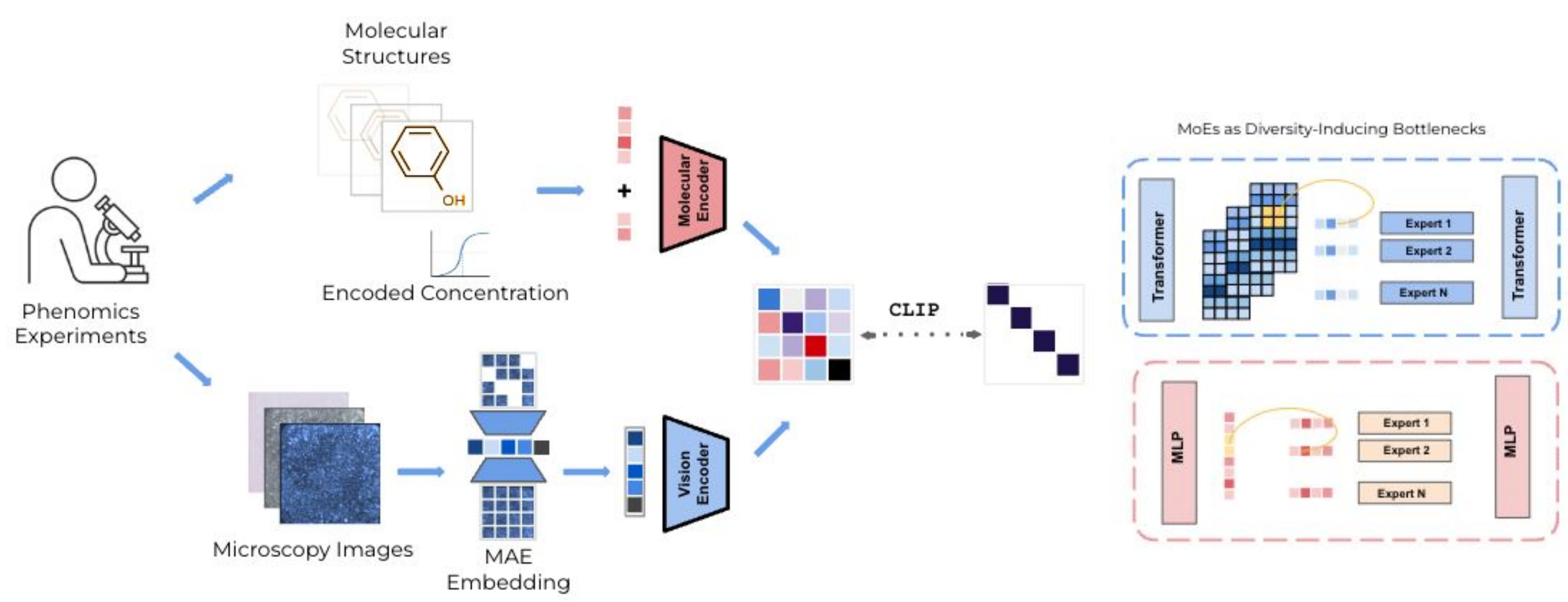
Incorporating MoEs in Contrastive Phenomic Retrieval

Karush Suri
Department of Electrical & Computer Engineering, University of Toronto

"MoEs benefit Contrastive Learning by inducing a higher level of diversity."

- *Contrastive Phenomic Retrieval* - Learning effects of molecules on cells using Contrastive Learning
- Addition of MoEs in encoders leads to diversity in visual features
- Induced diversity leads to 1.23x improvement over previous CLIP-based phenomic models
- Further validate improvement on downstream concentration prediction



Motivation

- MoEs are widely used for vision and multimodal learning
- A concrete understanding of their operation remains unclear
- Understand the benefits of MoEs in Contrastive Learning of phenomic representations
- Dissect the contributions of different components of MoEs

MoEs in Contrastive Phenomic Retrieval

- Contrastive learning of molecular representations and phenome lab experiments.
- Molecular fingerprints and microscopy images as input modalities.
- Maximize similarity in a joint embedding space using InfoNCE objective

