

Introduction

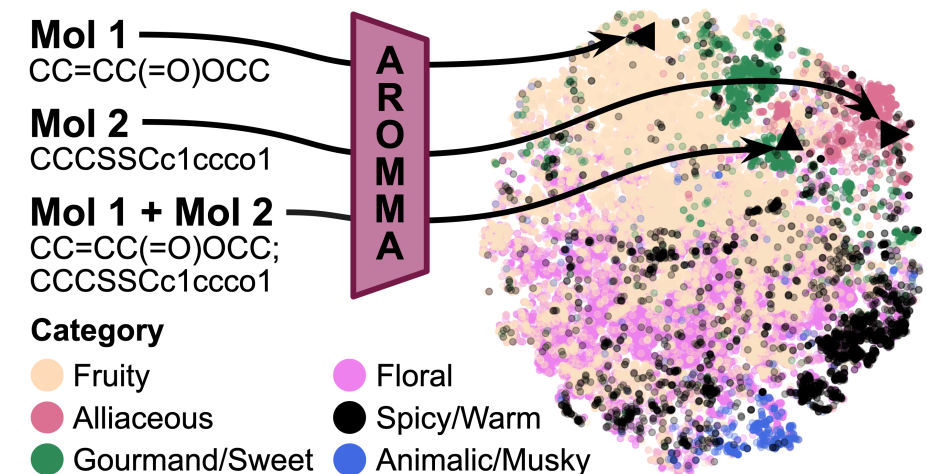
Open Problem in AI for Science: Can we map molecular structure to human olfactory perception?

Challenge

- public datasets (multi-label classification) are highly heterogeneous
 - descriptor set → no canonical representation, different cardinality (GS-LF=138, BP=74)
 - active label density → GS-LF (5 labels, fine-grained), BP (1.4 labels, coarse-grained)
- mixture perception reflect non-linear effects (agonism/antagonism)

Contribution

- unified embedding space that aligns single molecules and mixtures
- leverage a chemical foundation model (FM) providing robust molecular representations
- introduce an aggregator module capturing asymmetric component interactions
- mitigate heterogeneity in the label distribution of different datasets



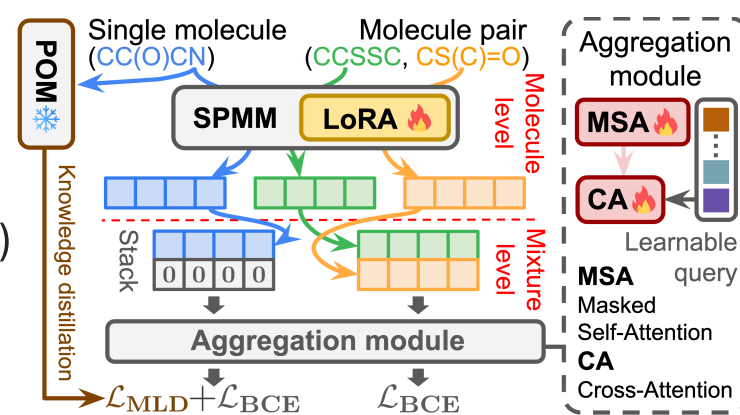
*GS-LF&POM (Lee, Science 2023), BP&MPNN-GNN (Sisson, ACS Omega 2025)

AROMMA

Aggregated Representations of **O**lfaction via **M**olecule and **M**ixture **A**lignment

Architecture

- molecule-level
 - chemical FM (SPMM)
 - knowledge distillation from pre-trained olfaction model (POM)
- mixture-level
 - asymmetric interaction (MSA)
 - permutation invariance (CA)



Training strategy

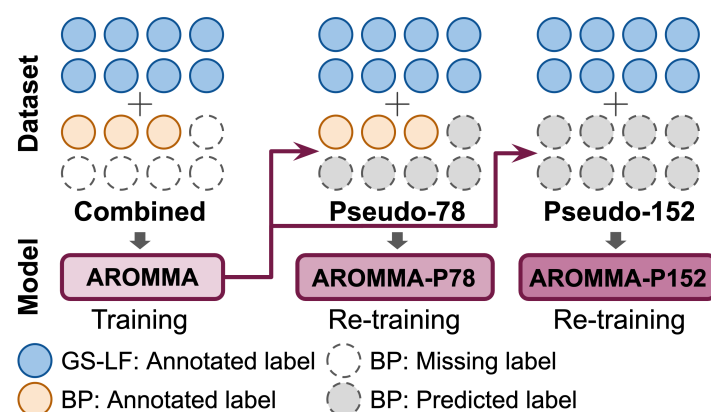
class-aware pseudo-labeling

- proportion of positive labels in GS-LF

$$\gamma_c = \frac{\sum_{j=1}^n \mathbb{I}(y_{jk} = 1)}{n}$$

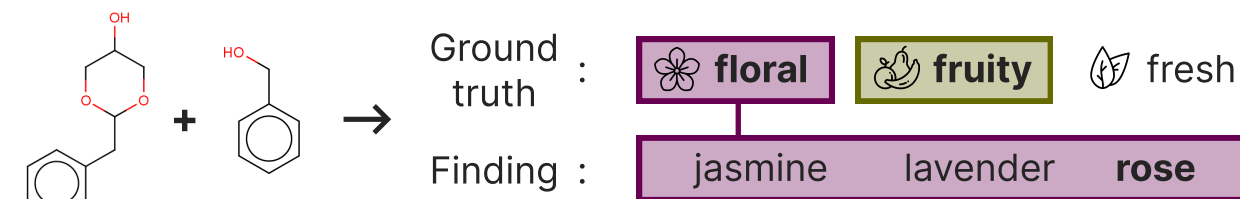
- class specific threshold τ_c for BP

$$\frac{\sum_{j=1}^n \mathbb{I}(p_{jc} \geq \tau_c)}{n} = \gamma_c$$



Results

Class-aware pseudo-labeling



Average AUROC (↑) scores on GS-LF (single molecules) +3.2%, BP (mixtures) +19.1%, and their combination

Model	Combined	GS-LF	BP
POM	-	0.874±0.001	-
MPNN-GNN	-	0.851±0.003	0.734±0.019
AROMMA (ours)	0.939±0.002	0.900±0.001	0.861±0.006
AROMMA-P78 (ours)	0.931±0.003	0.902±0.001	0.859±0.008
AROMMA-P152 (ours)	0.932±0.004	0.902±0.001	0.874±0.009

Ablation study on the main components of AROMMA

Embedder	Agg	KD	Combined	GS-LF	BP
POM(freeze)	CA		0.898±0.003	0.817±0.006	0.851±0.010
SPMM(freeze)	CA	✓	0.935±0.003	0.897±0.001	0.853±0.007
SPMM(LoRA)	PNA	✓	0.859±0.006	0.881±0.003	0.848±0.009
SPMM(LoRA)	CA		0.890±0.004	0.769±0.005	0.830±0.012
SPMM(LoRA) (ours)	CA	✓	0.939±0.002	0.900±0.001	0.861±0.006