Multimodal KDD 2023: International Workshop on Multimodal Learning

Optimizing Learning Across Multimodal Transfer Features for Modeling Olfactory Perception

Daniel Shin

Stanford university and Sony AI USA

Gao Pei

NAIST Japan and Sony AI Tokyo

Priyadarshini Kumari

Sony AI USA

Tarek Besold

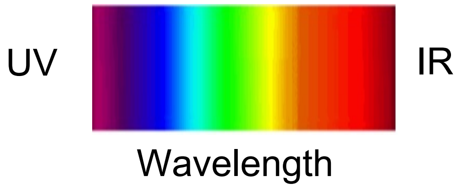
Sony AI Spain

Why is olfaction important?

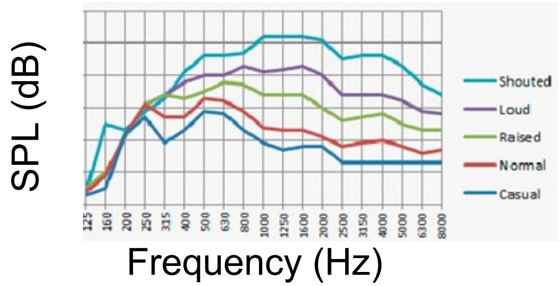
- Generating synthetic odorants
- Drug discovery
- Multimodal user interfaces
- Gastronomy – food recommendation/substitution
- Creating an immersive AR/VR system
- Providing a sense of smell to those who have anosmia

Image courtesy: Jennifer et.al. MIT news

Visual



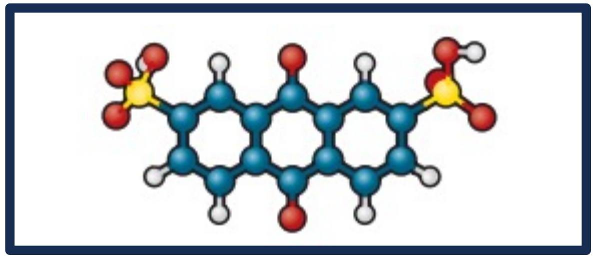
Auditory



Challenges

- Odor space is vast
- Biological mechanism is highly complex and little understood
- No intuitive set of molecular features for characterizing olfactory stimuli
- Olfaction is severely data-limited and highly skewed

Data-limited or Model-limited?



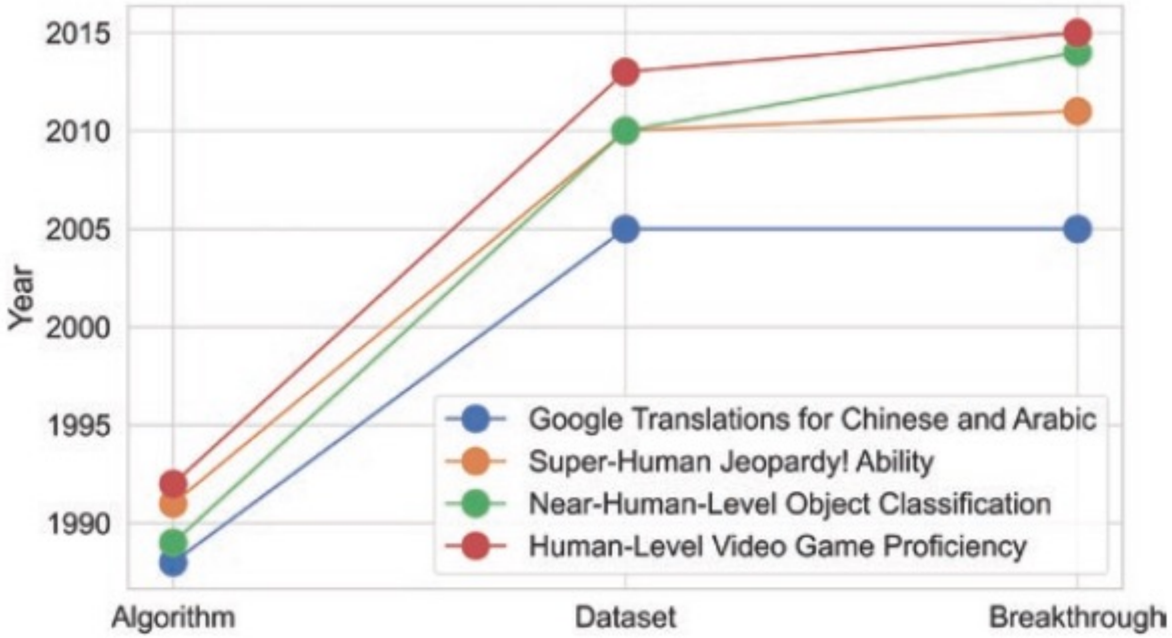
Molecular features
physical/chemical/structural information

