation space using PCA (Appendix E.5).

Finally, we vary the amount of downstream training data and evaluate all models on the same held-out test set across five folds (Appendix E.6). SCENT maintains a consistent advantage over the MS-only baseline as the training set is reduced, including in low-data regimes.

Ablation study. To assess the contributions of cross-modal alignment and the Transformer encoder, we compare SCENT with several ablated variants (Table 2). Replacing structure alignment with

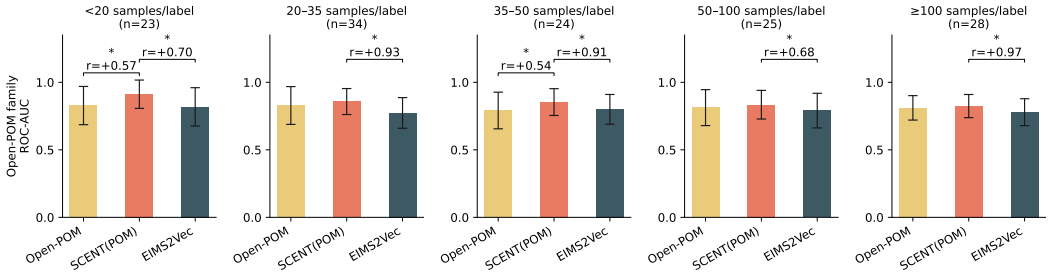


Figure 3: **Mean per-label ROC-AUC grouped by label frequency for Open-POM based models.** Each bar reports the mean AUC across labels within a given frequency bin, with a standard error of the mean error bar. Statistical significance is assessed via the Wilcoxon signed-rank test (* $p < .05$), and the r is the effect size. More results in Appendix E.3.

Table 2: **Ablation study of SCENT** (mean $\pm$ std over 5 folds). We ablate two components: the Transformer head and the training objective. *w/o CLIP 1* replaces the contrastive objective with masked reconstruction (MR); *w/o CLIP 2* replaces it with end-to-end supervised training (S); *w/o Transformer* retains contrastive alignment (C) but removes the Transformer head; *MSE* use MSE loss. EIMS2Vec uses in-spectrum peak co-occurrence (PC) and serves as the baseline.

Model Variant	Transformer	Objective	Micro-AUC↑	Macro-AUC↑	Adj.Precision@k↑
w/o CLIP 1	✓	MR	76.34 ± 0.03	55.39 ± 0.45	29.95 ± 0.20
w/o CLIP 2	✓	S	86.16 ± 0.54	78.98 ± 0.62	41.48 ± 0.86
w/o Transformer (Open-POM)	×	C	88.39 ± 0.12	83.04 ± 0.60	47.40 ± 0.59
w/o Transformer (MolFormer)	×	C	88.40 ± 0.08	83.15 ± 0.34	47.42 ± 0.59
Regression (Open-POM)	✓	MSE	88.25 ± 0.10	83.24 ± 0.25	46.28 ± 0.73
Regression (MolFormer)	✓	MSE	87.69 ± 0.08	82.06 ± 0.16	44.94 ± 0.38
EIMS2Vec	×	PC	87.98 ± 0.15	82.29 ± 0.61	46.49 ± 0.53

spectrum-only masked reconstruction (MR) substantially reduces performance, despite retaining the Transformer encoder. Directly training the same encoder on odor labels (S) also performs substantially worse than SCENT. The MSE regression results further support our choice of the contrastive objective. Together, these results indicate that the gains of SCENT do not arise from increased encoder capacity or spectrum-only representation learning, but depend strongly on supervision from molecular structure during alignment.

Removing the Transformer while retaining contrastive alignment leads to a smaller but consistent decrease across all metrics. Nevertheless, both aligned variants without the Transformer still outperform EIMS2Vec, showing that cross-modal structure alignment provides benefits independently of the Transformer architecture. The Transformer provides an additional improvement, suggesting that explicitly modeling interactions among spectral peaks complements the information transferred through structure alignment.

Prospective panel evaluation. We further evaluate whether spectrum–structure alignment remains beneficial on the prospective perceptual panel introduced by Lee et al. (2023). The panel contains odorants that are structurally or perceptually distinct from the training set and includes structure-odor discordant cases where structurally similar molecules can nevertheless evoke substantially different percepts. We use the 178 panel molecules for which matched EI-MS spectra are available and evaluate each molecule by the Pearson correlation between its predicted odor profile and the human panel mean value.

Averaging repeated correlations within each molecule, SCENT (Open-POM) outperforms EIMS2Vec across 178 molecules (mean r: 0.185 vs. 0.147, mean paired improvement: 0.039, 95% paired-bootstrap CI [0.014, 0.066], two-sided Wilcoxon signed-rank significant($p < 0.01$, Figure 4A)). Prediction difficulty is shared across spectral and structure-based models (Figure 4C). Molecules for which Open-POM, EIMS2Vec, and SCENT (Open-POM) all achieve $r < 0.10$ ($n = 47$) or all achieve $r > 0.30$ ($n = 28$) show no significant difference in nearest-neighbor Tanimoto similarity to GS-LF, but differ markedly in neighbor odor similarity, associating prediction difficulty with structure-odor discontinuities (Lee et al., 2023).

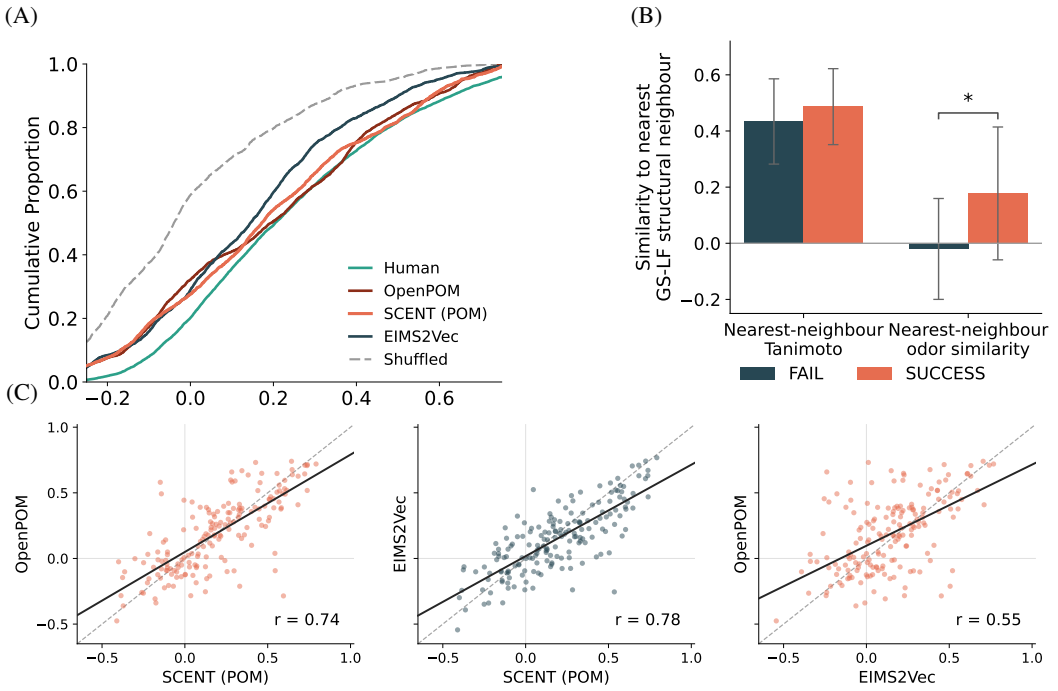


Figure 4: **Prospective panel evaluation.** **(A)** Cumulative density of plot showing the distribution of model prediction with the panel mean value (right side is better). **(B)** Molecules that are difficult for models (FAIL, $r < 0.10$) don't have significant structural difference with GS-LF data compared with well-predicted molecules (SUCCESS, $r > 0.30$), but the nearest chemical neighbor carries a far less similar odor profile using Wilcoxon rank-sum test. **(C)** Pairwise agreement in per-molecule prediction performance across model.

