Incorporating MoEs in Contrastive Phenomic Retrieval

Karush Suri

Abstract—*Contrastive Phenomic Retrieval* is the problem of learning a joint embedding space between molecules and cell phenomes using multimodal contrastive learning. Conditional computation strategies such as Mixtures of Experts (MoEs) have been widely adopted for contrastive learning. However, their efficacy and versatility arising from subnetwork computations remain an unanswered question. *How Do MoEs benefit contrastive phenomic retrieval?* We empirically show that MoEs benefit contrastive learning by providing a higher level of diversity in encoder representations. Expert networks act as diversity-inducing bottlenecks which prevent encoder embeddings from becoming too similar. Our experiments demonstrate that embeddings learned by MoEs are diverse and continue to adapt, thereby resulting in $1.23\times$ improvement over previous CLIP-based phenomic models. We further study the effectiveness of MoEs in few-shot downstream tasks of concentration prediction and show that learning diverse representations with MoEs helps distinguish between downstream candidate concentrations downstream.

Index Terms—Contrastive Learning, AI for Science, Drug Discovery

1 Introduction

Learning effects of molecules on cellular compositions is a pivotal aspect of discovering novel therapeutic candidates [1], [2]. Cellular compositions encode phenotypical effects which hold the key to reversing the effects of a particular disease [3]. These phenotypical effects are learned by leveraging contrastive learning over multiple molecular modalities (such as molecular structures, fingerprints, microsocopy lab experiment images, etc.) [4], [5], [6]. Once learned, similar molecules may then be retrieved from the joint embedding space of molecules and cell phenomes either using similarity search or amortized inference [7]. Identifying and retrieving such molecules from the joint embedding space of multimodal contrastive learning models is the problem of *contrastive phenomic retrieval* [8], [9].

Akin to multimodal contrastive learning, recent advances in contrastive phenomic retrieval leverage conditional computation strategies to stabilize and improve retrieval performance. One such strategy is the use of Mixtures of Experts (MoEs) [10]. MoEs serve as parameter-efficient bottlenecks wherein embeddings are distributed across different networks (or experts) using a router network. Each expert learns representations corresponding to a particular aspect of the embedding, hence expertizing across a set of features. Expert embeddings are then aggregated to yield network outputs. In the case of language and vision, MoEs aid in constructing distributed representations such that each token has access to multiple experts. This results in improved scalability over increasing number of data samples [11].

While MoEs have demonstrated significant improvements in both performance and parameter efficiency, a concrete understanding of their utility and versatility remain an unanswered question. Models of today adopt MoEs as black-box architectures with little intuition about their operation. Empirically, it is unclear as to how MoEs benefit contrastive learning and what is the central factor aiding them to construct rich representations across a wide range of modalities.

We aim to answer the above questions by studying the utility of MoEs in contrastive phenomic retrieval. We begin by identifying a key phenomenon in CLIP-based objectives wherein the addition of MoEs in encoders improves diversity in visual features. MoEs serve as diversity-inducing bottlenecks which yield higher cosine similarities between both encoder embeddings. Additionally, we show that this induced diversity by MoEs results in an absolute $1.23\times$ improvement over previous CLIP-based phenomic models. We further evaluate the utility of diversity induced by MoEs on few-shot downstream tasks of concentration prediction and observe that visual representations learned by MoEs better transfer downstream.

2 Related Work

Multimodal Contrastive Learning: Prior multimodal models build on CLIP [12] and combine samples from two or more domains to learn expressive representations [13], [14]. Contrastive learning methods maximize similarity between paired samples traditionally in language-vision domains [15]. Recent works in multimodal contrastive learning aim at reducing compute and parameter budgets for data-efficient training [16]. [17] utilize unimodal pretrained models for one or both modalities improving zero-shot performance with an order of magnitude fewer paired samples. [18] further demonstrate that utilization of an elementwise sigmoid loss allows contrastive learners to scale under label noise. By using a unimodal pretrained model to learn similarities, [19] demonstrate improved performance on zero-shot retrieval. [20] extend the multimodal framework to molecular multi-omics for cancer subtyping. Our treatment of multimodal

• *K. Suri is with the Department of Electrical and Computer Engineering, University of Toronto, ON, Canada.*
E-mail: karush.suri@mail.utoronto.ca

contrastive learning models is parallel to the aforesaid directions.