"Roasty"
"Sweet"
"Vanilla"
"Citrus"
"Fruity"
"Spicy"
"Alcoholic"

Perceptual descriptors

Benchmark Datasets

- Andrew Dravnieks, 1985 – 138 molecules rated using 146 semantic descriptors
- Keller and Vosshall BMC neuroscience 2016 - 480 molecules rated by 21 descriptors



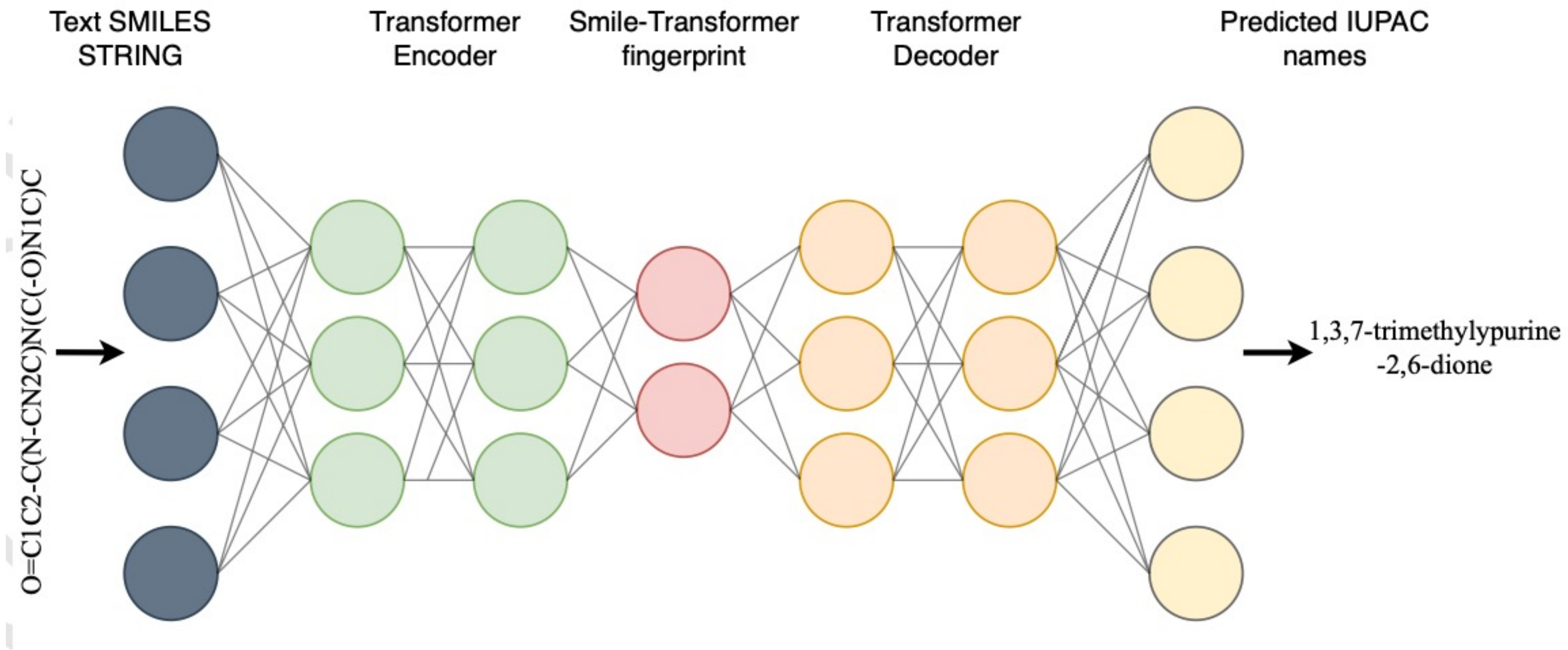
Transfer and multimodal features to address data scarcity

- Large molecular foundation models such as SMILES transformer, ChemBERT, and MolCLR are trained on PubChem and ZINC consisting of millions of molecules
- Un-supervised or self-supervised trained
- Demonstrated significant efficacy in drug discovery, protein folding, etc.
- How effective could a model be in a perceptual task without prior perceptual training?

