Molphenix: A Multimodal Foundation Model for PhenoMolecular Retrieval

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predicting molecular impact on cellular function is a core challenge in therapeutic design. Phenomic experiments, designed to capture cellular morphology, utilize microscopy based techniques and demonstrate a high throughput solution for uncovering molecular impact on the cell. In this work, we learn a joint latent space between molecular structures and microscopy phenomic experiments, aligning paired samples with contrastive learning. Specifically, we study the problem of *Contrastive PhenoMolecular Retrieval*, which consists of zero-shot molecular structure identification conditioned on phenomic experiments. We assess challenges in multi-modal learning of phenomics and molecular modalities such as experimental batch effect, inactive molecule perturbations, and encoding perturbation concentration. We demonstrate improved multi-modal learner retrieval through (1) a uni-modal pre-trained phenomics model, (2) a novel inter sample similarity aware loss, and (3) models conditioned on a representation of molecular concentration. Following this recipe, we propose *MolPhenix*, a molecular phenomics model. MolPhenix leverages a pre-trained phenomics model to demonstrate significant performance gains across perturbation concentrations, molecular scaffolds, and activity thresholds. In particular, we demonstrate an $8.1\times$ improvement in zero shot molecular retrieval of active molecules over the previous state-of-the-art, reaching 77.33% in top-1% accuracy. These results open the door for machine learning to be applied in virtual phenomics screening, which can significantly benefit drug discovery applications.

1 Introduction

Quantifying cellular responses elicited by genetic and molecular perturbations represents a core challenge in medicinal research [4, 48]. Out of an approximate 10^{60} druglike molecule designs, a small number are able to alter cellular properties to reverse the course of diseases [5, 22]. In recent years, microscopy-based cell morphology screening techniques, demonstrated potential for quantitative understanding of a molecule's biological effects. Experimental techniques such as cell-painting are used to capture cellular morphology, which correspond to physical and structural properties of the cell [6, 7]. Cells treated with molecular perturbations can change morphology, which is captured by staining and high throughput microscopy techniques. Perturbations with similar cellular impact induce analogous morphological changes, allowing to capture underlying biological effects in phenomic experiments. Identifying such perturbations with similar morphological changes can aid in discovery of novel therapeutic drug candidates [42, 24, 19].

Determining molecular impact on the cell can be formulated as a multi-modal learning problem, allowing us to build on a rich family of methods [35, 52, 45]. Similar to text-image models, paired data is collected from phenomic experiments along with molecules used to perturb the cells.

Submitted to Foundation Models for Science Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Contrastive objectives have been used as an effective approach in aligning paired samples from different modalities [35, 27]. A model that has learned a cross-modal joint latent space must be able to retrieve a molecular perturbant conditioned on the phenomic experiment. We identify this problem as *contrastive phenomolecular retrieval* (see Figure 2). Addressing this problem can allow for identification of molecular impact on cellular function, however, this comes with its own set of challenges. [15, 2, 54].

(1) Firstly, multi-modal paired phenomics molecular data suffers from lower overall dataset sizes and is subject to batch effects. Challenges with uniform processing and prohibitive costs associated with acquisition of paired data, leads to an order of magnitude fewer data points compared to text-image datasets [41, 9]. Furthermore, data is subject to random batch effects that capture non-biologically meaningful variation [28, 46]. **(2)** Paired phenomic-molecular data contains inactive perturbations that do not have a biological effect or do not perturb cellular morphology. It is difficult to infer a priori whether a molecule has a cellular effect, leading to the collection of paired molecular structures with unperturbed cells. These data-points are challenging to filter out without an effective phenomic embedding, as morphological effects are rarely discernible. These samples can be interpreted as misannotated, under the assumption of all collected pairs having biologically meaningful interactions. **(3)** Finally, a complete solution for capturing molecular effects on cells must capture molecular concentration. The same molecule can have drastically different effects along its dose response curve, thus making concentration an essential component for learning molecular impact.

In this work, we explore the problem of contrastive phenomolecular retrieval by addressing the above challenges circumvented in prior works. Our key contributions are as follows:

- We demonstrate significantly higher phenomolecular retrieval rates by utilizing a pretrained uni-modal phenomic encoder. Thus alleviating the data availability challenge, reducing the impact of batch effects, and identifying molecular activity levels.
- We propose a novel soft-weighted sigmoid locked loss (S2L) that addresses the effects of inactive molecules. This is done by leveraging distances computed in the phenomic embedding space to learn inter-sample similarities.
- We explore *explicit* and *implicit* methods to encode molecular concentration, assessing the model's ability to perform retrieval in an inter-concentration setting and generalize to unseen concentrations.

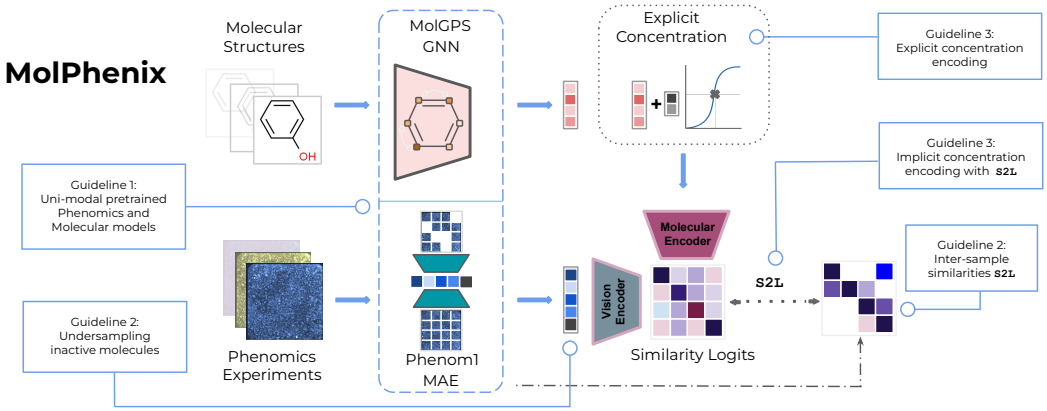


Figure 1: Illustration of proposed guidelines when incorporated in our *MolPhenix* contrastive phenomolecular retrieval framework. We address challenges by utilizing uni-modal pretrained MAE & MPNN models, inter-sample weighting with a dosage aware S2L loss, undersampling inactive molecules, and encoding molecular concentration.

Following these principles, we build *MolPhenix*, a multi-modal molecular phenomics model addressing contrastive phenomolecular retrieval (Figure 1). MolPhenix demonstrates large and consistent improvements in the presence of batch effects, generalizing across different concentrations, molecules, and activity thresholds. Additionally, MolPhenix outperforms baseline methods in zero-shot setting, achieving 77.33% top-1% retrieval accuracies on active molecules, which corresponds to a **8.1**× improvement over the previous state-of-the-art (SOTA) [40].

2 Related Work

Uni-modality Pretraining: Self-supervised methods have demonstrated success across a variety of domains such as computer vision, natural language processing and molecular representations [3, 36, 51]. In vision, contrastive methods have been used to minimize distance in the model's latent space of two views of the same sample [10, 43, 16, 18]. Reconstruction objectives have also permeated computer vision, such as masked autoencoders (MAE). MAEs typically utilize vision transformers to partition the image into learnable tokens and reconstruct masked patches [17, 14, 8, 12]. These methods have been extended to microscopy experimental data designed to capture cell morphology [50, 23]. Phenom1 utilizes a masked autoencoder with a ViT-L/8+ architecture and a custom Fourier domain reconstruction loss, yielding informative representations of phenomic experiments [23, 11]. From a representational perspective, Graph Neural Networks (GNN) have been used to predict molecular properties by reasoning over graph structures. A combination of reconstruction and supervised objectives have led to models generalizing to a diverse range of prediction tasks [31, 55, 47, 39]. Our work leverages uni-modal foundation models, which are used to generate embeddings of phenomic images and molecular graphs.

Multi-Modal Objectives: Multi-modal models combine samples from two or more domains, to learn rich representations and demonstrate flexible ways to predict sample properties [35, 1, 20]. Contrastive methods minimize distances between paired samples, traditionally in text-image domains. However, training these models is computationally expensive, requiring large datasets. Multiple contributions have allowed for a reduction in compute and data budgets by an order of magnitude. In *LiT*, the authors demonstrate that utilizing uni-modal pretrained models for one or both modalities matches zero-shot performance with an order of magnitude fewer paired examples seen [53]. Zhai et al. (2023) demonstrate that by replacing the softmax operation over cosine similarities with an element wise sigmoid loss, allows contrastive learners to improve performance under label noise regime [52]. By using a uni-modal pre-trained modal to calculate similarities between samples from one of the modalities, Srinivasa et al. (2023) have demonstrated improved performance on zero-shot evaluation [45]. In our work, we build along these directions in molecular phenomic multi-modal training.