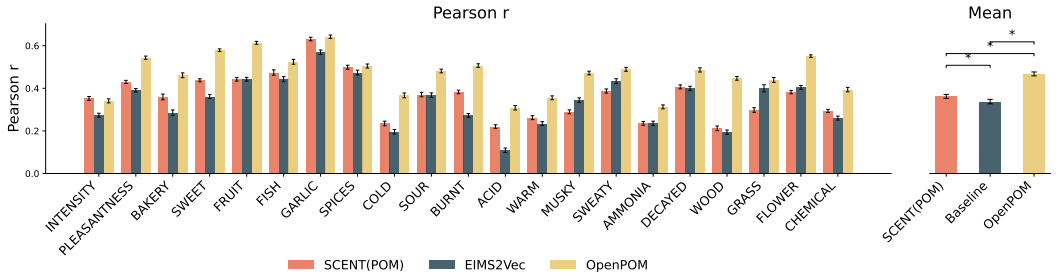


Figure 5: **Pearson correlation between predicted and human-rated perceptual scores on the DREAM dataset (Open-POM family).** Ridge regression (α=1 by grid search) is fitted on frozen embeddings to predict 21 continuous perceptual descriptors (mean $\pm$ SEM across 100 repeated train–test splits). Statistical significance is assessed via the two-sided Wilcoxon signed-rank test on 21 paired attribute (Wilcoxon, 1992) (* $p < .05$).

4.2 Human perceptual rating regression

Figure 5 and Figure F1 report the per-descriptor Pearson correlation between regression predictions and human ratings. Both SCENT variants outperform the unaligned EIMS2Vec baseline, indicating that spectrum-structure alignment improves the ability of MS-derived representations to capture human perceptual ratings, transferring perceptual grounding that is otherwise a key property of structure-based representation.

SCENT (Open-POM) remains below Open-POM, which is expected: Open-POM embeddings are directly optimized with odor supervision and therefore contain task-specific information that is not fully recovered by the present MS model. SCENT (MolFormer) is statistically indistinguishable from MolFormer, despite using only MS spectra at inference time. Since MolFormer embeddings are not trained with explicit olfactory supervision, this result indicates that structure-derived supervision improves the spectral representation to predict human perception, and that SCENT can effectively

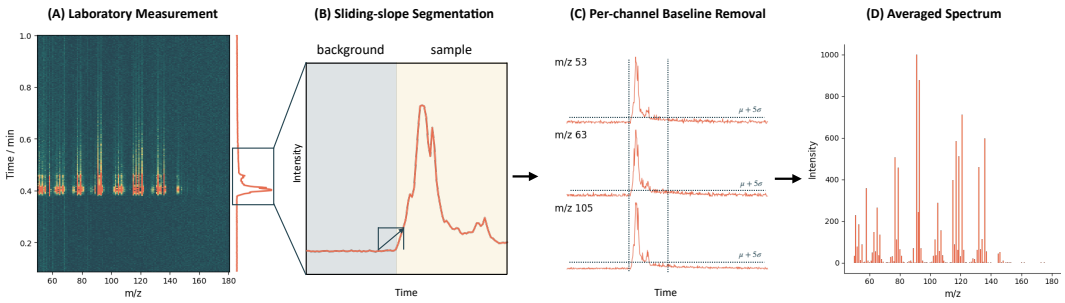


Figure 6: **Signal pre-processing pipeline for lab-collected spectra.** **(A)** Raw ion intensity is collected as a 2D matrix over m/z channels and time. **(B)** A sliding-slope algorithm segments the total ion current into background and sample regions. **(C)** We empirically selected a conservative threshold of $\mu_{\text{bg}} + 5\sigma_{\text{bg}}$ for per-channel background suppression. **(D)** Corrected intensities within the sample window are averaged over time to produce a single spectrum as model input.

transfer this information into the MS embedding space. Statistical comparisons use the 21 paired attribute-level correlations after averaging across splits.

To assess whether the observed trend depends on concentration handling, we perform two additional analyses in Appendix F.2 and Appendix F.3. First, we evaluate the frozen representations separately at three reported dilution levels in DREAM. Because different dilution levels contain different molecules, the absolute performance across concentrations is not directly comparable. Within each dilution subset, SCENT remains at or above the MS-only baseline, although the magnitude of the improvement varies across subsets, likely reflecting differences in chemical composition and perceptual signal quality. Second, we examine whether reported dilution provides information beyond the frozen spectral representation. Explicitly providing dilution improves condition-specific perceptual prediction, particularly when dilution interacts with the molecular representation.

4.3 Evaluation on independently acquired spectra

Applying SCENT to independently acquired laboratory spectra introduces a measurement-domain shift relative to library references, due to differences in peak intensities, background noise, and acquisition conditions. We therefore build a signal pre-processing pipeline and retrain the model on the common m/z range to evaluate whether the benefit of structure alignment persists under this shift.

Model retrain. To mitigate the influence of air-related ions in the low m/z region (small molecules) and low-volatility fragments in the high m/z region (heavy molecules), real measurements were restricted to the m/z range of [50, 180]. We retrain the model using the same m/z range consistently across NIST library data, ensuring alignment between training and inference conditions. No laboratory-measured spectrum is used for parameter adaptation. The retrained model is then evaluated on the same downstream tasks and datasets described in Section 3.2, with results reported in Appendix (Table G4).

Laboratory evaluation. To our knowledge, no public dataset pairs real-world direct EI-MS measurements with odor annotations. We therefore collected laboratory spectra for 50 single-compound samples, including 46 from the filtered GS-LF test set and 4 (not in the downstream GS-LF dataset) with labels sourced from GoodScent (2026). The full sample list is provided in Appendix G.2. The acquisition and preprocessing pipeline is shown in Figure 6 and detailed in Appendix G.1. The set is small by machine-learning standards because each entry is a physical acquisition rather than a database record. To quantify uncertainty arising from the small real-world validation set, we performed non-parametric bootstrap resampling over these 50 molecules. For each bootstrap replicate, we sampled 50 molecules with replacement, recomputed metrics for each of the five classification heads, and averaged the metric across heads. We repeated this procedure 10,000 times and reported the $2.5th$ and $97.5th$ percentiles as the 95% bootstrap confidence interval. Per-model results and 95% percentile confidence intervals are reported in Table 3 and Table G2. Relative to EIMS2Vec, SCENT (Open-POM) and SCENT (MolFormer) improve the observed Micro-AUC by 3.94 and 3.53 percentage points, respectively. Paired bootstrap resampling gives 95% confidence intervals

Table 3: **Performance comparison between EIMS2Vec and SCENT (Open-POM) on independently acquired spectra.** Results are reported as mean ± standard deviation across the five classification heads, together with 95% confidence intervals obtained by bootstrap resampling over the 50 measured molecules. Higher values indicate better performance.

Metric	Statistic	EIMS2Vec	SCENT (Open-POM)
Micro-AUC↑	mean ± std	84.69 ± 0.65	88.63 ± 0.29
	95% CI	$[82.85, 86.46]$	$[87.24, 90.06]$
Macro-AUC↑	mean ± std	78.47 ± 1.19	82.09 ± 0.61
	95% CI	$[74.43, 81.46]$	$[78.96, 84.48]$
Adj.P@5↑	mean ± std	42.32 ± 1.88	57.57 ± 1.14
	95% CI	$[36.74, 47.72]$	$[51.66, 63.62]$

of $[2.73, 5.21]$ and $[2.34, 4.80]$ for these differences, and the paired intervals for Macro-AUC and Adj.P@5 also exclude zero (Table G3). Overall, the benefit of structure alignment is preserved under independently acquired laboratory measurements.

5 Discussion

We introduced SCENT, a cross-modal contrastive learning framework that uses molecular structure as privileged training information during training while requiring only mass spectra at inference. Across odor descriptor prediction and continuous perceptual regression, aligning direct EI-MS representations with pretrained chemical structure embeddings consistently improves over MS-only representations, indicating that structure-derived chemical priors make the spectral representation more useful for olfactory prediction. We additionally evaluate transfer from library spectra to independently acquired laboratory measurements. SCENT retains its advantage over the MS-only baseline on the held-out laboratory spectra, showing that the benefit of structure alignment is preserved under the measurement-domain shift considered here. More broadly, our results show that chemical structure can serve as privileged supervision for learning perception-oriented representations of direct EI-MS. Our prospective panel analysis further associates shared prediction difficulty with lower odor-annotation similarity to nearby structures, consistent with structure–odor discontinuity as one challenge for both spectral and structure-based prediction.

Limitations. First, both alignment training and downstream evaluation rely on library-acquired spectra, which differ from real-world measurements in peak intensity distributions and background contamination. While our case study demonstrates encouraging transfer to lab-measured signals, the domain gap between library and real-world spectra remains a key challenge for application, and the small size of our real-world test set (50 samples) limits the strength of conclusions that can be drawn. Second, the current alignment framework is developed and evaluated on single-compound inputs. Extending SCENT to mixtures requires matched mixture-level spectral and perceptual data. Simply constructing synthetic mixed spectra does not provide a reliable odor target because mixture perception is not determined by a known linear combination of constituent percepts. Third, SCENT does not explicitly condition on concentration, and absolute concentration is not directly recoverable from the normalized relative-intensity EI-MS representation used here. Some studies (Pellegrino et al., 2025) already suggested that concentration-perception relationships can be strongly molecule specific, making them difficult to capture with a global rule shared across odorants. Explicitly providing the reported dilution yields an improvement, particularly through an interaction with the molecular representation. However, the present data contain only two conditions per molecule. Learning a generalizable molecule-specific concentration effects therefore remains an open challenge in computational olfaction.