Contributions

- Introduce a data-efficient olfactory perceptual model by leveraging multimodal transfer learning
- Investigate how different modality molecular representations contribute to olfactory perception modeling
- Introduce a label-balancer technique to address the problem of label skewness in the olfactory domain

Text-based transfer features

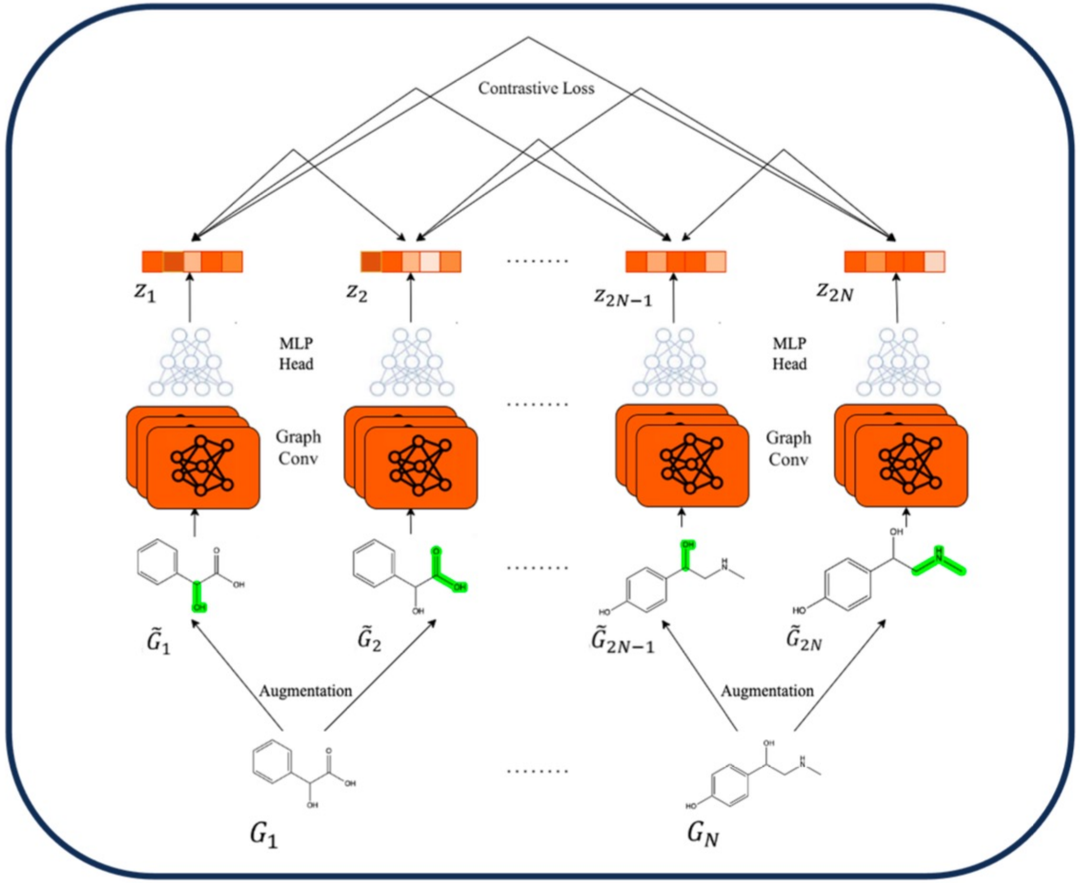


- Transformer trained on 83M SMILES from PubChem through self-supervised task of SMILES-IUPAC translation
- Both text-based molecular representations, SMILES and IUPAC, use a language model
- Intermediate transfer features obtained from pre-trained network are perceptually calibrated by fine-tuning using a small subset of perceptual descriptors

Graph-based transfer features

- MolCLR trained on 10M molecule graph through self-supervised contrastive loss
- Graph augmentation techniques – atom masking, bond deletion, and subgraph removal
- Model trained with NT-Xent loss

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}\{k \neq i\} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$



MolCLR perceptually calibrated on limited labeled data

[MolCLR Yuyang et. al. 2022]