Molecular-Phenomic Contrastive Learning: Prior works in contrastive phenomic retrieval have utilized the InfoNCE objective as a pre-training technique to construct uni-modal representations [32]. Recent methods have attempted to improve retrieval by using the InfoLOOB objective [34]. Specifically, CLOOME utilizes the InfoLOOB loss with hopfield networks for zero-shot retrieval on unseen data samples [37, 40]. Our work is parallel to the above directions, demonstrating a significant increase in molecular-phenomic retrieval by building on algorithmic improvements from the multi-modality literature.

3 Methodology

In this section, we explain key challenges facing phenomolecular retrieval and provide guidelines that are key methodological improvements behind the success of MolPhenix 1.

Preliminaries: Our setting studies the problem of learning multi-modal representations of molecules and phenomic experiments of treated cells [40]. The aim of this work is to learn a joint latent space which maps phenomic experiments of treated cells and the corresponding molecular perturbations into the same latent space. We consider a set of lab experiments $\mathcal{E}$ defined as the tuple $(\mathbf{X}, \mathbf{M}, \mathbf{C}, \mathbf{\Psi})$. Each experiment $\epsilon \in \mathcal{E}$ consists of data samples $\mathbf{x}_i \in \mathbf{X}$ (such as images) and perturbations $\mathbf{m}_i \in \mathbf{M}$ (such as molecules) which are obtained at a specific dosage concentration $\mathbf{c}_i \in \mathbf{C}$, while $\psi \in \mathbf{\Psi}$ denotes molecular activity threshold.

Figure 2 describes the problem of contrastive phenomolecular retrieval, where for a single image $\mathbf{x}_i$, the challenge consists of identifying the matching perturbation, $\mathbf{m}_i$, and concentration, $\mathbf{c}_i$, used to induce morphological effects. This can be accomplished in a zero-shot way by generating embeddings for $(\mathbf{m}_1, \mathbf{c}_1), ...(\mathbf{m}_j, \mathbf{c}_j)$ and $\mathbf{x}_i$ using functions $f_{\theta_m}(\mathbf{m}, \mathbf{c})$, $f_{\theta_x}(\mathbf{x})$ which map samples into $\mathbb{R}^d$.

Then, by defining a similarity metric between generated embeddings $\mathbf{z}_{x_i}$ and $\mathbf{z}_{m_i}$, f_{sim}, we can rank $(\mathbf{m}_1, \mathbf{c}_1)...(\mathbf{m}_j, \mathbf{c}_j)$ based on computed similarities. An effective solution to the contrastive phenomolecular retrieval problem would learn $f_{\theta_m}(\mathbf{m}, \mathbf{c})$ and $f_{\theta_x}(\mathbf{x})$ that results in consistently high retrieval rates of $(\mathbf{m}_i, \mathbf{c}_i)$ used to perturb $\mathbf{x}_i$.

In practice, the image embeddings are generated using a phenomics microscopy foundation MAE model [23, 17]. We use phenomic embeddings to marginalize batch effects, infer inter-sample similarities, and undersample inactive molecules. Activity is determined using consistency of replicate measurements for a given perturbation. For each sample, a p value cutoff $\psi \in \mathbf{\Psi}$ is used to quantify molecular activity. Only molecules below the p value cutoff ψ are considered active.

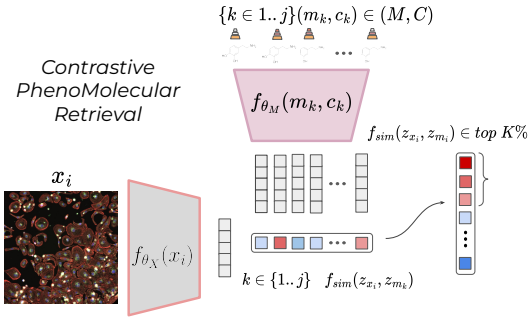


Figure 2: Illustration of the contrastive phenomolecular retrieval challenge. Image $\mathbf{x}_i$ and a set of molecules and corresponding concentrations $(\mathbf{m}_k, \mathbf{c}_k)$ get mapped into a $\mathbb{R}^d$ latent space. Their similarities get computed with f_{sim} and ranked to evaluate whether the paired perturbation appears in the top K%.

Prior methods in multi-modal contrastive learning utilize the InfoNCE loss, and variants thereof [32] to maximize the joint likelihood of $\mathbf{x}_i$ and $\mathbf{m}_i$. Given a set of $N \times N$ random samples $(\mathbf{x}_1, \mathbf{m}_1, \mathbf{c}_1), \cdots, (\mathbf{x}_N, \mathbf{m}_N, \mathbf{c}_N)$ 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:

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

Where $\mathbf{z}_x$, $\mathbf{z}_m$ correspond to phenomics and molecular embeddings respectively, τ is softmax temperature, and $\langle \cdot \rangle$ corresponds to cosine similarity.

Challenge 1: Phenomic Pretraining and Generalization

We find that using a phenomics foundation model to embed microscopy images allows for mitigation of batch effects, reduces the required number of paired data points, and improves generalization in the process. While `CLIP`, a hallmark model in the field of text-image multi-modality, was trained on 400 million curated paired data points, there is an order of magnitude fewer paired molecular-phenomic molecule samples [35]. Cost and systematic pre-processing of data make large scale data generation efforts challenging, and resulting data is affected by experimental batch effects. **Batch effects** induce noise in the latent space as a result of random perturbations in the experimental process, while biologically meaningful variation remains unchanged [33, 44]. Limited dataset sizes and batch effects make it challenging for contrastive learners to capture molecular features affecting cell morphology, yielding low retrieval rates [40].

We address data availability and generalization challenges by utilizing representations from a large **uni-modal pre-trained phenomic model**, θ_{Ph}, trained to capture representations of cellular morphology. θ_{Ph} is pretrained on microscopy images using a Fourier modified MAE objective, utilizing the ViT-L/8 architecture with methodology similar to Kraus et al. (2024) [17, 12, 23]. For simplicity in future sections, we refer to this model as *Phenom1*. This pretrained model allows a drastic reduction in the required number of paired multi-modal samples [53]. In addition, using phenomic representations alleviates the challenge of batch effects by averaging samples, $\mathbf{z}_x$, generated with the same perturbation $\mathbf{m}_i$ over multiple lab experiments ϵ_i. Averaging model representations $\frac{1}{N}\Sigma_{i \in N}^{1} \mathbf{z}_{x_i}$ allows marginalizing batch effect induced by individual experiments.

> **Guideline 1** *Utilizing pre-trained uni-modal encoder, θ_{Ph}, can be used to reduce the number of paired data-points compared to training θ without prior optimization. In addition, averaging phenomic embeddings $\mathbf{z}_x$ from matched perturbations can alleviate batch effects.*

To reason over molecular structures, we make use of features learned from GNNs trained on molecular property prediction [29]. We utilize a pretrained MPNN foundational model up to the order of 1B parameters for extracting molecular representations following a similar procedure to Sypetkowski et al. (2024) [47]. We refer to this model as *MolGPS*.

Challenge 2: Inactive Molecular Perturbations

The phenomics-molecular data collection process can result in pairing of molecular structures with unperturbed cells in cases where the molecule has no effect on cell morphology (Figure 3)

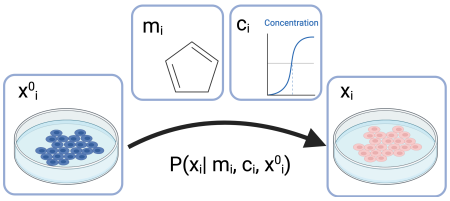


Figure 3: Data generation process of a phenomic experiment on cells $\mathbf{x_i}$ with molecular perturbations $\mathbf{m_i}$ and concentrations $\mathbf{c_i}$.

Since the morphological effects observed in cell $\mathbf{x}_i$ is conditioned on the perturbation, in the absence of a molecular effect $P(\mathbf{x_i}|\mathbf{x}_i^0, \mathbf{c_i}, \mathbf{m_i}) \sim P(\mathbf{x_i}|\mathbf{x}_i^0)$. In these samples, phenomic data will be independent, from paired molecular data, which results in misannotation under the assumption of data-pairs having an underlying biological relationship. We demonstrate how utilizing Phenom1 to undersample inactive molecules and learn continuous similarities between samples can alleviate this challenge.