Pretraining Molecule-Phenome Representations: Self-supervised methods have demonstrated pronounced success in vision, language and molecule representations [21], [22], [23]. In vision, these methods are used to minimize distance in the latent space of models between two views of the same sample [24], [25], [26], [27]. On the other hand, reconstruction-based objectives have also permeated representation learning such as masked autoencoders (MAE) [28]. MAEs allow for a higher degree of versatility by simulatenously masking and reconstructing segments of modality inputs [29]. These architectures typically utilize transformer architectures to partition the image into learnable tokens and reconstruct masked patches [30], [31], [32]. These methods have been extended to microscopy experimental data designed to capture cell morphology [33]. [34] utilize a masked autoencoder with a ViT-L/8+ architecture and a custom Fourier domain reconstruction loss, yielding informative representations of phenomic experiments [35]. Our work leverages analogous pretraining schemes.

Conditional Computation with MoEs: Recent advances in large model pretraining and contrastive learning employ MoEs [10]. These architectures serve as parameter-efficient alternatives which construct sets of distributed representations [36]. MoEs, being sparse gated units, only activate a portion of their parameters during training [37]. Recently, [38] make an effort to relax the sparsity constraints by proposing a soft MoE layer which adopts a weighted aggregation operation (eg- softmax) to yield output tokens. While contemporary models utilized MoEs for gaussian process regression and inference [39], recent methods have adopted these architectures for pretraining large foundational priors such as vision and language models [40]. [41] bootstrap continual MoE training from pretrained language embeddings to accelerate downstream language generalization. [42], on the other hand, incorporate MoEs in the multimodal setting by embedding vision and language via MoE encoders. [43] extend this idea further and embed audio, speech, vision and text in a joint latent space utilizing parallelized experts in MoEs. Finally, [44] utilize MoEs to learn interactions across different input modalities under constraints of partial labels. While the above works demonstrate scalable and performant use of MoEs, they fail to shed light on why these architectures benefit training. Current methods utilize experts as black-box architectures and provide little insights towards their operation. Among recent works, [45] attempt to explain MoEs by assessing their theoretical properties and non-linear expressivity of experts. MoE router learns cluster-center features which the experts can leverage to divide the problem into smaller sub-problems. Our empirical analysis in contrastive learning follows this line of work.

3 MoEs in Contrastive Phenomic Retrieval

Contrastive Phenomic Retrieval: We study the problem of learning multimodal representations of molecules and phenomic experiments of treated cells. Our setting considers a set of lab experiments $\mathcal{E}$ defined as the tuple $(\mathbf{X}, \mathbf{M}, \mathbf{C}, \mathbf{\Psi})$ wherein each experiment $e \in \mathcal{E}$ consists of cell images $\mathbf{x} \in \mathbf{X}$ and molecules as perturbations $\mathbf{m} \in \mathbf{M}$ obtained at specific dosage concentrations $\mathbf{c} \in \mathbf{C}$ with $\psi \in \mathbf{\Psi}$ denoting the threshold of molecular activity. Contrastive learning with CLIP utilizes an image encoder $f_{\theta_\mathbf{x}}$ with parameters $\theta_\mathbf{x}$ to generate the image embedding $\mathbf{z}_\mathbf{x} = f_{\theta_\mathbf{x}}(\mathbf{x})$, and a molecular encoder $f_{\theta_\mathbf{m}}$ with parameters $\theta_\mathbf{m}$ to generate the molecular embedding $\mathbf{z}_\mathbf{m} = f_{\theta_\mathbf{m}}(\mathbf{m}, \mathbf{c})$. Prior methods in multimodal contrasive learning utilize the InfoNCE loss to maximize the joint likelihood of $\mathbf{x}$ and $\mathbf{m}$. Given a set of $N \times N$ random samples $(\mathbf{x}_1, \mathbf{m}_1, \mathbf{c}_1), (\mathbf{x}_2, \mathbf{m}_2, \mathbf{c}_2), \ldots (\mathbf{x}_{N^2}, \mathbf{m}_{N^2}, \mathbf{c}_{N^2})$ containing N positive samples at k^{th} index and $(N-1) \times N$ negative samples, optimizing Equation 1 maximizes the likelihood of positive pairs while minimizing the likelihood of negative pairs.