Future work. Future work will build on the representation-learning framework introduced here by examining settings that are not fully captured by the current single-compound formulation. In particular, it will be important to understand how the learned spectral representations behave when perceptual outcomes depend more strongly on presentation conditions, and how the same framework can be adapted to more complex spectra such as mixtures. These directions would help clarify the scope and limits of structure-guided MS representation learning for olfactory prediction.

AI USE STATEMENT.

In this work, we used generative AI tools to clean and reformat datasets. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims. We used generative AI tools to summarize and analyze existing literature, brainstorm research ideas, and edit the manuscript to improve readability. All AI-assisted outputs were reviewed and verified by the authors against the original data, relevant literature, and experimental results where applicable. We take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT.

We provide the source code, model implementations, training and evaluation procedures, hyperparameters in the anonymous repository linked (See footnote in p. 2) and the appendix in the paper. Due to licensing restrictions, we are not permitted to redistribute EI-MS spectra from the NIST EI-MS Library. All public benchmark datasets used in this work are available from their original sources. The independently acquired laboratory EI-MS dataset is currently being prepared for public release. The molecular identities and perceptual annotations are provided in the Appendix, and the raw EI-MS measurements will be released in a future public version of the dataset.

REFERENCES

A. A. Barsainyan, R. Kumar, P. Saha, and M. Schmuker. OpenPOM - open principal odor map, 2023.

L. Chen, B. Xia, Y. Wang, X. Huang, Y. Gu, W. Wu, and Y. Zhou. Cmssp: A contrastive mass spectra-structure pretraining model for metabolite identification. *Analytical Chemistry*, 96(42): 16871–16881, 2024.

T. Debnath and T. Nakamoto. Extraction of sensing data for desired scent impressions using mass spectra of odorant molecules. *Scientific Reports*, 12(1):16297, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American chapter of the association for computational linguistics: human language technologies*, pages 4171–4186, 2019.

D. Feng, C. Li, W. Dai, and P. P. Liang. Smellnet: A large-scale dataset for real-world smell recognition. *arXiv:2506.00239*, 2025.

S. Goldman, J. Xin, J. Provenzano, and C. W. Coley. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of chemical information and modeling*, 64(7):2421–2431, 2023.

GoodScent. The good scents company. http://www.thegoodscentscompany.com/, 2026. Accessed: May 2026.

D. Hasebe, M. Alexandre, and T. Nakamoto. Exploration of sensing data to realize intended odor impression using mass spectrum of odor mixture. *PLOS one*, 17(8):e0273011, 2022.

K. Ito and T. Nakamoto. Improvement of odor impression predictive model using machine learning. In *2020 IEEE Sensors*, pages 1–4. IEEE, 2020.

H. Ji, H. Deng, H. Lu, and Z. Zhang. Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Analytical chemistry*, 92(13):8649–8653, 2020.

A. Kalia, Y. Zhou Chen, D. Krishnan, and S. Hassoun. Jestr: Joint embedding space technique for ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics*, 41(7):btaf354, 2025.

A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, leander schietgat, K. D. Grave, R. Norel, D. O. P. Consortium, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, and P. Meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.

B. K. Lee, E. J. Mayhew, B. Sanchez-Lengeling, J. N. Wei, W. W. Qian, K. A. Little, M. Andres, B. B. Nguyen, T. Moloy, J. Yasonik, J. K. Parker, R. C. Gerkin, J. D. Mainland, and A. B. Wiltschko. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661): 999–1006, 2023.

Leffingwell and Associates. Database of perfumery materials and performance, 2001. URL http://www.leffingwell.com/.

S. Liu, X. Zhang, A. Jiao, S. Sun, L. Dian, and X. Cui. Deep representation learning for electron ionization mass spectra retrieval. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 552–557, 2024.

Y. Nozaki and T. Nakamoto. Odor impression prediction from mass spectra. *PLoS one*, 11(6): e0157030, 2016.

R. Pellegrino, K. Samoilova, Y. Ihara, M. Andres, V. Singh, R. C. Gerkin, A. Koulakov, and J. D. Mainland. A quantitative framework for predicting odor intensity across molecule and mixtures. *bioRxiv*, 2025.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. 2021.

J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

K. J. Rossiter. Structure- odor relationships. *Chemical reviews*, 96(8):3201–3240, 1996.

B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, and A. B. Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

J. Suh, Y. Hong, and C. Park. A comparative study of machine learning models on molecular fingerprints for odor decoding. *Communications chemistry*, 8(1):278, 2025.

S.-H. Sung, J. M. Suh, Y. J. Hwang, H. W. Jang, J. G. Park, and S. C. Jun. Data-centric artificial olfactory system based on the eigengraph. *Nature communications*, 15(1):1211, 2024.

F. Taleb, M. Vasco, A. Ribeiro, M. Björkman, and D. Kragic. Can transformers smell like humans? *Advances in Neural Information Processing Systems*, 37:72032–72060, 2024.

J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science*, 5(4):700–708, 2019.

D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

J. Yan, X. Guo, S. Duan, P. Jia, L. Wang, C. Peng, and S. Zhang. Electronic nose feature extraction methods: A review. *Sensors*, 15(11):27804–27831, 2015.

M. Zhang, Y. Hiki, A. Funahashi, and T. J. Kobayashi. A deep position-encoding model for predicting olfactory perception from molecular structures and electrostatics. *npj Systems Biology and Applications*, 10(1):76, 2024.

Y. Zhang, K. Ding, Y. Wu, X. Zhuang, Y. Yang, Q. Zhang, and H. Chen. Breaking the modality barrier: Generative modeling for accurate molecule retrieval from mass spectra. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1561–1569, 2026.

A SCENT WORKFLOW

This section summarizes the experimental workflow. Detailed model architecture, objectives, and hyperparameters are provided in Section 3 and Appendix C.

Data division. We filter GS-LF dataset with molecules weight in 50-300 Da, remains 2,588 molecule. We first split the 2,588 molecule-spectrum pairs with valid MS spectra into a fixed 10% test set and a 90% training-validation pool using iterative stratified splitting of order 2 (Barsainyan et al., 2023). We don't use scaffold splits which would leave rare descriptors without test positives. Test set molecules are excluded from all upstream retraining and alignment data to avoid leakage.

Chemical encoder preparation. We use two chemical structure encoders as alignment targets: a leakage-controlled Open-POM model retrained after removing the fixed GS-LF test molecules from full GS-LF dataset, and the public MolFormer checkpoint used without fine-tuning. These encoders provide fixed molecular embeddings from SMILES strings. Prospective panel and DREAM molecules are not explicitly excluded from spectrum–structure alignment. Their panel ratings and DREAM continuous ratings are not used to train or select the upstream representations. DREAM ratings are used for downstream ridge-regression training and evaluation within the prescribed train–test splits, whereas panel ratings are used only for evaluation. Upstream Open-POM training uses GS-LF binary odor annotations, so these evaluations assess transfer to the specified perceptual ratings while allowing prior exposure to molecular structures and other annotations.

Alignment data construction. SMILES strings are retrieved from PubChem using CAS identifiers and canonicalized with RDKit after removing stereochemistry (`canonical=True, isomericSmiles=False`). Note that removing stereochemistry doesn't discard information accessible to the model since enantiomers produce identical EI-MS, so stereochemistry-dependent odour differences are unresolvable from EI-MS by any method. After discarding compounds without valid SMILES, 147,913 compound-spectrum pairs remain for alignment. Duplicate canonical SMILES for stereoisomer are retained as independent spectral entries. We therefore evaluate the same-canonical-SMILES collisions induced by this preprocessing. At the training batch size ($B = 512$), they account for only 0.135 entries per batch on average of off-diagonal in-batch comparisons. Thus, its contribution to the alignment objective is negligible under our training protocol. All spectral entries sharing the same canonical SMILES are assigned to the same partition during splitting, preventing entries collapsed by stereochemistry removal from appearing on both sides of a split.

EIMS2Vec spectrum embedding. For a spectrum, each observed peak j is represented by the corresponding center embedding $\mathbf{e}_{\text{cen},j}$, while peak intensity controls its contribution through power scaling. The resulting EIMS2Vec spectrum embedding is

$$u = \sum_j \tilde{x}_j^{0.5} \cdot e_{\text{cen},j}$$

where $\tilde{x}_j$ is the normalized intensity of peak j and $p = 0.5$ compresses the dynamic range.