To **undersample inactive molecules**, we extract the embeddings from Phenom1 and calculate the relative activity of each perturbation $(\mathbf{m}_i, \mathbf{c}_i) \in (\mathbf{M}, \mathbf{C})$. This is done using the rank of cosine similarities between technical replicates produced for a molecular perturbation against a null distribution. The null distribution is established by calculating cosine similarities from random pairs of Phenom1 embeddings generated with perturbation $(\mathbf{m}_j, \mathbf{c}_j), (\mathbf{m}_k, \mathbf{c}_k)$. Hence, we can compute a p-value and filter out samples likely to belong to the null distribution with an arbitrary threshold ψ.

In addition, by utilizing an inter-sample aware S2L **training objective**, the model can learn similarities between inactive molecules. S2L is grounded in previous work which demonstrates improved robustness to label noise (SigLip) and learnable inter-sample associations (CWCL) [52, 45]. Continuous Weighted Contrastive Loss (CWCL) provides better multi-modal alignment using a uni-modal pretrained model to suggest sample distances, relaxing the negative equidistant assumption present in InfoNCE [45]:

$$\mathcal{L}_{\text{CWCL, }\mathcal{M}\to\mathcal{X}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\sum_{j=1}^{N}\mathbf{w}_{i,j}^{\mathcal{X}}}\sum_{j=1}^{N}\mathbf{w}_{i,j}^{\mathcal{X}}\log\frac{\exp\left(\langle\mathbf{z}_{x_i},\mathbf{z}_{m_j}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle\mathbf{z}_{x_j},\mathbf{z}_{m_k}\rangle/\tau\right)}\right]. \tag{2}$$

CWCL weights logits with a continuous measure of similarity $\mathbf{w}^{\mathcal{X}}$, resulting in better alignment of embeddings $\mathbf{z}_{\mathbf{x}_i}$ and $\mathbf{z}_{\mathbf{m}_j}$ across modalities. In equation 2, $\mathbf{w}^{\mathcal{X}}$ is computed using a within modality similarity function such as $\mathbf{w}_{i,j}^{\mathcal{X}} = \langle z_{\mathbf{x}_i}, z_{\mathbf{x}_j}\rangle/2 + 0.5$. Note, the above formula is used only for mapping samples from modality $\mathcal{M}$ to $\mathcal{X}$ for which a pre-trained model $\theta_{\text{Ph}}$is available.

Another work, SigLIP, demonstrates robustness to label noise and reduces computational requirements during contrastive training [52]. It does so by avoiding computation of a softmax over the entire set of in-batch samples, instead relying on element-wise sigmoid operation:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\log\frac{1}{1+\exp\left(\mathbf{y}_{i,j}(-\alpha\,\langle\mathbf{z}_{\mathbf{x}_i},\mathbf{z}_{\mathbf{m}_j}\rangle + b)\right)}\right]. \tag{3}$$

In equation 3, α and b are learned, calibrating the model confidence conditioned on the ratio of positive to negative pairs. $\mathbf{y}_{i,j}$ is set to 1 if $i = j$ and -1 otherwise.

Inspired by prior works, we introduce S2L for molecular representation learning, which leverages inter-sample similarities and robustness to label noise to mitigate weak or inactive perturbations.

$$\mathcal{L}_{\text{S2L}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left[\frac{\mathbf{w}_{i,j}^{\mathcal{X}}}{1+\exp\left(-\alpha\mathbf{z}_{\mathbf{x}_i}.\mathbf{z}_{\mathbf{m}_j}+b\right)} + \frac{(1-\mathbf{w}_{i,j}^{\mathcal{X}})}{1+\exp\left(\alpha\mathbf{z}_{\mathbf{x}_i}.\mathbf{z}_{\mathbf{m}_j}+b\right)}\right]. \tag{4}$$

In the equation above, $\mathbf{z}_{\mathbf{x}_i}$ and $\mathbf{z}_{\mathbf{m}_j}$ correspond to latent representations of images and molecules, respectively. α and b correspond to learnable temperature and bias parameters for the calibrated sigmoid function. $\mathbf{w}_{ij}^{\mathcal{X}}$ is an inter-sample similarity function computed from images using the pretrained model θ_{Ph}. To compute $\mathbf{w}_{i,j}^{\mathcal{X}}$, we use the arctangent of L2 distance instead of cosine

similarity, as was the case for Equation 2 (more details in Appendix D.3). Intuitively, S2L can be thought of as shifting from a multi-class classification to a soft multi-label problem. In our problem setting, the labels are continuous and determined by sample similarity in the phenomics space.

> **Guideline 2** *When training a molecular-phenomic model, mitigating the effect of inactive molecules in training data distribution can be carried out by undersampling inactive molecules and using an inter-sample similarity aware,* `S2L` *loss (equation 4).*

Challenge 3: Variable Concentrations

Perturbation effect on a cell is determined by both molecular structure and corresponding concentration [49]. A model capturing molecular impact on cell morphology must be able to generalize across different doses, since variable concentrations can correspond to different data distributions.

We note that providing concentrations $\mathbf{c}_i$ as input to the model would benefit performance, as this would indicate the magnitude of molecular impact. However, we find that simply concatenating concentrations does not result in effective training due to its compressed dynamic range. To that end, we add concentration information in two separate ways: *implicit* and *explicit* formulations.

We add **implicit concentration** as molecular perturbation classes by using the S2L loss (Equation 4) to treat perturbation $\mathbf{m}_i$ with concentrations $\mathbf{c}_i$ and $\mathbf{c}_j$ as distinct classes. This pushes samples apart in the latent space proportionally to similarities between phenomic experiments.

We add **explicit concentration** c_i by passing it to the molecular encoder. We explore different formulation for dosage concentrations, $\mathbf{f}'(c_i)$, where $\mathbf{f}'$ maps $\mathbf{c_i} \rightarrow \mathbb{R}$. Encoded representations $\mathbf{f}'(c_i)$ are concatenated at the initial layer of the model. We find simple functional encodings $\mathbf{f}'$ (such as `one-hot` and `logarithm`) to work well in practice.

> **Guideline 3** *When training a molecular-phenomic model, conditioning on an (implicit and explicit) representation of concentration $\mathbf{f}'(\mathbf{c}_i)$ aids in capturing molecular impacts on cell morphology and improves generalization to previously unseen molecules and concentrations.*

4 Experimental Setup

In this section, we describe evaluation datasets used, and descriptions of the underlying data modalities. To assess phenomolecular retrieval, we use 1% recall metric unless stated otherwise, as it allows direct comparison between datasets with different number of samples. Additional implementation and evaluation details can be found in Appendix D.

Datasets: Our training dataset consists of fluorescent microscopy images paired with molecular structures and concentrations, which are used as perturbants. We assess models' phenomolecular retrieval capabilities on three datasets of escalating generalization complexity. First dataset, consisting of unseen microscopy images and molecules present in the training dataset. Second, a dataset consisting of previously unseen phenomics experiments and molecules split by the corresponding molecular scaffold. Finally, we evaluate on an open source dataset with a different data generating distribution [13]. In the case of the latter two datasets, the model is required to perform zero-shot classification, as it has no access to those molecules in the training data. This requires the model to reason over molecular graphs to identify structures inducing corresponding cellular morphology changes. Using methodology described in guideline 2 we report retrieval results for all molecules as well as on an active subset. Finally, all datasets are comprised of molecular structures at multiple concentrations ($.01, .1, 1.0, 10$, etc.) Additional details regarding the datasets can be found in Appendix C.

Modality Representations: In our evaluations, we consider different representations for molecular perturbations and phenomic experiments and quantitatively evaluate their impact.

- `Images`: Image encoders utilize 6-channel fluorescent microscopy images of cells representing phenomic experiments. Images are 2048×2048 pixels, capturing cellular morphology changes post molecular perturbation. We downscale each image to 256×256 using block mean downsampling.

- `Phenom1`: We characterize phenomic experiments by embedding high resolution microscopy images in the latent space of a phenomics model θ_{Ph} as described in guideline 1.
- `Fingerprints`: Molecular fingerprints utilize RDKIT [26], MACCS [25] and MORGAN3 [38] bit coding, which represent binary presence of molecular substructures. Additional information such as atomic identity, atomic radius and torsional angles are included in the fingerprint representations.
- `MolGPS`: We generate molecular representations from a large pretrained GNN. Specifically, we obtain molecular embeddings from a 1B parameter MPNN [29].