Multimodal training using label-balancer

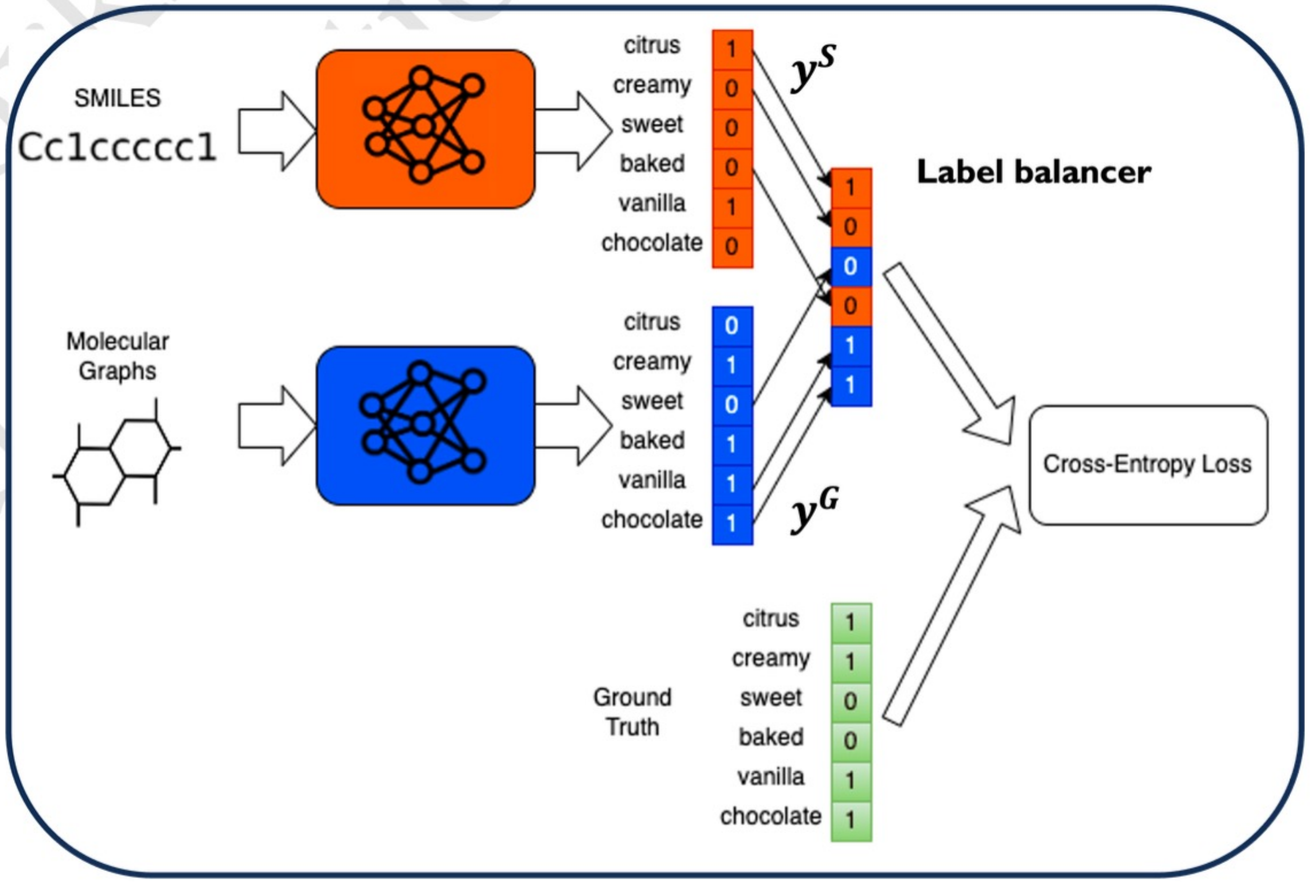
- MLP head : Combine graph and text-based molecular representations

$$z_M = f_M(f_S(z_S) \oplus f_G(z_G))$$

- f_M combines two modality features with optimal weights based on their perceptual relevance

- Label-Balancer: ensemble of models optimized for orthogonal subsets of perceptual labels

- Mitigate over-fitting and offers better generalization on rare-class test samples



Label-balancer: Objective function

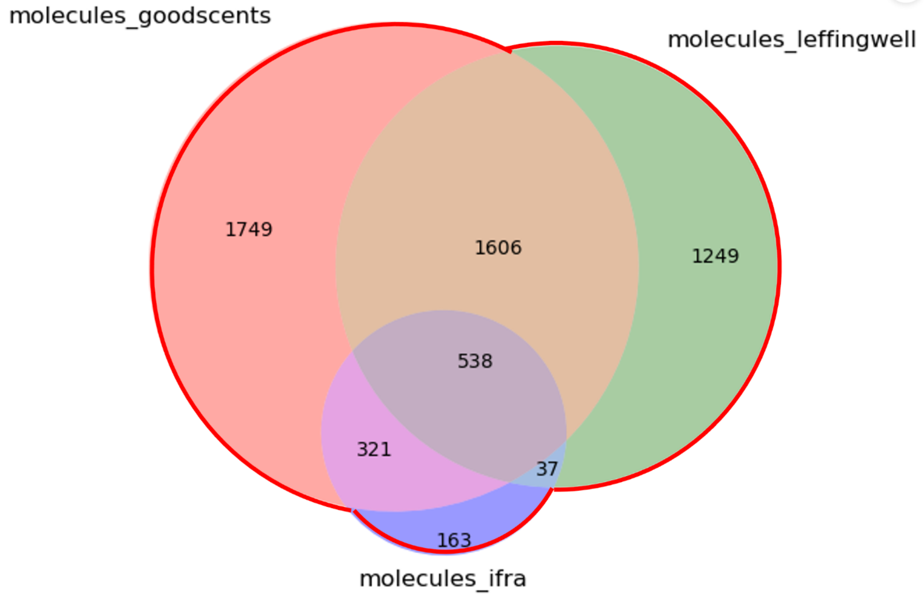
- Distributed objective function across different modalities