Spectrum-structure alignment. Each spectrum is represented as a variable-length sequence of observed non-zero peaks, ordered by increasing m/z. Spectra are zero-padded within each mini-batch, and padded positions are excluded from self-attention using a key-padding mask. SCENT initializes spectral peak representations from E_{cen} without the sum pooling. "A learnable `[CLS]` token is prepended and the sequence is processed by a two-layer Transformer encoder with four self-attention heads. The final `[CLS]` representation is linearly projected to the dimension of the corresponding frozen Open-POM or MolFormer embedding. The model is trained by minimizing the InfoNCE loss $\mathcal{L}_{\text{align}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{e^{s_{ii}}}{\sum_{j=1}^{N}e^{s_{ij}}} + \log\frac{e^{s_{ii}}}{\sum_{j=1}^{N}e^{s_{ji}}}\right)$.

The Transformer uses pre-LayerNorm blocks, so normalization is applied only to each sublayer's input while the intensity-scaled peak embeddings pass through the residual stream un-normalized. The intensity weighting is therefore not canceled by LayerNorm.

B HARDWARE

All computational experiments were conducted on a computing node equipped with an AMD EPYC 7502 32-Core Processor and an NVIDIA Quadro RTX 6000 GPU (24GB VRAM). The models were

implemented using the PyTorch framework and optimized with the Adam optimizer. The alignment training on full NIST2023 dataset takes approximately 1 hour per run. The odor prediction task takes 0.5 hour for 5 folds. The linear regression training doesn't need GPU source.

C HYPERPARAMETERS

During alignment, validation contrastive loss controls learning-rate scheduling and checkpoint selection. Training runs for the specified number of epochs, and the checkpoint with the lowest validation loss is retained. Figure C1 reports alignment validation AUC across the temperature grid, $\tau = 0.03$ also achieves the highest AUC among the evaluated settings.The full hyperparameter configuration for each stage of SCENT is reported in Table C1.

For the MSE ablation mentioned in Section 4.1, we replace InfoNCE with MSE without L2-normalizing either embedding. Training settings follow those reported in the Table C1, and the checkpoint with the lowest validation MSE is selected.

The hyperparameters of AE baseline mentioned in Section 4.1 are following the reported parameters in Hasebe et al. (2022).

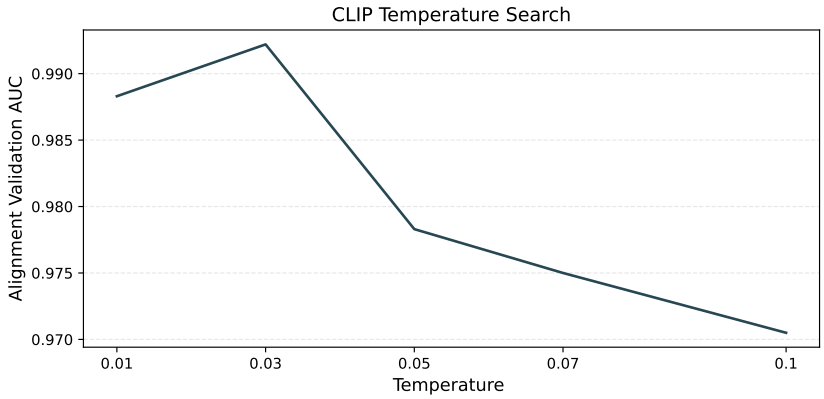


Figure C1: **Contrastive learning performance with different temperature τ.** When $\tau = 0.03$, alignment performance is highest.

D INTRODUCTION OF SIGNAL MODALITY

In this work, we focus on direct EI-MS. Our choice is motivated by its balance between interpretability with real-time accessibility for computational olfaction.

A mass spectrum represents a molecule as a discrete set of peaks, each defined by two quantities: its mass-to-charge ratio (m/z), an integer index reflecting the mass of a molecular fragment produced during ionization, and its intensity, which indicates the relative abundance of that fragment. Together, the pattern of peaks encodes a molecule's fragmentation pattern, which is a structured chemical signature that is reproducible across instruments and directly comparable to reference libraries such as NIST. In this section, we compare direct EI-MS with other sensing modalities and highlight the key rationale behind our selection.

- **Hard vs soft ionization**: Compared to online mass spectrometric techniques such as proton-transfer-reaction mass spectrometry (PTR-MS), which rely on softer ionization and often allow only compound-class level assignments, EI-based single-quadrupole MS enables direct comparison to established spectral libraries like NIST library. This allows us to have a certain degree toward molecular-level interpretability, with substantially reduced instrumental complexity and cost.

Table C1: Hyperparameter settings for training and downstream task.

Module	Hyperparameter	Symbol	Value
MS Encoder	Transformer Layers	L	2
	Attention Heads	H	4
	Embedding Dimension	d	500
	Dropout Rate	p	0.1
	Activation	-	ReLU
Alignment	Batch Size	B	512
	Encoder Learning Rate	η_{enc}	1e-4
	Projection Learning Rate	η_{proj}	1e-3
	Optimizer	-	Adam
	Temperature	τ	0.03
	Max Epochs	-	50
Downstream	Batch Size	B	64
	MLP Learning Rate	η_{cla}	1e-4
	w/o CLIP 2 Encoder Learning Rate	η_{enc_a}	1e-4
	MLP Hidden Dim Ratio	r_{mlp}	0.5
	Max Epochs	-	50

- **Time vs separation**: Unlike GC-EI-MS, which requires time-consuming separation with experts' experience (in the order of hours), direct EI-MS enables rapid acquisition (in the order of seconds).
- **Electronic sensor vs analytical equipment:** In the realm of rapid olfactory sensing, gas sensor arrays are widely adopted. However, they offer only non-specific digital signal fingerprints and lack explicit chemical meaning (Feng et al., 2025). In contrast, analytical instrumentation offers molecular-level information, establishing a rigorous chemical basis for olfactory computation.
- **EI-MS vs tandem MS:** Cross-modality spectrum–structure models have mainly been developed for tandem MS (2). The two modalities are not interchangeable: EI fragments radical cations at 70 eV, whereas MS/MS collisionally fragments soft-ionized, even-electron precursors. Encoders pretrained on tandem M therefore do not transfer directly to EI-MS.

E Additional results: odor descriptor prediction

E.1 Experiment details for prospective panel analysis

For Figure 4B, predictions and panel-mean ratings are separately mean-centered descriptor-wise across molecules before computing per-molecule Pearson correlations. Nearest GS-LF neighbors are identified using Morgan-fingerprint Tanimoto similarity ($r = 2$, 2048 bits). Neighbor odor similarity is computed as cosine similarity between continuous panel ratings and binary GS-LF annotations, with descriptor-wise centering performed separately within each dataset. Figure 4A uses leave-one-panelist-out reference means, pooling correlations across panelists, replicates, and molecules.

E.2 Comparison with retrieval baseline

A natural alternative is to first retrieve the molecule corresponding to a mass spectrum from a spectral library, and then perform structure-based odor prediction using the retrieved molecular identity. To establish the need for the proposed model, we evaluate on the retrieval-fine-tuned EIMS2Vec model using the checkpoint released by the original authors (Liu et al., 2024). The retrieved SMILES is then passed to Open-POM for odor prediction, with the results reported in the Table E1. The

Table E1: Comparison of forward odor prediction and retrieval-based pipelines on the held-out GS-LF test molecules. For open-set retrieval, all reference spectra sharing the query molecule's canonical SMILES are removed before search.

Model	Retrieval	Micro-AUC	Adj.P@5
EIMS2Vec	No	87.98 ± 0.15	46.49 ± 0.53
SCENT (Open-POM)	No	89.80 ± 0.09	50.60 ± 0.53
SCENT (MolFormer)	No	89.99 ± 0.18	50.62 ± 0.75
EIMS2Vec (retrieval fine-tune)	Yes	82.65 ± 0.14	38.38 ± 0.83

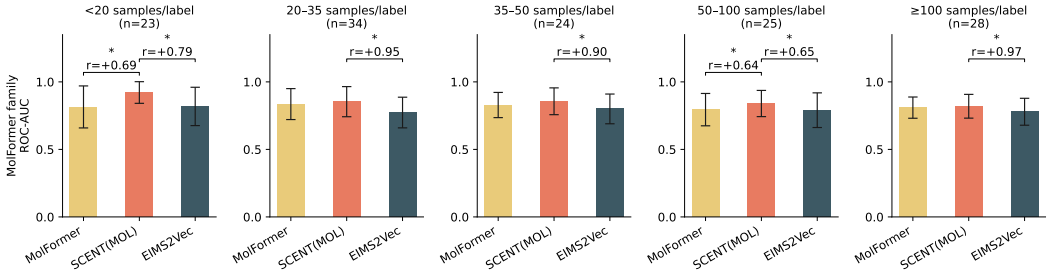


Figure E1: **Mean per-label ROC-AUC grouped by label frequency (MolFormer-based models).** Each bar reports the mean AUC across labels within a given frequency bin (error bars indicate standard error of the mean). SCENT (MolFormer) consistently outperforms EIMS2Vec across all frequency bins, with statistically significant gains even in low-frequency categories. Statistical significance is assessed via the Wilcoxon signed-rank test (* $p < .05$), and the r is the effect size.

comparison is performed in an open-set setting: for each query, all reference spectra sharing its canonical SMILES are removed from the search index, so the correct answer is never directly retrievable. When the analyte is absent from the reference library, direct perceptual prediction with SCENT outperforms the evaluated 2-stage retrieval-based pipeline.