5 Results and Discussion

To evaluate the effectiveness of Guidelines 1, 2, and 3 we carry out evaluation in two different settings: (1) cumulative concentrations, and (2) held-out concentrations, testing the models' ability to generalize to new molecular doses. Finally, we perform comprehensive ablations testing model performance with varying data, model, and optimization parameters. The comprehensive set of results can be found in Tables 10, 11, 12, and 13.

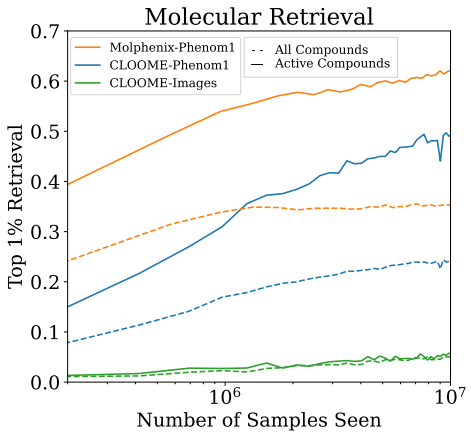


Figure 4: Comparison of training phenomic encoder from scratch and utilizing pre-trained Phenom1 unseen dataset. X-axis plotted on logarithmic scale.

5.1 Evaluation on cumulative concentrations:

We demonstrate improvements in phenomolecular recall due to usage of a phenomics pre-trained foundation model, identify that MolPhenix set of design choices results in higher final performance, and more data efficient learning. Figure 4 demonstrates recall accuracy on all molecules and an active subset for CLOOME and MolPhenix models, as a function of training samples seen.

We observe a large performance gap between models trained on Phenom1 embeddings as opposed to images, emphasizing the utility of using a pre-trained encoder for microscopy images (Table 1). We note that provision of Phenom1 (CLOOME-Phenom1 Vs CLOOME-Images) significantly improves both active and all molecule retrieval by $\mathbf{5.69}\times$ and, $\mathbf{4.75}\times$ respectively (Table 1).

Furthermore, we identify that while all molecules retrieval stagnates throughout training, the performance on an active subset keeps improving, underscoring the importance of identification of the active subset. Finally, we compare CLOOME and MolPhenix trained using Phenom1 embeddings and find there is a consistent retrieval performance gap, throughout training, with a **1.26** $\times$ final improvement (Figure 4, Table 1). Compared to CLOOME [40] trained directly on images, MolPhenix achieves an average improvement of $\mathbf{8.78}\times$ on active molecules on the unseen dataset. These results verify the effectiveness of Guideline 1 in accelerating training, and the importance of Guidelines 2 and 3 in recall improvements over CLOOME.

Table 1: Impact of pre-trained Phenom1 and MolGPS on CLOOME and MolPhenix for a matched number of seen samples (Top), where we observe an **8.1** $\times$ improvement of MolPhenix over the CLOOME baseline for active unseen molecules. SOTA results trained with a higher number of steps by utilizing the best hyperparameters (Bottom *). We note that MolPhenix's main components such as S2L and embedding averaging relies on having a pre-trained uni-modal phenomics model.

Method	Modality	Active Molecules			All Molecules		
		Unseen Im.	Unseen Im. + Mol.	Unseen Dataset	Unseen Im.	Unseen Im. + Mol.	Unseen Dataset
CLOOME	Images & Multi-FPS	$.0756 \pm .0042$	$.0787 \pm .0065$	$.0528 \pm .0057$	$.0547 \pm .0028$	$.0661 \pm .0020$	$.0223 \pm .0014$
CLOOME	Phenom1 & Multi-FPS	$.4659 \pm .0042$	$.5057 \pm .0014$	$.2065 \pm .0146$	$.3009 \pm .0053$	$.2474 \pm .0013$	$.1337 \pm .0045$
MolPhenix	Phenom1 & Multi-FPS	$\mathbf{.7807 \pm .0025}$	$.6365 \pm .0014$	$.3545 \pm .0097$	$\mathbf{.5253 \pm .0029}$	$\mathbf{.3655 \pm .0017}$	$.2163 \pm .0021$
MolPhenix	Phenom1 & MolGPS	$.7646 \pm .0014$	$\mathbf{.6387 \pm .0056}$	$\mathbf{.4160 \pm .0016}$	$.5012 \pm .0002$	$.3511 \pm .0004$	$\mathbf{.2508 \pm .0026}$
MolPhenix*	Phenom1 & MolGPS	$\mathbf{.9689 \pm .0017}$	$\mathbf{.7733 \pm .0036}$	$\mathbf{.5860 \pm .0082}$	$\mathbf{.5583 \pm .0007}$	$\mathbf{.3824 \pm .0016}$	$\mathbf{.2809 \pm .0060}$

Table 2: Top-1% recall accuracy with use of the proposed MolPhenix guidelines, such as Phenom1 and embedding averaging. We omit explicit concentration from this experiment.

Loss	Active Molecules			All Molecules		
	Unseen Images	Unseen Im. + Mol.	Unseen Dataset	Unseen Images	Unseen Im. + Mol.	Unseen Dataset
CLIP	$.3373 \pm .0043$	$.4228 \pm .0008$	$.1514 \pm .0038$	$.1761 \pm .0043$	$.1867 \pm .0022$	$.0734 \pm .0022$
Hopfield-CLIP	$.2578 \pm .0042$	$.3559 \pm .0042$	$.1256 \pm .0092$	$.1531 \pm .0046$	$.1709 \pm .0029$	$.0673 \pm .0020$
InfoLOOB	$.3351 \pm .0011$	$.4206 \pm .0031$	$.1563 \pm .0028$	$.1746 \pm .0003$	$.1860 \pm .0029$	$.0745 \pm .0019$
CLOOME	$.3572 \pm .0026$	$.4348 \pm .0039$	$.1658 \pm .0063$	$.1968 \pm .0029$	$.2005 \pm .0026$	$.0911 \pm .0022$
DCL	$.6363 \pm .0025$	$.6177 \pm .0047$	$.3184 \pm .0087$	$.3277 \pm .0047$	$.2562 \pm .0008$	$.1364 \pm .0067$
CWCL	$.7091 \pm .0045$	$.6529 \pm .0020$	$.3556 \pm .0094$	$.3635 \pm .0064$	$.2696 \pm .0019$	$.1526 \pm .0058$
SigLip	$.7763 \pm .0045$	$.6401 \pm .0065$	$.3396 \pm .0042$	$.3729 \pm .0039$	$.2544 \pm .0014$	$.1470 \pm .0038$
S2L (ours)	$\mathbf{.9097 \pm .0020}$	$\mathbf{.6759 \pm .0012}$	$\mathbf{.4181 \pm .0012}$	$\mathbf{.4688 \pm .0009}$	$\mathbf{.2852 \pm .0001}$	$\mathbf{.1838 \pm .0007}$

Table 3: Top-1% recall accuracy across different concentration encoding choices with use of the proposed MolPhenix guidelines, such as Phenom1 and embedding averaging.

Implicit Concentration	Explicit Concentration	Active Molecules			All Molecules		
		Unseen Im.	Unseen Im. + Mol.	Unseen Dataset	Unseen Im.	Unseen Im. + Mol.	Unseen Dataset
✗	✗	$.7350 \pm .0071$	$.6509 \pm .0104$	$.3333 \pm .0004$	$.3610 \pm .0025$	$.2668 \pm .0034$	$.1532 \pm .0007$
✓	✗	$.9097 \pm .0020$	$.6759 \pm .0012$	$.4181 \pm .0012$	$.4688 \pm .0009$	$.2852 \pm .0001$	$.1838 \pm .0007$
✓	sigmoid	$.9423 \pm .0011$	$.7155 \pm .0016$	$.4573 \pm .0022$	$.5071 \pm .0024$	$.3441 \pm .0026$	$.2144 \pm .0026$
✓	logarithm	$.9426 \pm .0066$	$.7451 \pm .0050$	$.4727 \pm .0056$	$.5183 \pm .0027$	$.3700 \pm .0036$	$.2275 \pm .0032$
✓	one-hot	$\mathbf{.9430 \pm .0029}$	$\mathbf{.7490 \pm .0052}$	$\mathbf{.4850 \pm .0020}$	$\mathbf{.5433 \pm .0030}$	$\mathbf{.3819 \pm .0032}$	$\mathbf{.2384 \pm .0049}$

We evaluate the impact of different loss objectives on the proposed MolPhenix training framework. Table 2 presents top-1% retrieval accuracy across different contrastive losses utilized to train molecular-phenomics encoders on cumulative concentrations. Compared to prior methods, the proposed `S2L` loss demonstrates improved retrieval rates in cumulative concentration setting. Label noise and inter-sample similarity aware losses such as CWCL and SigLip also demonstrate improved performance. The effectiveness of S2L can be attributed to smoothed inter-sample similarities and implicit concentration information.