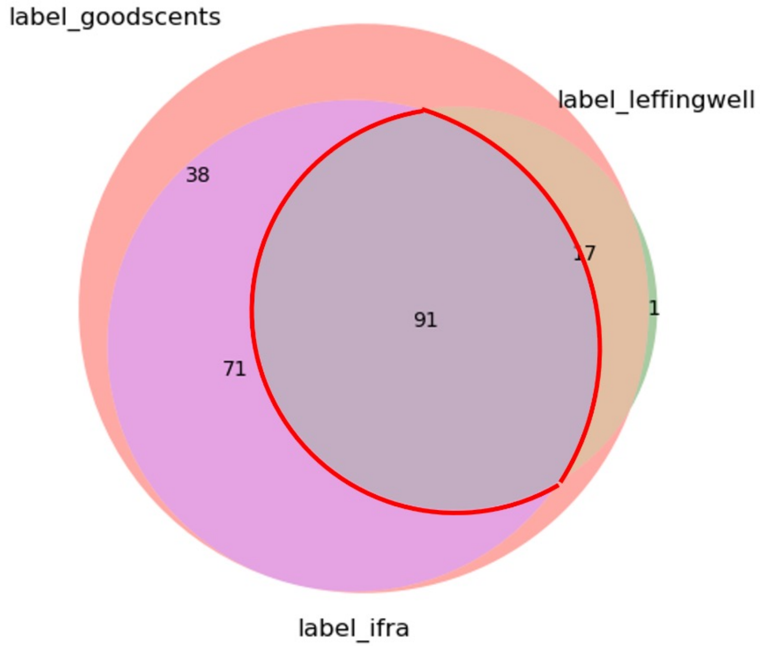
$$L_{ce} = -\sum_{i=1}^{L} logp_i^y \mathbb{1}_{y^S}(i) - logp_i^y \mathbb{1}_{y^G}(i)$$

- y_S and y_G are complementary label subsets optimized by SMILES transformer and GNN, respectively
- Division of labels among different modalities enables learning diverse and perceptually effective features
- Better generalization on sparsely represented class samples