E.3 Label grouped AUC

We further analyze model performance stratified by label frequency across the five training folds to understand where SCENT's gains are concentrated. Figure E1 and Figure 3 reports mean per-label ROC-AUC within five frequency bins, evaluated using the Wilcoxon signed-rank test for pairwise significance (Wilcoxon, 1992). Of the 138 labels, 4 have no positive test samples, leaving 134 for evaluation (see Figure H2A). The results show that SCENT consistently outperforms EIMS2Vec across all bins, with the largest absolute gains in low- and medium-frequency categories, suggesting that chemical structure alignment is particularly beneficial when training signal per label is scarce.

E.4 Performance on chemical similarity group

To better evaluate whether SCENT generalizes beyond close analogues of its training molecules, we stratify the held-out test set by chemical novelty. We define novelty as the maximum Tanimoto similarity (Morgan fingerprint, $r = 2$, 2048 bits) between a test molecule and any molecule in the union of the five training folds, and score each molecule by the ROC-AUC of its 138 descriptor predictions. We group test molecules into quartiles of this similarity and additionally report the subset with similarity below 0.4.

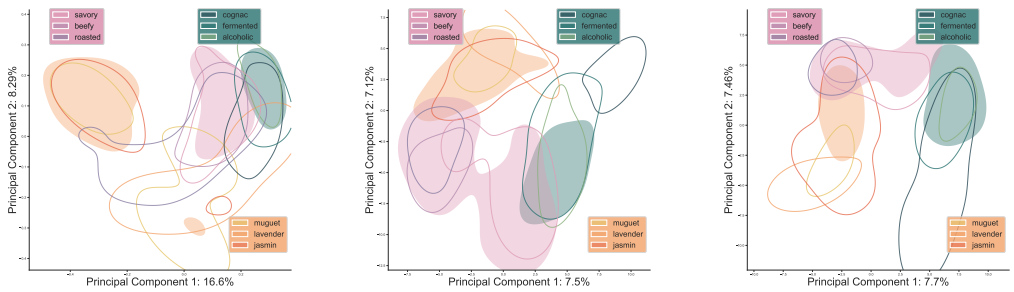
As shown in Table E2, all models perform worse on less familiar molecules, including the two structure-based models that receive the ground-truth SMILES, but the advantage of structure alignment grows as novelty increases. On the most dissimilar subset (Q4), both SCENT variants significantly outperform EIMS2Vec (two-sided paired Wilcoxon signed-rank test, $p = 0.006$ and $p = 0.009$). Because the split is iterative-stratified rather than scaffold-based, this analysis captures graded novelty within the benchmark rather than a regime without analogues. Some test molecules reach a high similarity without being duplicates, because Morgan fingerprints saturate on homologous series.

Table E2: **Per-molecule ROC-AUC on the held-out GS-LF test set, stratified by chemical novelty.** Novelty is the maximum Tanimoto similarity ($r = 2$, 2048 bits) of a test molecule to the union of the five training folds. Mean ± std over the five fold-specific classification heads reported.

			MS input			SMILES input	
Group	n	Mean NN Tanimoto	EIMS2Vec	SCENT (Open-POM)	SCENT (MolFormer)	Open-POM	MolFormer
Q1	75	0.41	84.44 ± 0.15	87.63 ± 0.08	**88.63** ± 0.17	87.15 ± 0.32	85.83 ± 0.20
Q2	75	0.57	87.49 ± 0.32	90.05 ± 0.32	89.92 ± 0.11	**90.34** ± 0.16	88.50 ± 0.27
Q3	74	0.67	88.81 ± 0.37	90.74 ± 0.21	91.26 ± 0.30	**91.52** ± 0.17	90.69 ± 0.26
Q4	75	0.90	91.06 ± 0.31	91.62 ± 0.13	91.81 ± 0.30	**92.44** ± 0.23	91.87 ± 0.30
NN Tanimoto < 0.4	30	0.33	80.05 ± 0.73	84.68 ± 0.20	**86.13** ± 0.41	84.55 ± 0.47	82.21 ± 0.61

E.5 PCA VISUALIZATION FOR FULL GS-LF DATASET

We visualize the low-dimensional representations of the MS embedding using PCA for different odorants and models. In the aligned space, the embedding spaces of both SCENT variants show markedly improved separation between odor categories: the ethereal cluster (cognac, fermented, alcoholic) and the meaty cluster (savory, roasted) become more spatially distinct, with reduced overlap between their respective areas. At the same time, we also observe that there is a decrease in the explained variance ratio of the first two principal components (PCs) after alignment (compared to Figure E2A). The result prove that alignment redistributes information across the embedding dimensions.



(A) EIMS2Vec (Liu et al., 2024) (B) SCENT (MolFormer) (C) SCENT (Open-POM)

Figure E2: **Visualization of the latent space of different MS encoder models.** The shade highlights the space associated with coarse-grained descriptors, while the contour lines mark regions with more precise descriptors. We employ PCA to reduce the dimensionality of the corresponding embeddings.

E.6 DATA SCALING

Figure E3 reports test loss, micro-AUC, macro-AUC and test precision@k across training set fractions. Both metrics confirm that SCENT maintains a consistent advantage over EIMS2Vec at all data scales, with the gap being most pronounced at small fractions (10–20%), further supporting the conclusion that alignment pretraining transfers chemical priors not recovered by the evaluated MS-only baselines.

F ADDITIONAL RESULTS: HUMAN RATING REGRESSION

F.1 STATISTICAL TEST RESULTS OF PEARSON r

To better understand the statistical significance of the perceptual regression results, we assess statistical significance as follows. For each model and perceptual attribute, we first average the test-set Pearson correlations over the 100 repeated train–test splits. Pairwise model comparisons then use

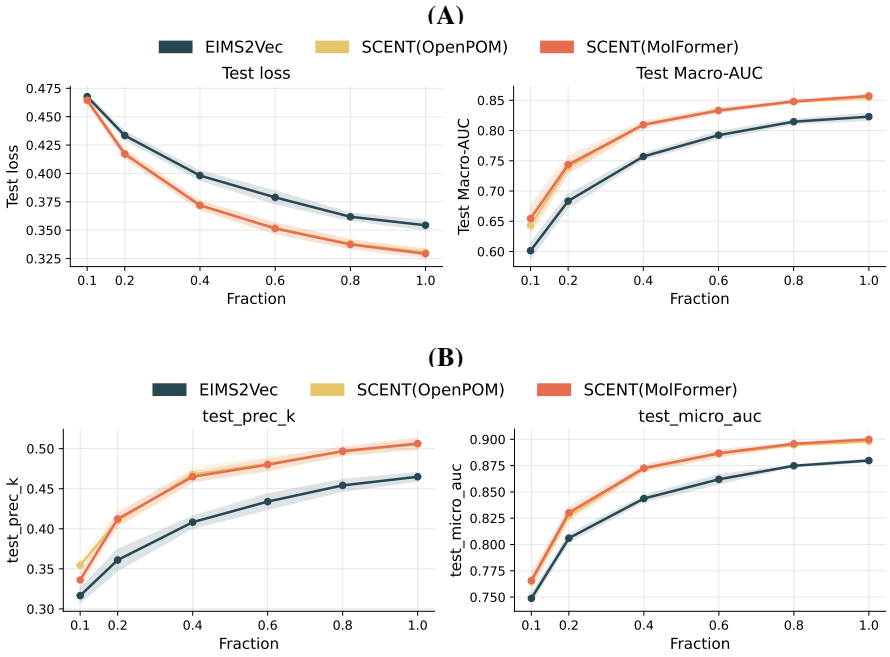


Figure E3: **Performance across training-set sizes.** SCENT (Open-POM), SCENT (MolFormer), and EIMS2Vec are evaluated using varying fractions of the GS-LF training set (0.1-1.0). Results are mean ± std over five folds. Both SCENT variants remain above EIMS2Vec across training-set