Finally, in Table 3, we observe recall improvements when considering both molecular structures and concentration. We note the importance of the addition of implicit concentration, further confirming the importance of considering molecular effects at different concentrations as different classes. Explicitly encoding molecular concentration with `one-hot`, `logarithm` and `sigmoid` yields improved recall performance, where `one-hot` performs the best in a cumulative concentration setting. These findings verify the efficacy of implicit and explicit concentration encoding outlined in Guideline 3.

Table 4: Top-1% recall accuracy of different loss objectives while using the proposed MolPhenix guidelines, such as Phenom1 and embedding averaging.

Loss	Unseen Im.	Unseen Im. + Mol.	Unseen Dataset
CLIP	.2109	.2425	.1519
Hopfield-CLIP	.1581	.2034	.1198
InfoLOOB	.2122	.2496	.1501
CLOOME	.2164	.2461	.1479
DCL	.4717	.4027	.2841
CWCL	.5731	.4403	.3232
SigLip	.5718	.4217	.3021
S2L (ours)	**.8334**	**.4615**	**.3792**

Table 5: Top-1% recall accuracy across different concentration encoding choices while using the proposed MolPhenix guidelines, such as Phenom1 and embedding averaging.

Implicit Concentration	Explicit Concentration	Unseen Im.	Unseen Im. + Mol.	Unseen Dataset
✗	✗	.5942	.4315	.3129
✓	✗	**.8334**	.4615	**.3792**
✓	sigmoid	.8256	**.4692**	.3765
✓	logarithm	.7953	.4466	.3664
✓	one-hot	.7489	.4088	.3379

Results are averaged across experiments for each dropped concentration, and across three seeds. Recall is reported for active molecules, while the results for all molecules can be found in Table 13.

5.2 Evaluation on held-out concentrations:

Next, we evaluate recall on held-out concentrations to obtain a measure of generalization performance. This evaluation allows us to capture the utility of our models for prediction of unseen concentrations,

hence resembling *in-silico* testing. We omit concentrations from the training set and evaluate recall at the excluded data, where we observe a drop in retrieval performance for unseen concentrations. Similar to cumulative concentration results, we find that using `S2L` improves recall over other losses and outperforms CLOOME by up to 126% (Table 4). While `one-hot` encoding exhibits significant improvements in cumulative concentrations, its expressivity on unseen concentrations is limited (Table 5) and `sigmoid` encoding provides a sufficient representation of concentration.

5.3 Ablation Studies

We assess the importance of our design decisions by conducting an ablation study over our proposed guidelines. Figure 5 presents the variation of top-1% recall accuracy across key components such as cutoff p value, fingerprint type, and embedding averaging. We observe that employing a lower cutoff p value yields improved generalization for unseen dataset, while employing a higher cutoff appears to be optimal for unseen images + unseen molecules. For molecular structure representations, we find that using embeddings from the large pretrained MPNN graph based model (e.g., MolGPS) surpasses traditional fingerprints. Finally, utilization of embedding averaging demonstrates improved recall. More ablations over model size, projector dimension, and batch size can be found in Appendix E.5.

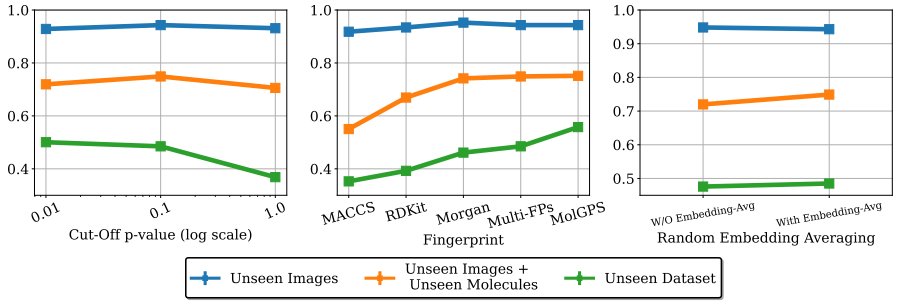


Figure 5: Ablations of top-1 % recall accuracy with **(bottom-left)** cutoff p value, **(bottom-center)** fingerprint type, and **(bottom-right)** embedding averaging.

6 Conclusion

In this work, we investigate the problem of *contrastive phenomolecular retrieval* by constructing a joint multi-modal embedding of phenomic experiments and molecular structures. We identify a set of challenges afflicting molecular-phenomic training and proposed a set of guidelines for improving retrieval and generalization. Empirically, we observed that contrastive learners demonstrate higher retrieval rates when using representations from a high-capacity uni-modal pretrained model. Use of inter-sample similarities with a label noise resistant loss such as `S2L` allows us to tackle the challenge of inactive molecules. Finally, adding implicit and explicit concentrations allows models to generalize to previously unseen concentrations. MolPhenix demonstrates an **8.1**$\times$ improvement in zero shot retrieval of active molecules over the previous state-of-the-art, reaching 77.33% in top-1% accuracy. In addition, we conduct a preliminary investigation on MolPhenix's ability to uncover biologically meaningful properties (activity prediction, zero-shot biological perturbation matching, and molecular property prediction in Appendix E.1, E.2, and E.3, respectively.). We expect a wide range of applications for MolPhenix, particularly in drug discovery. While there's a remote chance of misuse for developing chemical weapons, such harm is unlikely, with our primary focus remaining on healthcare improvement.

Limitations and Future Work: While our study covers challenges in phenomolecular recall, we leave three research directions for future work. (1) Future investigations could consider studying additional modalities such as text, genetic perturbations and chemical multi-compound interventions. (2) While we propose and evaluate our guidelines on previously conducted phenomic experiments, we note that a rigorous evaluation would evaluate model predictions in a wet-lab setting. (3) In addition, our work makes the assumption that the initial unperturbed cell state x_i^0 can be marginalized by utilizing a single cell line with an unperturbed genetic background. Future works can relax this assumption, aiming to capture innate intercellular variation.

References

[1] 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.

[2] S. Albelwi. Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

[3] 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.

[4] C. Bock, P. Datlinger, F. Chardon, M. A. Coelho, M. B. Dong, K. A. Lawson, T. Lu, L. Maroc, T. M. Norman, B. Song, G. Stanley, S. Chen, M. Garnett, W. Li, J. Moffat, L. S. Qi, R. S. Shapiro, J. Shendure, J. S. Weissman, and X. Zhuang. High-content crispr screening. *Nature Reviews Methods Primers*, 2(1), Feb. 2022.

[5] 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*, 16(1):3–50, Jan. 1996.

[6] M. Boutros, F. Heigwer, and C. Laufer. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.

[7] 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*, 11(9):1757–1774, 2016.

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

[9] S. N. Chandrasekaran, J. Ackerman, E. Alix, D. M. Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino, P. J. Byrne, H. Ceulemans, C. Ch'ng, B. A. Cimini, D.-A. Clevert, N. Deflaux, J. G. Doench, T. Dorval, R. Doyonnas, V. Dragone, O. Engkvist, P. W. Faloon, B. Fritchman, F. Fuchs, S. Garg, T. J. Gilbert, D. Glazer, D. Gnutt, A. Goodale, J. Grignard, J. Guenther, Y. Han, Z. Hanifehlou, S. Hariharan, D. Hernandez, S. R. Horman, G. Hormel, M. Huntley, I. Icke, M. Iida, C. B. Jacob, S. Jaensch, J. Khetan, M. Kost-Alimova, T. Krawiec, D. Kuhn, C.-H. Lardeau, A. Lembke, F. Lin, K. D. Little, K. R. Lofstrom, S. Lotfi, D. J. Logan, Y. Luo, F. Madoux, P. A. Marin Zapata, B. A. Marion, G. Martin, N. J. McCarthy, L. Mervin, L. Miller, H. Mohamed, T. Monteverde, E. Mouchet, B. Nicke, A. Ogier, A.-L. Ong, M. Osterland, M. Otrocka, P. J. Peeters, J. Pilling, S. Prechtl, C. Qian, K. Rataj, D. E. Root, S. K. Sakata, S. Scrace, H. Shimizu, D. Simon, P. Sommer, C. Spruiell, I. Sumia, S. E. Swalley, H. Terauchi, A. Thibaudeau, A. Unruh, J. Van de Waeter, M. Van Dyck, C. van Staden, M. Warchoł, E. Weisbart, A. Weiss, N. Wiest-Daessle, G. Williams, S. Yu, B. Zapiec, M. Żyła, S. Singh, and A. E. Carpenter. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023.