Label-balancer technique outperforms classical multi-modality fusion approaches

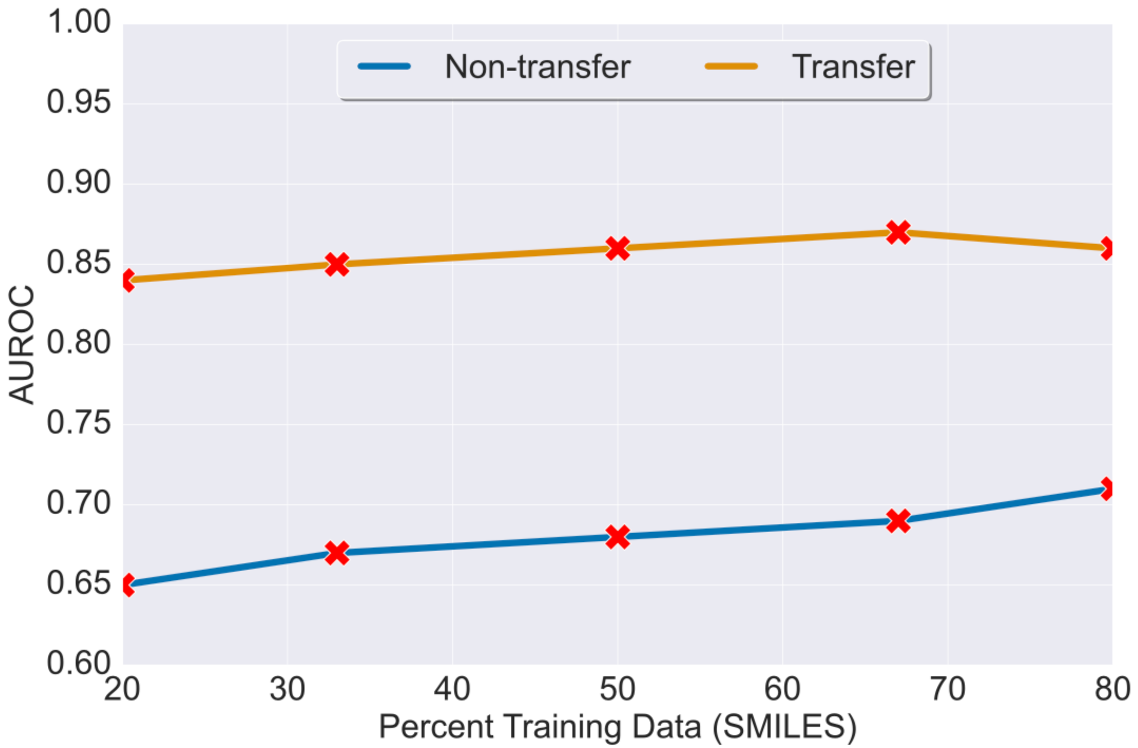
Evaluation - Dataset

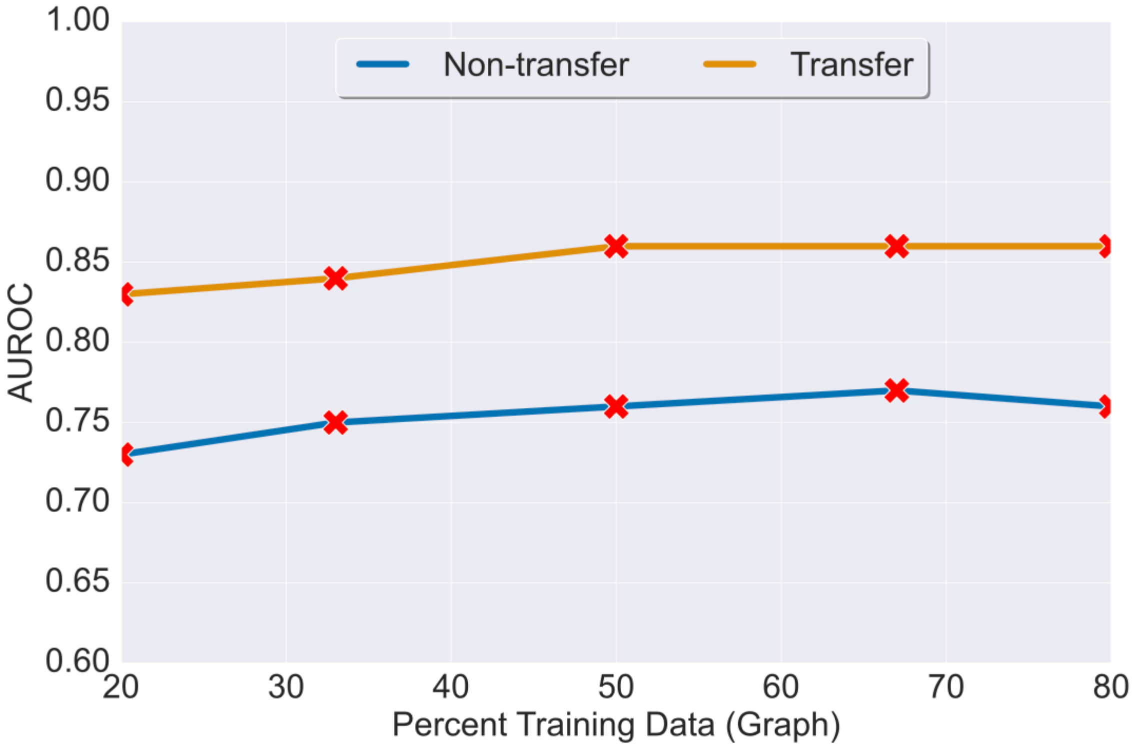




- Dataset source – a) Goodscents b) Leffingwell c) IFRA
- Curated dataset – 5663 molecules gathered from three data sources described by 91D perceptual descriptors

Evaluation – Perceptual effectiveness of transfer features





- Pre-trained features are effective even without any prior perceptual training
- Our model with just 25% training data > state-of-the-art model with 100% training data