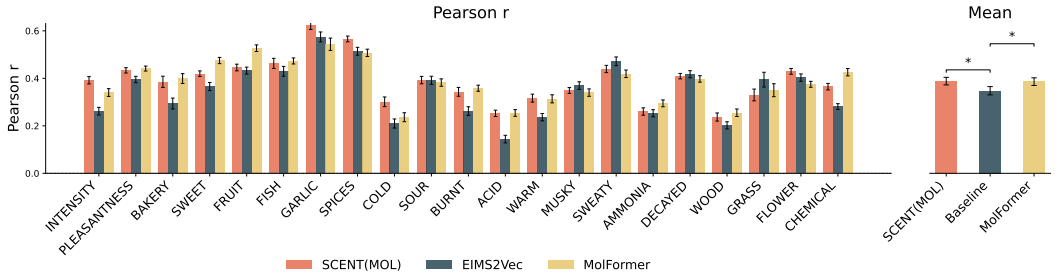


Figure F1: **Pearson correlation between predicted and human-rated perceptual scores on the DREAM dataset (MolFormer family).** Ridge regression (α=1 by grid search) is fitted on frozen embeddings to predict 21 continuous perceptual descriptors (mean ± SEM across 100 repeated train–test splits). SCENT (MolFormer) significantly outperforms the unaligned baseline and is statistically indistinguishable from MolFormer despite requiring no molecular structure at inference. Statistical significance is assessed via the Wilcoxon signed-rank test (Wilcoxon, 1992) (* $p < .05$). More results in Appendix F.1.

a two-sided Wilcoxon signed-rank test (Wilcoxon, 1992) on the resulting 21 paired attribute-level mean correlations (n = 21).

As shown in Table F1, both variants of SCENT significantly outperform EIMS2Vec baseline, suggesting that the alignment introduces the structure-related odor information from both supervised and self-supervised chemical model. At the same time, SCENT (MolFormer) also achieves a performance without significant gap with MolFormer, while SCENT (Open-POM) still remains a gap with its chemical teacher. That is expected given that Open-POM embeddings are directly optimized with odor supervision and therefore contain task-specific perceptual structure that cannot be fully recovered from MS signals alone.

F.2 Concentration sensitivity analysis

To evaluate how SCENT prediction align with human rating in different concentration level, we analyses concentration sensitivity for 3 concentration levels in DREAM dataset by training regression on

Table F1: Pairwise Wilcoxon signed-rank tests on per-attribute Pearson r (n=21).

Group	Comparison	W	*p*-value	Sig.
Open-POM-based	SCENT (Open-POM) vs. EIMS2Vec	55.0	0.035	*
	SCENT (Open-POM) vs. Open-POM	3.0	<0.001	*
MolFormer-based	SCENT (MolFormer) vs. EIMS2Vec	29.0	0.002	*
	SCENT (MolFormer) vs. MolFormer	115.0	1.000	—

each level. Since the molecules in different concentration levels are different, the comparison only happened within each concentration group. The experiment answers the question of how similar the model's predictions are to human perception at different concentrations if the model does not encode concentration information. The Pearson correlation (mean±std) are shown in the table below, suggesting that the benefit of chemical alignment is retained across most of the evaluated concentration regimes, while performance becomes less distinguishable at the most dilute condition.

Table F2: Concentration sensitivity analysis on the DREAM dataset. Pearson correlation coefficients are reported as mean ± standard deviation.

Dilution	Sample Number	EIMS2Vec	SCENT (POM)	SCENT (MOL)	Open-POM	MolFormer
1/10	102	0.184 ± 0.174	0.263 ± 0.147	0.300 ± 0.151	0.368 ± 0.083	0.279 ± 0.126
1/1,000	266	0.261 ± 0.077	0.296 ± 0.085	0.325 ± 0.092	0.415 ± 0.095	0.344 ± 0.103
1/100,000	283	0.307 ± 0.148	0.318 ± 0.139	0.330 ± 0.145	0.400 ± 0.147	0.295 ± 0.138

F.3 Effect of explicit dilution information

We investigate whether reported dilution provides information beyond the frozen spectral representation. Unlike the main regression experiment, which uses one highest-available-concentration profile per molecule, this analysis retains both measured dilution conditions for 320 molecules, yielding 640 molecule–dilution observations. We use the 250 molecule-disjoint DREAM train–test splits and keep all observations from the same molecule on the same side of each split.

Let $\mathbf{z}_m \in \mathbb{R}^D$ denote the frozen spectral embedding of molecule m, and let c_{mj} denote the standardized log dilution of its j-th presentation condition. We compare three ridge-regression models.

The concentration-blind baseline uses only the molecular representation:

$$\hat{\mathbf{y}}_{mj} = W_z \mathbf{z}_m + b.$$

The dilution-concatenation model adds dilution as an auxiliary scalar:

$$\hat{\mathbf{y}}_{mj} = W_z \mathbf{z}_m + w_c c_{mj} + b.$$

Finally, the interaction model additionally includes the element-wise embedding–dilution product:

$$\hat{\mathbf{y}}_{mj} = W_z \mathbf{z}_m + w_c c_{mj} + W_{\text{int}}(c_{mj} \mathbf{z}_m) + b.$$

The dilution slope allowing the dilution effect to vary with the molecular representation while keeping the spectral encoder frozen.

Table F3 shows that providing dilution improves prediction for all three encoders. Concatenating dilution with the frozen embedding shows that presentation condition provides predictive information beyond the molecular representation itself. For the two SCENT representations, adding the embedding-dilution interaction provides a further improvement, indicating that the effect of dilution is better modeled as dependent on the molecular representation rather than as a single global shift shared across molecules.

Table F3: Effect of explicit dilution information on molecule-disjoint perceptual regression. Both measured dilution conditions are retained for each molecule. Results are mean ± standard deviation across 250 DREAM splits.

Condition	SCENT (Open-POM)	SCENT (MolFormer)	EIMS2Vec
$\mathbf{z}$	0.301 ± 0.039	0.325 ± 0.040	0.254 ± 0.047
$[\mathbf{z}, c]$	0.326 ± 0.037	0.349 ± 0.039	0.288 ± 0.046
$[\mathbf{z}, c, c\mathbf{z}]$	$\mathbf{0.362 \pm 0.034}$	$\mathbf{0.371 \pm 0.036}$	$\mathbf{0.290 \pm 0.043}$

G Additional results: case study

G.1 Data collection and pre-processing.

Laboratory spectra were acquired using an Agilent 5973 inert (G2579A) quadrupole mass spectrometer. A 2-m methyl-deactivated fused-silica capillary (0.15-mm internal diameter) was used as a direct atmospheric inlet and flow restrictor, with flow driven by the vacuum system of the mass spectrometer. The capillary was enclosed in a heated transfer line capable of temperatures up to 150 °C, with an estimated flow rate of 0.6–1.1 $\mathrm{mL\,min^{-1}}$ depending on operating temperature.

The raw MS signals are collected in laboratory air, comprising a background period (about 0.3 min) followed by sample exposure (about 0.7 min), with an overall measurement duration of around 60 seconds. For each time point, we compute the total ion intensity by summing signals across all m/z channels, forming a temporal intensity trace that reflects the sample elution profile. We apply a sliding-slope algorithm with a 20-time point window to this temporal intensity trace to identify key boundaries of the sample region, where the steepest increase indicates the onset of sample elution and the subsequent increase of slope marks the tailing start and signal termination points. This procedure separates the sample window from background regions. For each m/z channel, we estimate noise statistics by computing the mean μ_{bg} and standard deviation σ_{bg} over time. Baseline correction is then performed by subtracting a threshold defined as $\mu_{\mathrm{bg}} + 5\sigma_{\mathrm{bg}}$, which suppresses background noise while preserving statistically significant ion signals. Finally, the corrected intensities within the sample window are averaged over time to form the model input.

G.2 Real-world dataset

We provide the molecules used for the real-world test set in Table G1. The first 46 molecule labels are from the GS-LF dataset. The labels for the last four molecules are from GoodScent (2026). All those 50 molecules will not appear in any stage of training in real-world adaption stage to avoid data leakage.

G.3 Chemical model ablation

To better understand the effect of selecting chemical model as alignment target, we test the real-world samples in SCENT (MolFormer), and the results are shown in Table G2. Due to the unsupervised training paradigm only on SMILES string, the MolFormer family performance not as well as Open-POM family, but still outperform the MS-only baseline, indicating the advantage of introducing structure alignment.