[10] 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*, 33:22243–22255, 2020.

[11] 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*, pages 7480–7512. PMLR, 2023.

[12] 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.

[13] 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*, pages 2023–02, 2023.

[14] C. Feichtenhofer, Y. Li, K. He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[15] A. Fürst, E. Rumetshofer, J. Lehner, V. T. Tran, F. Tang, H. Ramsauer, D. Kreil, M. Kopp, G. Klambauer, A. Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35:20450–20468, 2022.

[16] 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.

[17] 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*, pages 16000–16009, 2022.

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

[19] 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*, 59(3):1163–1171, 2019.

[20] 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.

[21] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[22] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. L. Chin, S. A. Strawbridge, M. Garcia-Patino, R. Kruger, A. Sivakumaran, S. Sanford, R. Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski, D. Y. Rayat, H. Jackson, K. Harford, A. Anjum, M. Zakir, F. Wang, S. Tian, B. Lee, J. Liigand, H. Peters, R. Q. R. Wang, T. Nguyen, D. So, M. Sharp, R. da Silva, C. Gabriel, J. Scantlebury, M. Jasinski, D. Ackerman, T. Jewison, T. Sajed, V. Gautam, and D. S. Wishart. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, Nov. 2023.

[23] 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.

[24] O. Z. Kraus, B. T. Grys, J. Ba, Y. Chong, B. J. Frey, C. Boone, and B. J. Andrews. Automated analysis of high-content microscopy data with deep learning. *Molecular systems biology*, 13(4):924, 2017.

[25] H. Kuwahara and X. Gao. Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *Journal of Cheminformatics*, 13:1–12, 2021.

[26] G. Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

[27] F. Lanusse, L. Parker, S. Golkar, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Ohana, M. Pettee, B. R.-S. Blancard, T. Tesileanu, K. Cho, and S. Ho. Astroclip: Cross-modal pre-training for astronomical foundation models, 2023.

[28] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[29] D. Masters, J. Dean, K. Klaser, Z. Li, S. Maddrell-Mander, A. Sanders, H. Helal, D. Beker, A. Fitzgibbon, S. Huang, et al. Gps++: Reviving the art of message passing for molecular property prediction. *arXiv preprint arXiv:2302.02947*, 2023.

[30] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

[31] O. Méndez-Lucio, C. Nicolaou, and B. Earnshaw. Mole: a molecular foundation model for drug discovery, 2022.

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

[33] H. S. Parker and J. T. Leek. The practical effect of batch on genomic prediction. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

[34] 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*, pages 5171–5180. PMLR, 2019.

[35] 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. PMLR, 2021.

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

[37] 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.

[38] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[39] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang. Self-supervised graph transformer on large-scale molecular data, 2020.

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

[41] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[42] 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*, 25(5):611–618, 2018.

[43] 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*, NIPS'16, page 1857–1865, Red Hook, NY, USA, 2016. Curran Associates Inc.

[44] C. Soneson, S. Gerster, and M. Delorenzi. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS one*, 9(6):e100335, 2014.

[45] 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*, 36, 2023.

[46] M. Sypetkowski, M. Rezanejad, S. Saberian, O. Kraus, J. Urbanik, J. Taylor, B. Mabey, M. Victors, J. Yosinski, A. R. Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4284–4293, 2023.

[47] M. Sypetkowski, F. Wenkel, , F. Poursafaei, N. Dickson, K. Suri, P. Fradkin, and D. Beaini. On the scalability of foundational models for molecular graphs. *arxiv*, 2024.

[48] 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*, 21(12):899–914, 2022.

[49] R. M. Walmsley and N. Billinton. How accurate is in vitro prediction of carcinogenicity? *British Journal of Pharmacology*, 162(6):1250–1258, Feb. 2011.

[50] 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, June 2023.

[51] 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.

[52] 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*, pages 11975–11986, 2023.

[53] 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.

[54] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022.

[55] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, and G. Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

A NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we claim that we build a multi-modal molecular-phenomics model and demonstrate improvements over prior works. This is done by taking using a uni-modal pre-trained phenomics model, tackling inactive molecules by undersampling and learning inter-sample similarities. In addition, we take into account concentration in our model training. We demonstrate comprehensive results supporting these claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the conclusion, we have a limitations subsection discussing future research directions and assumptions in our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work does not contain proofs or theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our work documents our design decisions in detail and has comprehensive details about the underlying dataset. We document all our hyperparameter choices and model architectural decisions. Our evaluation is performed on a publicly accessible dataset RXRX3, allowing for benchmarking of other methods. To reproduce the pre-trained phenomics model, we base our architecture on the work from [23], for which they have also provided access to a snakker model, namely Phenom-Beta via a web platform hosted on the BioNeMo platform `https://www.rxrx.ai/phenom`. To reproduce the pre-trained molecular model, we based our architecture on [47], for which the authors provide all the code and data needed to reproduce it. We further note that the molecular model can be replaced by simple molecular fingerprints with only a slight drop in performance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: As part of the submission, we are unable to provide code to reproduce model training due to use of its proprietary nature. The training dataset is also an asset of a private institution, meaning that we are unable to be made publicly accessible. The unseen dataset RXRX3 is, however, open source and can be used by the community to evaluate public phenomics models.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details regarding our hyperparameter choices in the Appendix C. In addition we document the use of scaffold splitting for Unseen Molecules & Images dataset. Unseen Dataset RXRX3 is publicly accessible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our reported results are averaged over 3 random seeds used to initialize the model and dictating stochasticity during model training. We report most standard deviations in the main text, and the remaining ones are all present in the Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on compute time for each experiment in Appendix D.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research described does not violate the NeurIPS Code of Ethics. Our experiments do not include human subjects, we follow fair use of data, privacy, and do not release model weights for mitigating impact measures.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work discusses the potential in which MolPhenix can have positive societal impact and we touch on the extenralities in our concluding statements.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: In our work we do not release model weights or the underlying code.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Assetts used are referenced and licenses checked or otherwise not released publicly.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing not human subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing not human subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

B Assumption of the Initial Cell State

There is an important distinction between phenomics - molecule and text - image contrastive training although there are initial similarities. In the text - image domain the two modalities are directly generated by the same latent variable which is the underlying semantic class. Whereas in phenomics - molecule, the observed phenomics variable is actually conditioned on molecular structure and the initial state. There are two important conclusions from this: (1) This indicates that if molecular structure has no effect on the initial cell state, there will not be a positive pairing between the molecular structure and morphological patterns captured by phenomics, making it indistinguishable from a control image. (2) There is an underlying assumption that the initial cell state x_i^0 is constant. In accordance with this assumption we utilize experiments with a fixed cell line, *HUVEC-19*, and a constant genetic background. Future works can relax this assumption by taking into account phenomics experiments of the cells prior to the perturbation. This can allow the models to generalize beyond a single cell line and to diverse genetic backgrounds.

C Dataset

Models have been trained using our in house training set and we have conducted our evaluation on two novel datasets and an open-source molecule dataset [13]:

- `Training Set`: Our training dataset comprises 1,316,283 pairs of molecules and concentration concentration combinations, complemented by fluorescent microscopy images generated through over 2,150,000 phenomic experiments.
- `Evaluation set 1 - Unseen Images + Seen Molecules`: The first set consists of unseen images and seen molecules. Unseen microscopy images are associated with 15,058 pairs of molecules and concentrations from the training set and selected randomly.
- `Evaluation set 2 - Unseen Images + Unseen Molecules`: The second set includes previously unseen molecules, and images (consisting of 45,771 molecule and concentration pairs). Predicting molecular identities of previously unseen molecular perturbations corresponds to zero-shot prediction. Scaffold splitting was used to split this validation dataset from training ensuring minimal information leakage.
- `Evaluation set 3 - Unseen Dataset`: Finally, we utilize the RXRX3 dataset [13], an open-source out of distribution (OOD) dataset consisting of 6,549 novel molecule and concentration pairs associated with phenomic experiments. The distribution of molecular structures differs from previous datasets, making this a challenging zero-shot prediction task.