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

Here, τ denotes the softmax temperature and $\langle \cdot \rangle$ denotes the cosine similarity.

Conditional Computation with MoEs: MoE architectures distribute embeddings across expert networks. Our study focusses on Soft MoEs [38] which utilize a series of *dispatch* and *aggregate* operations. We denote the input embedding $\mathbf{Z} \in \mathbb{R}^{b \times d}$, where b is the number of tokens and d is their dimension. Each MoE layer uses a set of s expert networks $(g_1, g_2, \ldots, g_s)$ applied on individual tokens. Each expert processes p slots, each consisting of a d-dimensional set of parameters, $\mathbf{\Phi} \in \mathbb{R}^{d \times (s.p)}$. The input slots $\tilde{\mathbf{Z}} \in \mathbb{R}^{(s.p) \times d}$ arise from a weighted sum of all b tokens as presented in Equation 2.

$$\mathbf{D}_{ij} = \frac{\exp((\mathbf{Z}\mathbf{\Phi})_{ij})}{\sum_{i'=1}^{b} \exp((\mathbf{Z}\mathbf{\Phi})_{i'j})} \quad ; \quad \tilde{\mathbf{Z}} = \mathbf{D}^{\mathrm{T}}\mathbf{Z} \quad (2)$$

Here, $\mathbf{D}$ are the dispatch weights. Upon application of expert functions on each slot, we obtain the output slots $\tilde{\mathbf{Y}}_i = g_i(\tilde{\mathbf{Z}}_i)$. Finally, a weighted sum across all $(s.p)$ slots results in the output tokens $\mathbf{Y}$ with $\mathbf{A}$ denoting the aggregate weights in Equation 3.

$$\mathbf{A}_{ij} = \frac{\exp((\mathbf{Z}\mathbf{\Phi})_{ij})}{\sum_{j'=1}^{s.p} \exp((\mathbf{Z}\mathbf{\Phi})_{ij'})} \quad ; \quad \mathbf{Y} = \mathbf{A}\tilde{\mathbf{Y}} \quad (3)$$

Experimental Setup: We begin by studying the behavior of CLIP during contrastive pretraining of phenome lab experiments. Figure 1 (left) presents our setting wherein we consider three input modalities, (1) microscopy images, (2) molecule structure fingerprints and (3) concentrations as one-hot encoded context. Our empirical analysis utilizes biological phenome maps consisting of raw cellular microscopy images [46]. Since raw microscopy images have an order of magnitude more dimensions than molecule fingerprints, we additionally pretrain a unimodal MAE to extract compact latent representations of visual inputs. Both encoders are trained simultaneously using the CLIP objective [12].

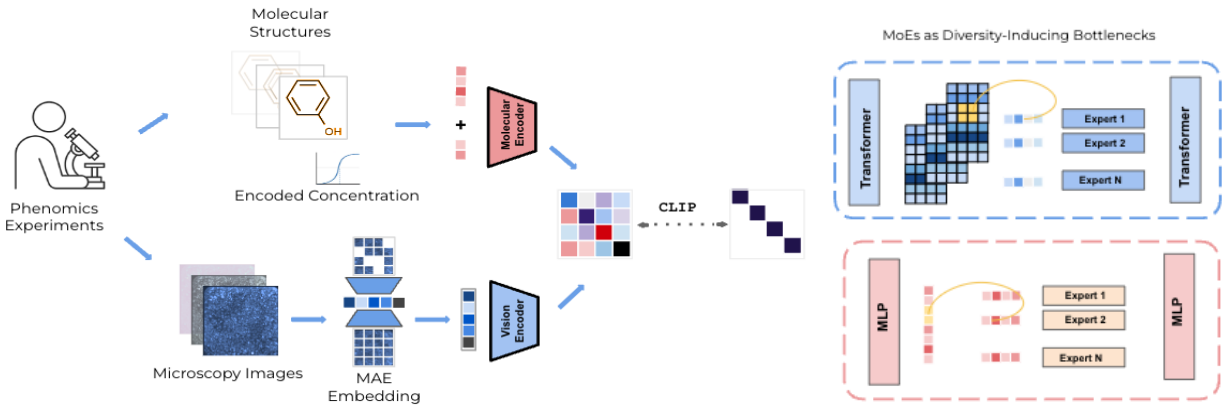

Fig. 1. **(left)** Illustration of contrastive phenomic retrieval wherein we use microsocopy images, molecule structure fingerprints and one-hot encoded concentrations as input modalities. CLIP learns a joint embedding space of phenome lab experiments using a vision encoder and a molecular encoder. **(right)** Implementation of Soft MoEs in different encoder architectures. We combine MoEs with MLPs and Transformers where molecule features act as tokens.