G.4 Paired bootstrap comparison

To better evaluate the uncertainty arising from the small laboratory set, we resample the 50 molecules with replacement 10,000 times, evaluate all models on each resampled set, average each metric over the five classification heads, and report the $2.5th$ and $97.5th$ percentiles. Because every model is scored on the same molecules, molecule difficulty shifts all scores together. This widens each marginal interval but cancels in the difference between models. Table G3 therefore reports the paired difference SCENT − EIMS2Vec per replicate. All six intervals exclude zero, including Macro-AUC, whose marginal intervals overlap.

Table G1: **Compounds used for evaluation on independently acquired laboratory EI-MS spectra.** Reference odor annotations are obtained from GS-LF and GoodScents. SMILES are canonical non-isomeric representations generated with RDKit.

Canonical SMILES	Formula	Descriptor Number	Odor Descriptors (OD)
CC(=O)O	$C_2H_4O_2$	3	pungent, sharp, sour
CC(C)=O	C_3H_6O	4	apple, ethereal, pear, solvent
O=Cc1ccccc1	C_7H_6O	9	almond, bitter, cherry, fruity, nutty, oily, powdery, sharp, sweet
CCCC=O	C_4H_8O	7	chocolate, cocoa, fruity, green, malty, musty, pungent
CCCCO	$C_4H_{10}O$	7	alcoholic, balsamic, fermented, oily, solvent, sweet, winey
CC(=O)C(C)=O	$C_4H_6O_2$	5	buttery, caramellic, creamy, pungent, sweet
CC12CCC(CC1)C(C)(C)O2	$C_{10}H_{18}O$	5	camphoreous, cooling, fresh, herbal, medicinal
CCC(C)=O	C_4H_8O	6	camphoreous, ethereal, fruity, ketonic, solvent, sweet
CC1=CCC2CC1C2(C)C	$C_{10}H_{16}$	13	aromatic, camphoreous, cooling, earthy, fresh, herbal, mint, pine, sharp, sweet, terpenic, warm, woody
COc1ccccc1	C_7H_8O	4	anisic, ethereal, phenolic, sweet
CCCC(=O)OCC	$C_6H_{12}O_2$	8	banana, buttery, cognac, ethereal, fruity, juicy, pineapple, ripe
CCCCC(C)=O	$C_6H_{12}O$	10	banana, dairy, ethereal, fermented, fruity, ketonic, rummy, sweet, winey, woody
C=CC(C)(C)O	$C_5H_{10}O$	5	earthy, ethereal, fruity, herbal, oily
CCCCCCOC(C)=O	$C_8H_{16}O_2$	8	apple, banana, fatty, fresh, fruity, green, pear, sweet
C=CCOC(=O)Cc1ccccc1	$C_{11}H_{12}O_2$	5	fruity, honey, pineapple, rummy, sweet
C=CC(O)CCCCC	$C_8H_{16}O$	12	earthy, fruity, grassy, green, herbal, melon, mushroom, musty, oily, sweet, vegetable, violet
Cc1nccnc1C	$C_6H_8N_2$	10	caramellic, cocoa, coffee, green, meaty, musty, nutty, potato, powdery, roasted
C=C(C)C1CC=C(C)CC1	$C_{10}H_{16}$	11	aromatic, citrus, fresh, herbal, lemon, orange, phenolic, pine, sweet, terpenic, woody
CC(=O)c1cnccn1	$C_6H_6N_2O$	8	chocolate, coffee, hazelnut, musty, nutty, popcorn, potato, roasted
CCC=CCCO	$C_6H_{12}O$	15	apple, cortex, earthy, fatty, floral, fresh, fruity, grassy, green, herbal, leafy, melon, oily, pungent, vegetable
Cc1ccccc1	C_7H_8	1	sweet
O=Cc1ccco1	$C_5H_4O_2$	6	almond, caramellic, phenolic, spicy, sweet, woody
CCOC(C)=O	$C_4H_8O_2$	10	brandy, ethereal, fruity, grape, green, rummy, sharp, sweet, weedy, winey
C=CC(C)=CCC=C(C)C	$C_{10}H_{16}$	12	anisic, citrus, floral, green, herbal, lemon, sweet, terpenic, tropical, vegetable, warm, woody
CC1=CCC(=C(C)C)CC1	$C_{10}H_{16}$	9	anisic, citrus, fresh, fruit skin, herbal, lemon, pine, sweet, woody
C=CC(C)=O	C_4H_6O	5	ethereal, gassy, pungent, sharp, sweet
CC(C)O	C_3H_8O	3	alcoholic, musty, woody
CCCC=CC=O	$C_6H_{10}O$	5	fresh, fruity, green, leafy, vegetable
CC(C)CC(=O)O	$C_5H_{10}O_2$	9	cheesy, dairy, fatty, fruity, pungent, ripe, sour, sweaty, tropical
COc1ccccc1O	$C_7H_8O_2$	9	meaty, medicinal, phenolic, savory, smoky, spicy, sweet, vanilla, woody
CC1CCC(C(C)C)C(O)C1	$C_{10}H_{20}O$	11	bitter, camphoreous, cooling, earthy, hay, herbal, mint, musty, pine, sweet, woody
OCCc1ccccc1	$C_8H_{10}O$	6	dry, floral, fresh, honey, rose, sweet
COc1cc(C=O)ccc1O	$C_8H_8O_3$	4	chocolate, creamy, sweet, vanilla
CC1=CCC2C(C1)C2(C)C	$C_{10}H_{16}$	10	citrus, fresh, herbal, medicinal, phenolic, pine, solvent, sweet, terpenic, woody
CCCCC(CC)CO	$C_8H_{18}O$	7	citrus, fermented, floral, fresh, oily, rose, sweet
CCCC(C)=O	$C_5H_{10}O$	10	banana, dairy, ethereal, fermented, fruity, ketonic, rummy, sweet, winey, woody
C=C1C2CCC(C2)C1(C)C	$C_{10}H_{16}$	13	camphoreous, citrus, cooling, fresh, green, herbal, mint, oily, pine, spicy, sweet, terpenic, woody
CC12CCC(CC1=O)C2(C)C	$C_{10}H_{16}O$	9	camphoreous, cooling, fresh, green, herbal, medicinal, mint, phenolic, woody
C=C(C)C1CC=C(C)C(=O)C1	$C_{10}H_{14}O$	4	herbal, mint, spicy, sweet
CSSC	$C_2H_6S_2$	4	cabbage, onion, sulfurous, vegetable
CSC	C_2H_6S	14	berry, cabbage, cooked, creamy, fishy, fruity, green, onion, pungent, radish, sulfurous, sweet, tomato, vegetable
CCCCCCC=O	$C_7H_{14}O$	10	aldehydic, citrus, cognac, fatty, fresh, green, herbal, nutty, ozone, sweet
CCCCCC=O	$C_6H_{12}O$	11	aldehydic, clean, fatty, fresh, fruity, grassy, green, leafy, sweaty, vegetable, woody
C=CC(=C)CCC=C(C)C	$C_{10}H_{16}$	7	balsamic, celery, herbal, rose, spicy, terpenic, woody
C=Cc1ccccc1	C_8H_8	5	balsamic, floral, gassy, pungent, sweet
CCCCC=O	$C_5H_{10}O$	6	berry, chocolate, fermented, fruity, nutty, pungent
Cc1ccccc1O	C_7H_8O	5	herbal, leathery, medicinal, musty, phenolic
CC=Cc1ccc(OC)cc1	$C_{10}H_{12}O$	2	medicinal, sweet
O=CC=Cc1ccccc1	C_9H_8O	8	aldehydic, aromatic, balsamic, cinnamon, honey, powdery, spicy, sweet
CC1CCC(C(C)C)C(=O)C1	$C_{10}H_{18}O$	13	anisic, berry, camphoreous, cooling, dry, fresh, fruity, green, herbal, mint, peach, sweet, woody

Table G2: **Performance comparison between EIMS2Vec and SCENT (MolFormer) on independently acquired spectra.** Results are reported as mean $\pm$ standard deviation across the five classification heads, together with 95% confidence intervals obtained by bootstrap resampling over the 50 measured molecules. Higher values indicate better performance.

Metric	Statistic	EIMS2Vec	SCENT (MolFormer)
Micro-AUC↑	mean $\pm$ std	84.69 ± 0.65	88.22 ± 0.41
	95% CI	$[82.85, 86.46]$	$[86.60, 89.83]$
Macro-AUC↑	mean $\pm$ std	78.47 ± 1.19	81.74 ± 0.83
	95% CI	$[74.43, 81.46]$	$[78.28, 84.50]$
Adj.P@5↑	mean $\pm$ std	42.32 ± 1.88	54.03 ± 1.22
	95% CI	$[36.74, 47.72]$	$[48.01, 60.08]$

G.5 Library-to-laboratory performance comparison

To quantify the performance change from library to laboratory spectra, we evaluate the same five fold-specific checkpoints on paired spectra for the 50 laboratory samples, keeping odor labels and evaluation preprocessing fixed. The metrics are computed within each of the five classification heads and averaged. Table G4 reports performance under both input conditions and the paired change, $\Delta_{lib} = M_{library} - M_{lab}$.