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\langle \mathbf{z}_{\mathbf{x}_i}, \mathbf{z}_{\mathbf{m}_i}\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle \mathbf{z}_{\mathbf{x}_i}, \mathbf{z}_{\mathbf{m}_k}\rangle/\tau)}\right] + \left[\log\frac{\exp(\langle \mathbf{z}_{\mathbf{x}_i}, \mathbf{z}_{\mathbf{m}_i}\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle \mathbf{z}_{\mathbf{m}_i}, \mathbf{z}_{\mathbf{x}_k}\rangle/\tau)}\right].$$

- MoE modules to process visual features within encoder blocks
- Soft-MoE with a soft router combined with linear experts in the penultimate layer

$$\mathbf{A}_{ij} = \frac{\exp((\mathbf{Z\Phi})_{ij})}{\sum_{j'=1}^{s.p}\exp((\mathbf{Z\Phi})_{ij'})} \quad ; \quad \mathbf{Y} = \mathbf{A}\tilde{\mathbf{Y}}$$

Experiments

- Pretraining with contrastive learning to recall molecules
- Addition of MoEs induces diversity and is performant

Method	Without Soft MoEs					With Soft MoEs				
	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)
CLIP [12]	0.73±0.0003	0.80±0.0007	0.89±0.0004	0.15±0.00	5.62±0.001	**0.90±0.004**	**0.94±0.003**	**0.97±0.001**	**-0.02±0.03**	**4.73±0.007**
Hopfield-CLIP [47]	0.180±0.01	0.18±0.01	0.33±0.016	0.07±0.003	8.48±0.02	**0.24±0.013**	**0.25±0.015**	**0.43±0.02**	**0.014±0.009**	**8.45±0.02**
InfoLOOB [48]	0.492±0.009	**0.58±0.01**	0.740±0.009	0.56±0.03	7.46±0.06	**0.499±0.016**	0.57±0.02	**0.743±0.018**	**0.45±0.04**	**7.31±0.09**
NtXent [25]	0.27±0.01	0.30±0.02	0.46±0.02	0.51±0.02	8.76±0.15	**0.41±0.01**	**0.46±0.01**	**0.64±0.01**	**0.40±0.02**	**8.49±0.10**

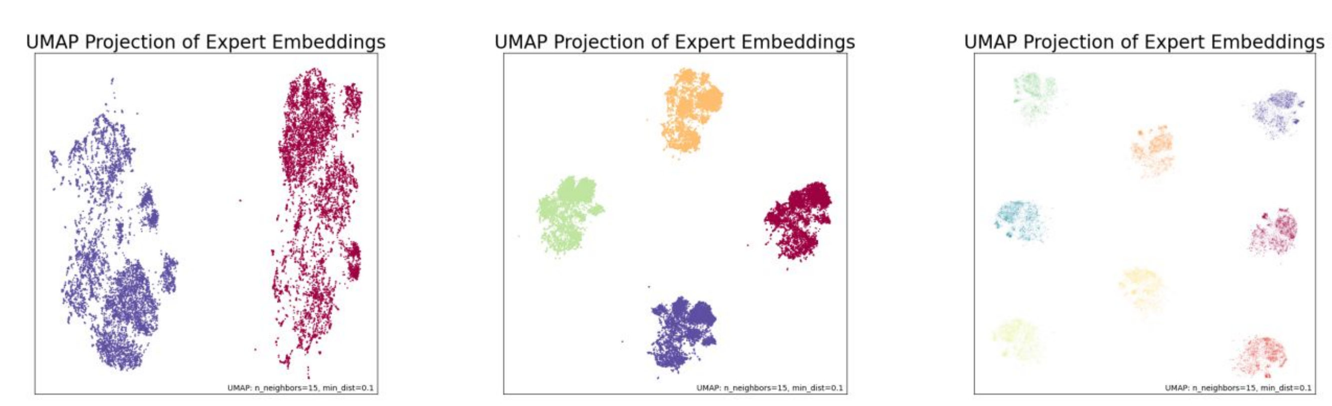
TABLE 1

- Downstream concentration prediction of heldout molecules
- Soft and Sparse MoEs are performant for high and low concentrations