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
Molecular retrieval performance of different contrastive learning objectives. CLIP-based objectives result in similar visual features. These methods require the use of diversity-inducing strategies such as MoEs. Bold entries denote highest values across corresponding columns while shaded entries denote highest values across the table.

4 Experiments

Does lack of feature diversity affect other CLIP-based objectives? We answer this question by comparing retrieval performance of recent CLIP-based objectives. In Table 1 we observe that CLIP-based objectives, in the absence of MoEs, have high cosine similarities. Specifically, InfoLOOB [48] and NtXent [25] have approximately $3\times$ higher similarities compared to CLIP. Furthermore, all the presented objectives explicitly require the use of MoEs to reduce similarity overfitting and boost performance. Thus, high cosine similarities of CLIP-based methods combined with their reliance on MoE architectures indicate that lack of diversity plagues recent methods as well that aim to adopt or improve over CLIP.

Our analysis aims to answer two main questions; (1) Does the addition of representation diversity benefit performance?, (2) Which components of MoEs benefit diversity and performance? Figure 2 (top) answers our first question by comparing CLIP variants with the addition of Soft MoEs. In our comparison, we include common design decisions which are required to stabilize contrastive learning (such as dropout, weight decay and sparsity). Additionally, we include methods which benefit unsupervised learning models in general (such as periodic resetting of parameters and pruning network weights). Across the board, we note that CLIP with the addition of soft MoEs outperforms prior design decisions leading to $2-17\%$ improvements over baselines. Furthermore, MoE-based methods induce the most diversity as indicated by lowest cosine similarity values between the encoder embeddings.

While MoEs demonstrate improvements in performance and scalability, it is unclear as to which components of these architectures contribute towards their benefits. We answer our second question by ablating between different MoE architectures. We begin by varying the number of input slots for each expert while keeping the number of experts fixed. Figure 2 (bottom) presents the variation of retrieval performance for varying number of input slots. While MoEs with smaller slots per experts induce diversity, these fall short of boosting molecular retrieval performance. This arises as a lack of expressivity within the model architecture. MoEs with larger number of slots better induce diversity and boost performance with 32 slots per expert providing a 7.6% improvement over the smallest architecture. Figure 3 (top) extends our analysis to varying number of experts while keeping the slots per expert fixed. More experts per encoder benefit diversity the most. Empirically, both encoders consisting of 8-16 experts are found to be stable with the most representation diversity arising from distributed representations of each expert.

In Figure 3 (bottom) we answer the key question of *where should the Soft MoE layer be placed?* Traditional language architectures place MoEs in the penultimate layers of the model. We see that phenomic models benefit the most in performance and diversity when MoEs are placed within the encoder blocks. Placing MoEs in the input layer leads to instability wherein encoders fail to capture representations propagating from different expert heads. On the other hand, placing MoEs in the projection head leads to coarser embeddings wherein outputs are too similar. Encoder blocks serve as the right balance between stability and diversity by accumulating and distributing embeddings between layers.

Few-Shot Concentration Prediction: Predicting the concentration of molecular perturbations is a key ingredient

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**

Fig. 2. **(left)** Comparison of molecular retrieval performance for CLIP-based models. Across different design choices, MoEs consistently demonstrate improved recall performance. **(right)** Retrieval performance for varying number of slots per expert. More slots provide expressivity to experts.

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**

Fig. 3. Variation of molecular retrieval performance with **(left)** varying number of experts and **(right)** different placements of MoE layer. Larger experts placed in encoder blocks are found to benefit retrieval performance.