Table G3: **Paired bootstrap comparison on the 50 independently acquired laboratory spectra.** Each of 10,000 replicates resamples the 50 molecules with replacement and applies the same resampled set to all models. Entries are the mean paired difference in percentage points with the $2.5th$ and $97.5th$ percentiles.

Metric	SCENT (Open-POM) − EIMS2Vec	SCENT (MolFormer) − EIMS2Vec
Micro-AUC	+3.94 [+2.73, +5.21]	+3.53 [+2.34, +4.80]
Macro-AUC	+3.62 [+1.59, +6.14]	+3.27 [+0.97, +5.92]
Adj.P@5	+15.25 [+9.73, +20.91]	+11.71 [+6.10, +17.30]

Table G4: Laboratory performance on the 50 measured compounds (MS input within [50, 180]), with the paired change relative to library spectra on the same compounds. Δ_{lib} denotes change relative to library MS within [50, 180] input on the same split.

Model	Input	Micro-AUC↑	Δ_{lib}	Macro-AUC↑	Δ_{lib}	Adj.Precision@k↑	Δ_{lib}
EIMS2Vec	MS[50,180]	84.69 ± 0.65	↓0.44	78.47 ± 1.19	↓0.24	42.32 ± 1.88	↓1.14
SCENT (Open-POM)	MS[50,180]	88.63 ± 0.29	↓0.61	82.09 ± 0.61	↓0.94	57.57 ± 1.14	↑0.09
SCENT (MolFormer)	MS[50,180]	88.22 ± 0.41	↓0.74	81.74 ± 0.83	↓1.26	54.03 ± 1.22	↓0.68

H LABEL DISTRIBUTION

GS-LF dataset has a long-tail feature. To better understand how label distribution affects the results, and to what extend the real-world dataset label distribution similar with the training dataset, we visualize the label distribution among the full GS-LF dataset and real sample validation set. According to Figure H1, "Both datasets exhibit a long-tailed label distribution. The laboratory set covers common descriptors and also includes descriptors that are rare in GS-LF.

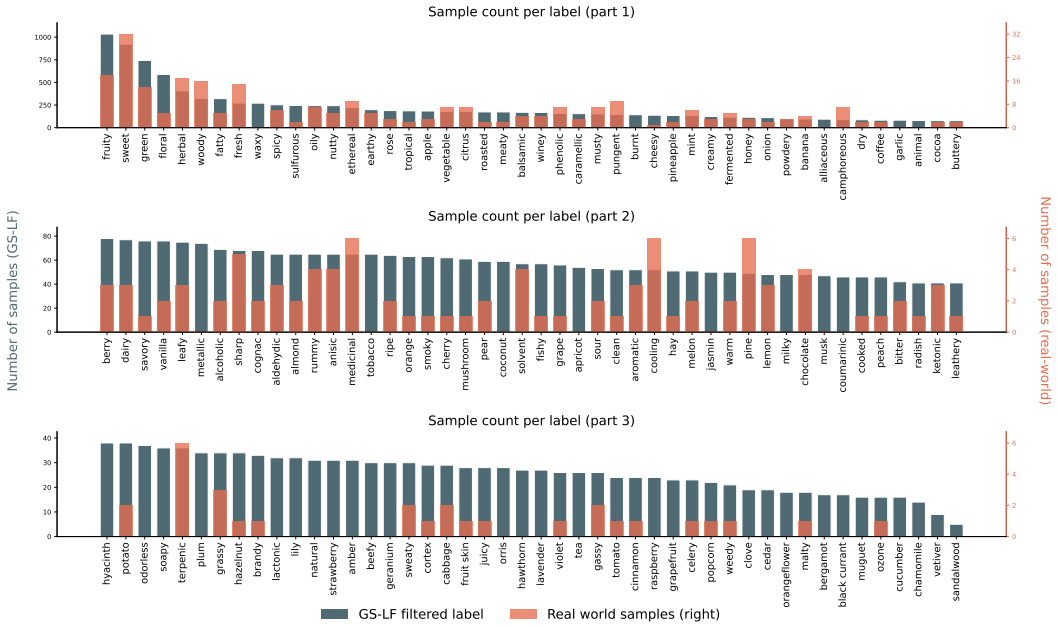
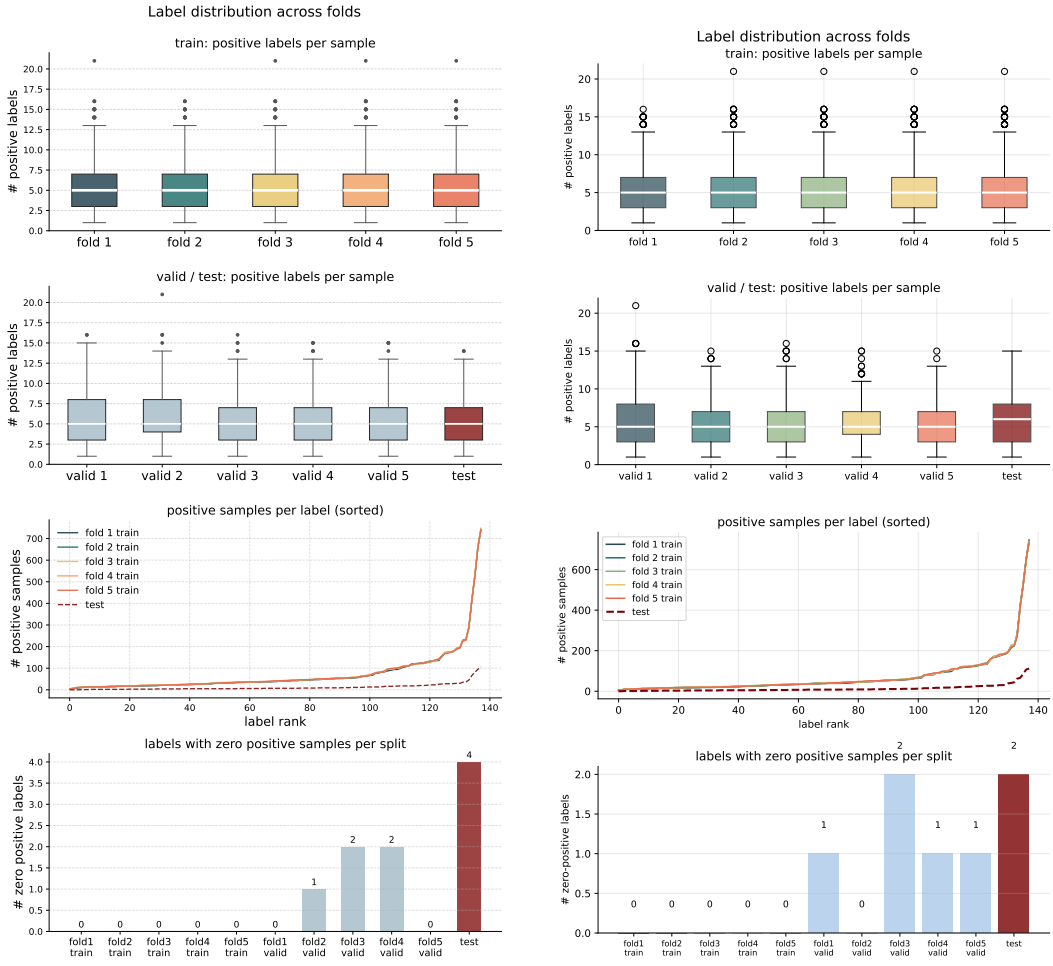


Figure H1: **Label distribution among filtered GS-LF dataset and laboratory measured sample dataset.**

For each fold and test set for odor prediction task, the label distribution for full spectra experiment and [50, 180] experiment are shown in Figure H2A and Figure H2B respectively. The top panels show the distribution of positive labels per sample across folds. Both training and validation splits maintain a consistent median of approximately 5 positive labels per sample, with the test set showing



(A) **Full spectrum**.

(B) **m/z in [50, 180] downstream task**.

Figure H2: Label distributions of the downstream task.

a similar range, confirming that iterative stratified splitting preserves the label co-occurrence structure across all splits. The bottom-left panel shows the number of positive samples per label sorted by frequency, reflecting the long-tail nature of the GS-LF dataset. The bottom-right panel reports the number of labels with zero positive samples in each split. All training folds contain no zero-positive labels, while a small number appear in validation folds and the test set, which is expected given the class imbalance in rare odor categories.