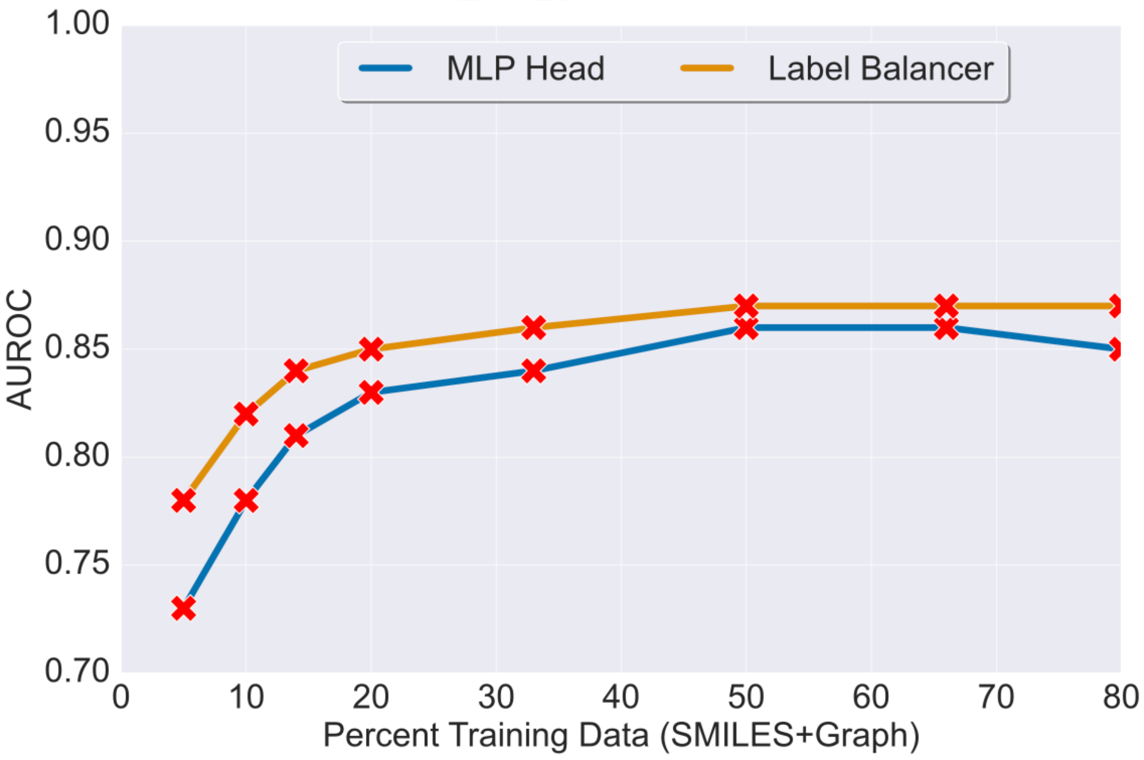
Evaluation – Perceptual effectiveness of multimodal features

Features	Multimodal	Test AUROC
SMILES (S) [53]	✗	0.71
Graph (G) [40]	✗	0.76
MORDRED (M) [31]	✗	0.80
$S \oplus G$	✓	0.81
$S \oplus M$	✓	0.83
$G \oplus M$	✓	0.84
$S \oplus G \oplus M$	✓	0.81
$S \odot G \odot M$	✓	0.84
$S \mid\mid G \mid\mid M$	✓	0.84

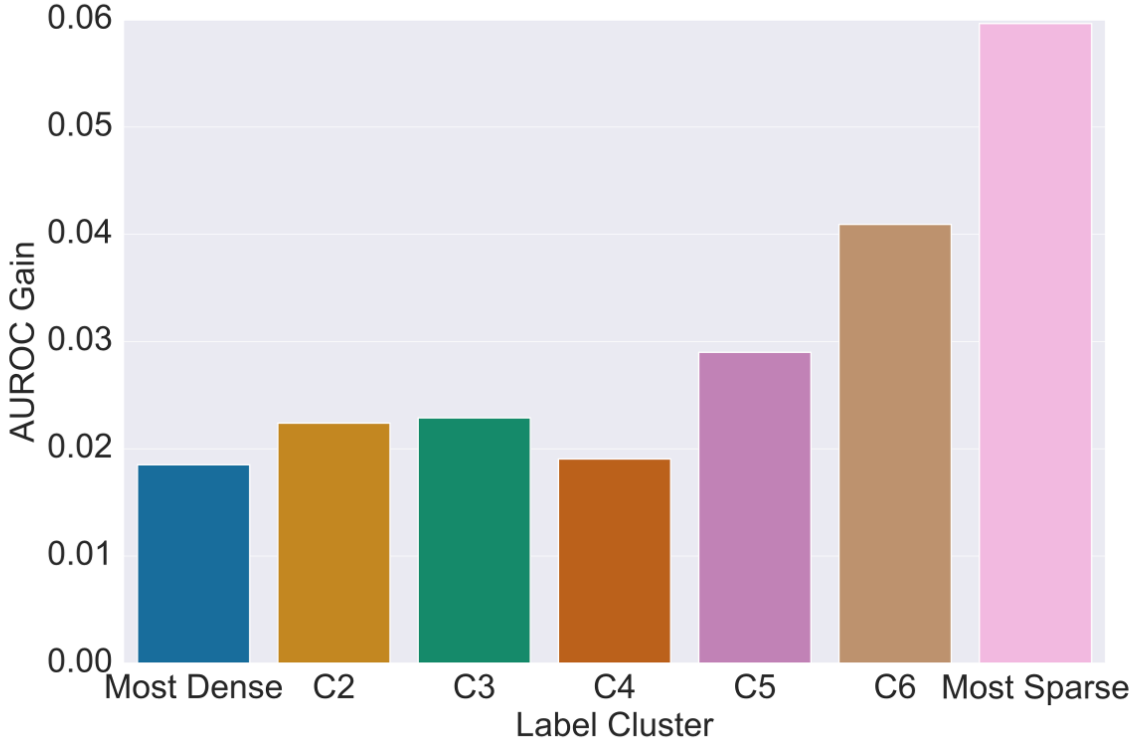
Performance comparison with and w/o multimodal learning

Limited enhancements from combining graph and text modalities due to insufficient complementary information.