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
Few-shot concentration prediction performance of downstream classifiers trained with embeddings of CLIP, Soft MoEs and Sparse MoEs.

in evaluating the chemical validity of CLIP embeddings. We consider the task of concentration prediction, wherein we train a logistic binary classifier in few shots to classify whether a molecule belongs to a concentration set or not. Our classifiers are trained from frozen embeddings of both molecular and vision encoders. In order to prevent passing ground-truth labels, we mask out true concentrations as inputs to the molecular encoder. Since concentrations exhibit an inherent discrete sparse structure, an ideal embedding would be the one that captures correlations in as less as a few bits.

Table 2 presents our results for few-shot concentration prediction using embeddings of CLIP-like models. While MoE models present improved diversity and downstream classification performance, sparse MoEs are found to be more performant. Compared to CLIP and sparse MoEs, addition of soft MoEs presents the most diverse embeddings. Compared to CLIP and soft MoEs, addition of sparse MoEs presents diverse as well as performant embeddings.

5 Conclusion

We studied the utility of MoEs through the lens of contrastive phenomic retrieval. By utilizing lab experiment images and molecular structure fingerprints as modalities, we learned a joint embedding space over molecules and cell phenomes. We observed that CLIP-based objectives are prone to a lack of feature diversity. MoEs address this key phenomenon in training of CLIP by providing a higher-level of diversity in encoder representations. Our experiments empirically validated the efficacy of MoEs in induing diversity and improving retrieval performance by $1.23\times$ over previous CLIP methods. Additionally, we showed that MoE representations demonstrate a positive transfer on few-shot and zero-shot downstream task of concentration prediction.

Limitations and Future Work: While our setting studies molecule-cellular interactions, we reserve two aspects for future work. Firstly, our setting can be extended to accommodate additional modalities such as molecule structures. These would inform the model of the geometry of each molecule. Lastly, MoE representations can be transferred to additional downstream tasks which require a higher degree of versatility such as protein binding and target rescue.

References

[1] P. Akarapipad, K. Kaarj, Y. Liang, and J.-Y. Yoon, "Environmental toxicology assays using organ-on-chip," *Annual Review of Analytical Chemistry*, vol. 14, pp. 155–183, 2021.

[2] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: A molecular modeling perspective," *Medicinal Research Reviews*, vol. 16, no. 1, p. 3–50, Jan. 1996.

[3] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto, and M. Mercola, "Phenotypic drug discovery: recent successes, lessons learned and new directions," *Nature Reviews Drug Discovery*, vol. 21, no. 12, pp. 899–914, 2022.

[4] J. Simm, G. Klambauer, A. Arany, M. Steijaert, J. K. Wegner, E. Gustin, V. Chupakhin, Y. T. Chong, J. Vialard, P. Buijnsters *et al.*, "Repurposing high-throughput image assays enables biological activity prediction for drug discovery," *Cell chemical biology*, vol. 25, no. 5, pp. 611–618, 2018.

[5] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter, "Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes," *Nature protocols*, vol. 11, no. 9, pp. 1757–1774, 2016.

[6] M. Hofmarcher, E. Rumetshofer, D.-A. Clevert, S. Hochreiter, and G. Klambauer, "Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks," *Journal of chemical information and modeling*, vol. 59, no. 3, pp. 1163–1171, 2019.

[7] P. Fradkin, P. Azadi, K. Suri, F. Wenkel, A. Bashashati, M. Sypetkowski, and D. Beaini, "How molecules impact cells: Unlocking contrastive phenomolecular retrieval," *Advances in Neural Information Processing Systems*, 2024.

[8] A. Sanchez-Fernandez, E. Rumetshofer, S. Hochreiter, and G. Klambauer, "Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures," *Nature*, 2023.

[9] C. Q. Nguyen, D. Pertusi, and K. M. Branson, "Molecule-morphology contrastive pretraining for transferable molecular representation," *bioRxiv*, pp. 2023–05, 2023.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[11] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[12] 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*. PMLR, 2021, pp. 8748–8763.

[13] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," 2022.

[14] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, "Language is not all you need: Aligning perception with language models," 2023.

[15] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[16] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International conference on machine learning*. PMLR, 2020, pp. 4182–4192.

[17] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," 2022.