Method	Concentration=1.0							Concentration=0.1						
	AUROC (↑)	Precision (↑)	Recall (↑)	Accuracy (↑)	RBF Similarity (↑)	Cosine Similarity (↓)	Training Error (↓)	AUROC (↑)	Precision (↑)	Recall (↑)	Accuracy (↑)	RBF Similarity (↑)	Cosine Similarity (↓)	Training Error (↓)
CLIP	0.62±0.19	0.59±0.16	0.58±0.15	0.58±0.15	0.2815	0.1579	0.71±0.18	0.97±0.04	0.96±0.09	0.96±0.09	0.95±0.08	0.2872	0.1552	0.19±0.16
CLIP+Soft MoEs	0.79±0.12	0.74±0.10	0.74±0.10	0.74±0.10	**0.6367**	**-0.0269**	0.58±0.20	0.98±0.25	**0.97±0.03**	**0.97±0.03**	**0.97±0.03**	**0.6366**	**-0.0265**	**0.13±0.13**
CLIP+Sparse MoEs	**0.81±0.11**	**0.76±0.09**	**0.76±0.09**	**0.76±0.09**	0.6094	-0.0243	**0.56±0.20**	**0.98±0.24**	**0.97±0.03**	**0.97±0.03**	**0.97±0.03**	0.6103	-0.0247	0.14±0.13

TABLE 2

- Experts learn diverse interpretable features of cells from microscopy-level visual input



Ablation Studies

- Which components of MoEs contribute towards their efficacy?
- Soft MoEs with larger number of slots per expert are more performant

Method	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)
CLIP	0.73±0.0003	0.80±0.0007	0.89±0.0004	0.15±0.00	5.62±0.001
CLIP+Dropout	0.67±0.00	0.73±0.0003	0.84±0.00	0.11±0.00	5.94±0.00
CLIP+Weight Decay	0.73±0.00	0.80±0.00	0.90±0.00	0.14±0.00	5.60±0.001
CLIP+Resets	0.38±0.11	0.41±0.12	0.56±0.13	0.03±0.003	7.76±0.58
CLIP+Pruning	0.069±0.007	0.006±0.006	0.12±0.011	0.07±0.06	25.14±11.46
CLIP+Soft Pruning	0.61±0.08	0.67±0.09	0.80±0.08	0.13±0.0008	6.56±0.87
Sparse CLIP	0.70±0.00	0.77±0.00	0.87±0.00	0.14±0.00	5.76±0.001
CLIP+Sparse MoE	0.88±0.006	0.93±0.004	0.96±0.003	0.003±0.003	4.95±0.08
CLIP+Soft MoE	**0.90±0.004**	**0.94±0.003**	**0.97±0.001**	**-0.02±0.03**	**4.73±0.007**

Slots	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)
2	0.83±0.10	0.87±0.11	0.90±0.11	-0.03±0.02	4.80±0.66
4	0.90±0.02	0.94±0.003	0.97±0.001	-0.01±0.001	4.62±0.07
8	0.72±0.12	0.76±0.13	0.80±0.13	**-0.03±0.019**	4.80±0.35
16	0.89±0.23	0.92±0.25	0.93±0.26	0.004±0.04	5.84±1.41
32	**0.906±0.004**	**0.946±0.003**	**0.97±0.001**	**-0.03±0.001**	**4.52±0.007**

- Moderately larger number of experts lead to more diversity
- Placing the MoE in penultimate layers of encoder block leads to stability

Experts	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)
2	0.90±0.002	0.94±0.002	0.97±0.001	0.01±0.001	4.59±0.04
4	0.90±0.002	0.94±0.002	0.97±0.001	-0.008±0.001	4.57±0.05
8	**0.906±0.002**	**0.946±0.002**	**0.976±0.001**	**-0.03±0.001**	**4.56±0.08**
16	**0.904±0.08**	**0.946±0.002**	**0.976±0.001**	-0.01±0.00	**4.56±0.03**

Placement	Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)	Cosine Similarity (↓)	Test Error (↓)
Input Layer	0.05±0.00	0.04±0.00	0.10±0.001	0.01±0.001	9.016±0.003
Projection Head	0.06±0.03	0.05±0.002	0.10±0.006	0.13±0.005	9.66±0.18
Encoder Block	**0.90±0.004**	**0.94±0.003**	**0.97±0.001**	**-0.02±0.03**	**4.73±0.007**