C.1 Concentration Details

Additional details regarding the number of molecules at significant concentrations of each evaluation set are available in Table 6.

Table 6: Separated number of molecules for different concentrations at various pvalue cut-offs

	pvalue=1.0			pvalue=.1			pvalue=.01		
Concentration	Unseen Im.	Unseen Im. + Mol.	Unseen Data	Unseen Im.	Unseen Im. + Mol.	Unseen Data	Unseen Im.	Unseen Im. + Mol.	Unseen Data
.1	1497	1109	0	387	170	0	161	68	0
.25	1775	1111	1638	600	203	237	334	121	165
1.0	2721	11392	1639	1259	734	390	672	390	268
2.5	1787	4018	1636	1329	644	516	929	413	375
3.0	74	10454	0	12	1540	0	4	729	0
5.0	3	50	0	0	27	0	0	20	0
1.0	2712	11392	1636	2544	8117	792	2116	4815	625
25.0	0	2916	0	0	1734	0	0	950	0
Unique molecules	3026	14256	1639	2729	9857	823	2309	5778	642

D Implementation Details

In our experiments we report the top 1% recall metric as it is agnostic to the size of the dataset used. Across different datasets, top 1 metric can correspond to varying levels of difficulty due to the number of negatives evaluated. Top 1% can be used to compare models with different batch sizes, datasets, and evaluations with different number of concentrations.

D.1 Hyperparameters

Our design choices and utilized hyperparameters for is presented in Table 7. We set batch size to 512 through experiments presented in top section of Table 1 and Figure 4 since training CLOOME model on images is not efficient compared to using pretrained models. In addition, results presented at bottom section of Table 1 are based on the best parameters found through described ablation studies (section E.5).

Table 7: Hyperparameter values utilized in our proposed MolPhenix training framework.

Parameter	Value
number of seeds	3
learning rate	1e-3
weight decay	3e-3
optimizer	`AdamW`
training batch size	8192
validation batch size	12000
embedding dim	512
model size	medium (38.7 M)
model structure	6 ResNet Blocks + 1 Linear layer + 1 ResNet Block + 1 Linear layer
epochs	100
self similarity clip val	.75
learnable temperature initialization	2.302
learnable bias initialization	-1.0
Distance function	arctangent of l2 distance

D.2 Resource Computation

We utilized an NVIDIA A100 GPU to train Molphenix using Phenom1 and MolGPS embeddings, which takes approximately $\sim$4.75 hours each. For loss comparison experiments, we run each model using 3 different seeds and 8 different losses, resulting in a total of 114 hours of GPU processing time. For concentration experiments we train 7 runs, one for each concentration, with 3 seeds each totaling 21 runs per set of parameters. With 25 sets of parameters evaluated (13), that amounts to 2,500 A100 compute hours. Moreover, we employed 8 NVIDIA A100 GPUs to train CLOOME model on phenomics images, with an average of 40 hour usage per run. Across three seeds, that amounts to $\sim$ 1000 hours of A100 GPU usage (8 GPUs for 40 hours 3 times).

Note that, without accounting for the time to train Phenom1, MolPhenix is 8.4 $\times$ faster than the CLOOME baseline.

D.3 S2L Distance function

To calculate inter sample distances, we utilize arctangent of l2 distances between Phenom1 embeddings. More specifically, we calculate distances with

$$\arctan(\|z_{\mathbf{x}_i} - z_{\mathbf{x}_j}\|_2^2/c) * \frac{4}{\pi} - 1, \tag{5}$$

where c is a constant indicating the median l2 distance between a null set of embeddings. Empirically, we've found that setting similarities below a threshold k to 0 improves model performance: $\lceil w \rceil^k$.

Usage of arctan-l2 distances is motivated by an observation that cosine similarities do not effectively separate inactive molecules from other molecular pairs (Figure 6). To alleviate inactive molecule challenge, we require significant separation of CDF curves of inactive perturbations (p value > .9) and active molecules (p < .01). We observe that in both the plots using arctangent and cosine similarities achieves this purpose. However, if we compare high p-value curves with high-low, we find that in the case of cosine similarities they are almost identical. This indicates that the distribution of

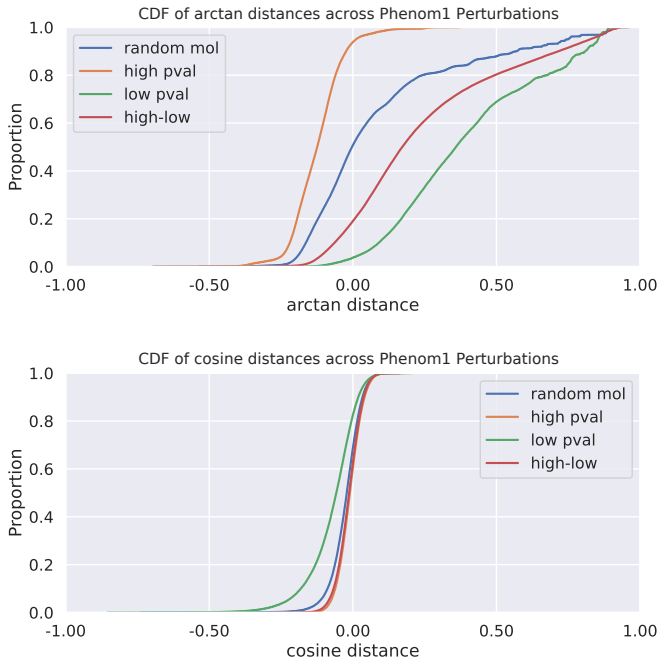


Figure 6: Plotted are cumulative densities of distance metrics for cosine similarity and arctangent of l2 distances between embeddings. Random mol corresponds to Phenom1 distances between random molecules, high pval corresponds to distances between molecules with high p-values, low pval corresponds to distances between active molecules with low p-values, finally high-low corresponds to distances between active and inactive molecules.

cosine similarities between active - inactive molecules is almost identical to that of inactive - inactive molecules. In contrast, when using arctangent similarities, we observe that the two CDF curves are well separated.

This property of l2 distances can inform our model training to identify inactive-inactive molecules. These results informed our decision to utilize arctangent of l2 distances to specify sample similarities for the S2L loss.

E Additional Results

E.1 Predicting molecular activity

Given the significance of identifying active molecules, we evaluate the ability of the chemical encoder to predict molecular activity. To do so, we assessed whether embeddings generated from the chemical encoder can be used to predict calculated p-values for unseen molecules. We fit a logistic regression on molecular embeddings from the training set, classifying whether a molecular perturbation and concentration have a p-value below .01. We find that the trained logistic regression is capable of predicting molecular activity on two downstream datasets with a non-overlapping set of molecules, Figure 8. In addition, we provide a u-map of molecular embedding for the unseen dataset RXRX3, colored by p-value. We qualitatively observe a clustering of active molecules using a U-map (Figure 7). It demonstrates that predicting compounds activity is possible using MolPhenix chemical encoder as molecules representations are distinct, independent of the experimented dosage concentration.