[18] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[19] R. S. Srinivasa, J. Cho, C. Yang, Y. M. Saidutta, C.-H. Lee, Y. Shen, and H. Jin, "Cwcl: Cross-modal transfer with continuously weighted contrastive loss," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[20] Z. Yang, Z. Chen, Y. Matsubara, and Y. Sakurai, "Moclim: Towards accurate cancer subtyping via multi-omics contrastive learning with omics-inference modeling," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 2895–2905.

[21] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," 2023.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[23] S. Zaidi, M. Schaarschmidt, J. Martens, H. Kim, Y. W. Teh, A. Sanchez-Gonzalez, P. Battaglia, R. Pascanu, and J. Godwin, "Pre-training via denoising for molecular property prediction," 2022.

[24] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[25] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 1857–1865.

[26] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020.

[27] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[29] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1332–1344.

[30] C. Feichtenhofer, Y. Li, K. He *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35 946–35 958, 2022.

[31] S. Cao, P. Xu, and D. A. Clifton, "How to understand masked autoencoders," *arXiv preprint arXiv:2202.03670*, 2022.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] R. Xie, K. Pang, G. D. Bader, and B. Wang, "Maester: Masked autoencoder guided segmentation at pixel resolution for accurate, self-supervised subcellular structure recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023.

[34] O. Kraus, K. Kenyon-Dean, S. Saberian, M. Fallah, P. McLean, J. Leung, V. Sharma, A. Khan, J. Balakrishnan, S. Celik, D. Beaini, M. Sypetkowski, C. V. Cheng, K. Morse, M. Makes, B. Mabey, and B. Earnshaw, "Masked autoencoders for microscopy are scalable learners of cellular biology," 2024.

[35] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," in *International Conference on Machine Learning*. PMLR, 2023, pp. 7480–7512.

[36] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts," *arXiv preprint arXiv:2407.06204*, 2024.

[37] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[38] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," *arXiv preprint arXiv:2308.00951*, 2023.

[39] J. W. Ng and M. P. Deisenroth, "Hierarchical mixture-of-experts model for large-scale gaussian process regression," *arXiv preprint arXiv:1412.3078*, 2014.

[40] K. M. Lo, Z. Huang, Z. Qiu, Z. Wang, and J. Fu, "A closer look into mixture-of-experts in large language models," *arXiv preprint arXiv:2406.18219*, 2024.

[41] T. Zhu, X. Qu, D. Dong, J. Ruan, J. Tong, C. He, and Y. Cheng, "Llama-moe: Building mixture-of-experts from llama with continual pre-training," *arXiv preprint arXiv:2406.16554*, 2024.

[42] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *arXiv preprint arXiv:2111.02358*, 2021.

[43] Y. Li, S. Jiang, B. Hu, L. Wang, W. Zhong, W. Luo, L. Ma, and M. Zhang, "Uni-moe: Scaling unified multimodal llms with mixture of experts," *arXiv preprint arXiv:2405.11273*, 2024.

[44] H. Yu, P. P. Liang, R. Salakhutdinov, and L.-P. Morency, "Mmoe: Mixture of multimodal interaction experts," *arXiv preprint arXiv:2311.09580*, 2023.

[45] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding mixture of experts in deep learning," *arXiv preprint arXiv:2208.02813*, 2022.

[46] M. M. Fay, O. Kraus, M. Victors, L. Arumugam, K. Vuggumudi, J. Urbanik, K. Hansen, S. Celik, N. Cernek, G. Jagannathan *et al.*, "Rxrx3: Phenomics map of biology," *Biorxiv*, pp. 2023–02, 2023.

[47] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve *et al.*, "Hopfield networks is all you need," *arXiv preprint arXiv:2008.02217*, 2020.

[48] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5171–5180.

[49] G. Landrum *et al.*, "Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling," *Greg Landrum*, vol. 8, no. 31.10, p. 5281, 2013.

6 Appendix

6.1 Pretraining with Representation Diversity

Our pretraining strategy closely follows the setup of standard CLIP training. In the case of smaller models, we only utilized 1 encoder block with tuned learning rates. Explicit learning rate schedules were not used. For larger models exceeding 250M parameters, we use custom learning rate schedules to kickstart training. A common strategy we

utilize for 250M and 1B parameter models is the use of warm starts learning rate. Learning rate is set to 1e-6 and increased by a factor of 1.01 every 100th step for the first 1000 steps.