Evaluation – Label-balancer training technique



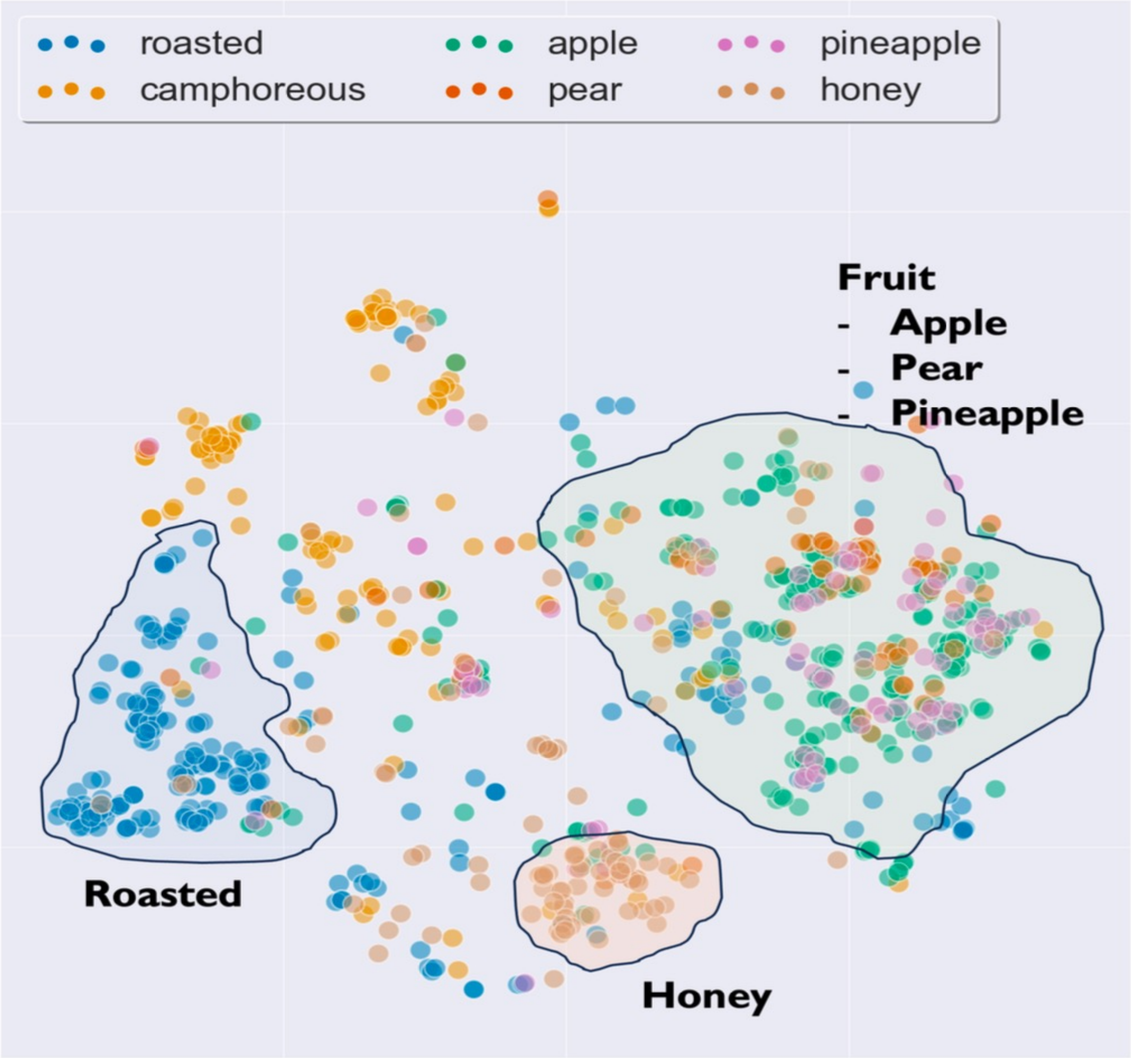
Performance comparison b/w classical multimodal fusion technique (MLP head) and label balancer



Performance gain by the label balancer over MLP head on most-dense to most-sparse classes

Label balancer consistently outperforms the MLP head across all training dataset sizes and yields higher gains for sparse classes than for dense ones

Evaluation – Learned embedding



- Clusters are diffused as each molecule is described by multiple labels
- Perceptually similar classes appear closer to each other than distinct ones

Future Directions

- Construct and assess molecular foundation model trained on tabular representation, incorporating chemical and physical properties

- Examine label-balancer efficacy across diverse modalities and their combination

- Explore novel approaches for robust and generalizable multilabel and multimodal training for modeling human smell perception