We experimented with two warms starts strategies. In the first case, which we denote as fast increment, we increase the learning rate every 100th step. In the second case, which we call slow increment, we increase the learning rate every 500th step. We empirically found fast increment to work well and stabilize CLIP training.

Similar to warm starts learning, we decay learning rates for larger model sizes. Following 30,000 steps of training, learning rates for 250M and 1B parameter models are decayed by a factor of 0.99 at intervals of every 100 steps.

In the case of encoder architectures, we found wider encoders to work best. Increasing the depth of MLP and Transformer encoders also yielded favourable results. In the case of MLP encoders, we tuned the number of blocks between 2, 4, 6 and 8 with 8 being the best configuration. For Transformer models, we tuned the number of encoder blocks between 2, 4, 8, 12, 16 and 24 with 24 being the best configuration. As for width, we tuned between 256, 512, 1024, 2048 and 4096. In the case of MLP models, 4096 was found to work the best whereas for transformer models we were only able to accomodate 2048 due to the large memory requirements posed by attention layers.

For all architectures, weight decay was found to benefit training while dropout had little to no effect. In the case of MoEs, we kept the number of experts fixed to 8 and the number of slots per expert fixed to 32. In general, larger number of slots per expert were found to benefit training.

6.2 Few-Shot Concentration Prediction

Our few-shot concentration prediction experiment utilized the 250M parameter model for probing. We freeze both the encoders and add a linear layer along with sigmoid nonlinearity to predict the threshold of concentration. Since we are predicting the concentration modalities themselves, we blocked the one-hot encoded context input and only provided an empty mask in its place. This allowed us to fairly evaluate our model's latent space of cell concentration information. Thus, we only provided the image and molecule fingerprint embedding to our downstream classification head.

The task, thus being of binary prediction, only required a handful of samples. In the case of concentration = 1.0, we utilized a total of 300 shots. For concentration = 0.1, only 20 shots were utilized. We specifically chose these concentrations as these possessed the most number of noisy embeddings, noisy molecule samples and active molecule perturbations. While both CLIP and CLIP+Soft MoE models were stable, the addition of MoEs accelerated downstream learning across both concentrations.

6.3 Dataset Details

We note that our pretraining data distribution consists of 1.6M paired samples containing 205456 unique molecules spanning a total of 14 unique concentrations. While the image encoder utilizes embeddings of raw microsocopy images as input, the molecular encoder utilizes fingerprints generated from valid molecule SMILES as input. We utilize the ECFP fingerprinting technique [49] in order to generate and cache molecule fingerprints. Our input to the molecular encoder consists of MACCS keys and the hashed MORGAN fingerprinting with an atomic radius of 2. When combined together, the input of size 2215 dimensions is passed to the molecular encoder.

Our pretrained MAE encoder downscales each image into an embedding of 768 dimensions which are fed as input to the CLIP image encoders. In addition to the image embedding, we encode the concentration of each phenome experiment and concatenate it with the input embedding. Since we have 13 unique concentrations out of which 8 are the most frequent ones, we utilize one-hot encoding as our encoding strategy to generate the concentration embedding. Other encoding choices, such as log and sigmoid, were also tried but one-hot was empirically found to be more informative.

Both image and molecule encoders utilize the same architecture consisting of 2-4 blocks of MLP/Transformer layers with GeLU activations and skip connections. Each encoder is followed by a LayerNorm projection which is passed to the projection head. Similar to the original CLIP architecture, the projection head acts as a dimensionality-reduction bottleneck wherein the final output embedding is projected and normalized before computation of similarity matrix. Output embeddings of encoders are of size 1024-2048 dimensions wherein larger dimensions are found to be more performant.

To make our experiments reproducible, we setup our code base on the open-source RXRX3 dataset[1], the largest open-sourced dataset consisting of biological phenome maps [46]. The dataset consists of 2.2M images from the HUVEC cell line. The dataset additionally consists of 1674 known chemical entities across 8 different unique concentrations each. Each image is of dimension $2048 \times 2048 \times 6$ consisting of 6 channels corresponding to experiment stains Hoechst, ConA, Phalloidin, Syto14, MitoTracker and WGA. Each phenome experiment consists of 1536-well plate density containing 1 imaging site per well.

1. https://rxrx3.rxrx.ai/downloads