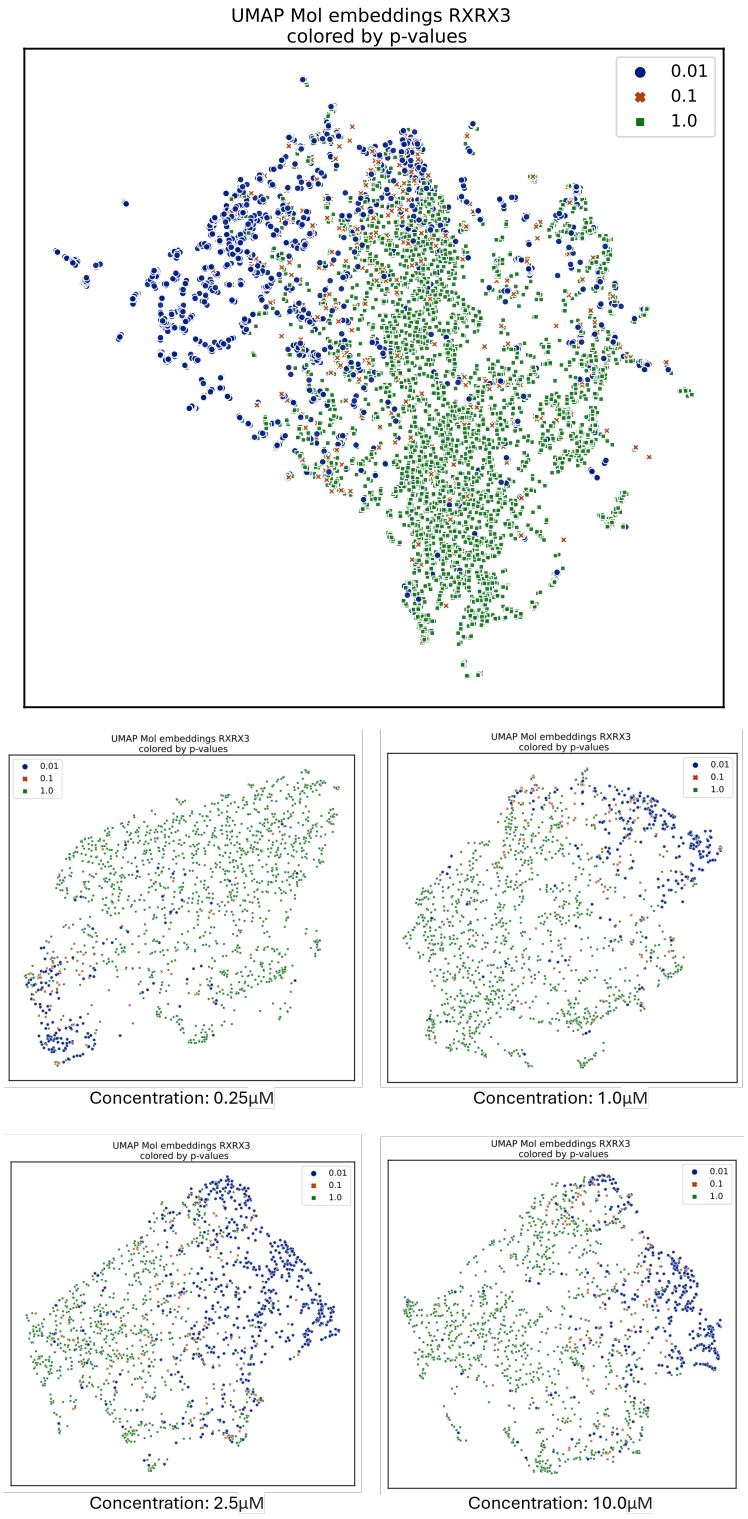


Figure 7: U-map demonstrating dimensionality reduction of the chemical embeddings of unseen dataset RXRX3. First two dimensions are visualized and points are colored corresponding to their activity measured in phenomics experiments. Activity is evaluated using p-values calculated using technical replicability of Phenom1 embeddings. Top plot shows the u-map figure of all chemical embeddings, and bottom figure contains u-map figure of representations at specific concentrations.

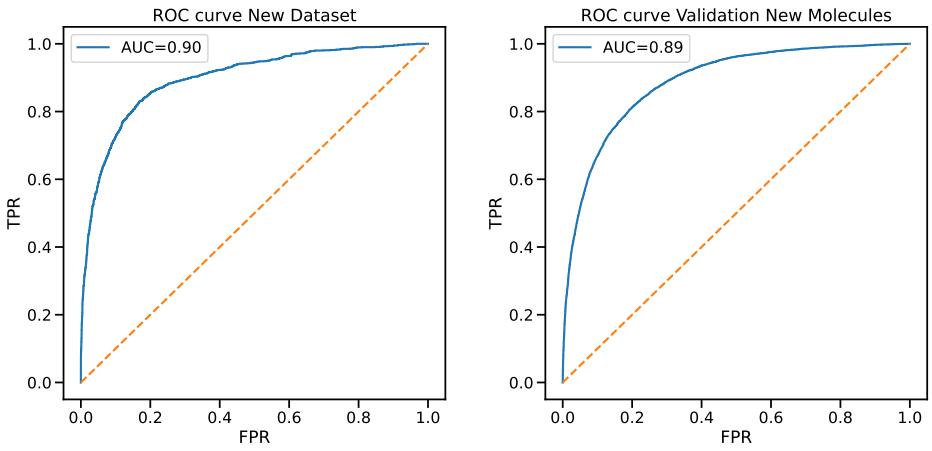


Figure 8: **Top left:** ROC AUC of logistic regression predicting molecular activity on new dataset. **Top right:** ROC AUC of logistic regression predicting molecular activity on validation dataset with new molecules and new images.

E.2 Zero Shot Biological Validation

We conduct a preliminary investigation into whether MolPhenix can be used to identify biological relationships without the need for conducting the underlying experiments. To this end, we evaluate on a subset of ChEMBL with curated pairs of gene knockouts and molecular perturbants [30]. These pairs of perturbations were curated due to the similarity of their effects on cells, although these might not always be captured through phenomic experiments. Thus, there is maximum performance that can be reached through just phenomic data, which we assume to be achieved by experimental data embedded using Phenom1.

To evaluate MolPhenix's ability to identify previously known biological associations directly from data, we embed phenomics experiments from gene knockouts using the vision encoder. To perform in-silico screening, we then embed the molecular structures associated with positive pairs using the chemical encoder. Generating molecular embeddings and the corresponding concentrations does not utilize any experimental data. We then calculate cosine similarities between embeddings of phenomics experiments evaluating gene knockouts, and representations of the chemical representations along with encoded concentrations. Using the computed cosine similarities we are then able to assess whether MolPhenix is capable of identifying known associations between gene knockouts and molecular structures. Since there is no information on molecular concentration at which the cells must be treated with, we repeat the experiment across 4 concentrations. To get a null distribution of cosine similarities we take pairs of genes knockouts and molecules for which there are no annotated relationships. We calculate a cut-off for a low and high percentiles, and then evaluate what percentage of pairs of genes and molecules with known relationships exceed the set thresholds.

Figure 9 demonstrates that in-silico screening using MolPhenix Molecular encoder is capable of recovering a significant portion of known interactions. This is performed without the use of experimental data on the molecular encoder. It is difficult to estimate an upper bound on the expected performance due to uncertainty in the quality of curation of known pairs, presence of unknown associations between genes and molecules, and uncertainty regarding molecular concentration. There is a clear trend however that MolPhenix molecular encoder is capable of recovering a meaningful fraction of these interactions.

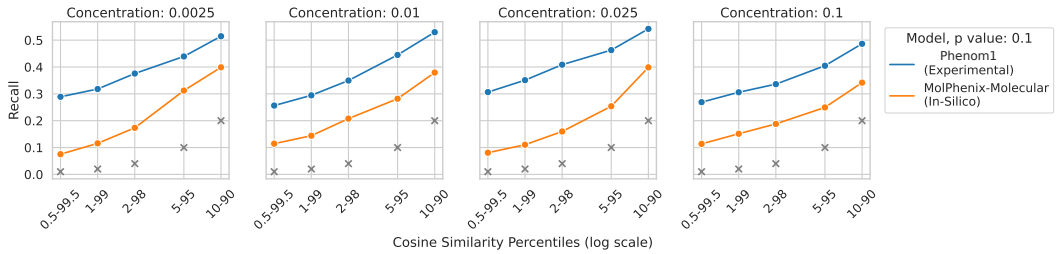


Figure 9: Evaluation of 0-shot ChEMBL identification of gene knockout and molecular phenomic similarities. On the X axis are percentile ranges, at which points the threshold is computed for cosine similarities. On the y axis is plotted total recall of recovered known interactions. Grey x plotted for each range indicate baseline recall. Orange line indicates MolPhenix-Molecular encoding of chemical compounds and MolPhenix-Vision for encoding gene knockout phenomics experiment. Blue line indicates Phenom1 encoding of phenomics experiments for both the molecular perturbation and gene knockouts. In-silico encoding of molecular perturbation, as well as the corresponding concentration, recovers a significant fraction of observed interactions.

E.3 Molecular Property Prediction

We expand our evaluation with additional experiments supporting the utility of MolPhenix beyond retrieval. We conduct a KNN evaluation of the MolPhenix latent space, assessing the learned embedding on 35 molecular property prediction tasks across the Polaris and TDC datasets (Table 8 and 9). We find that MolPhenix trained with fingerprint embeddings consistently outperforms standalone input fingerprints, demonstrating that the MolPhenix latent space effectively clusters molecules according to their biological properties. We observed an interesting effect where prediction quality is positively correlated with implied dosage, indicating that MolPhenix learns dosage-specific effects. In addition, utilizing

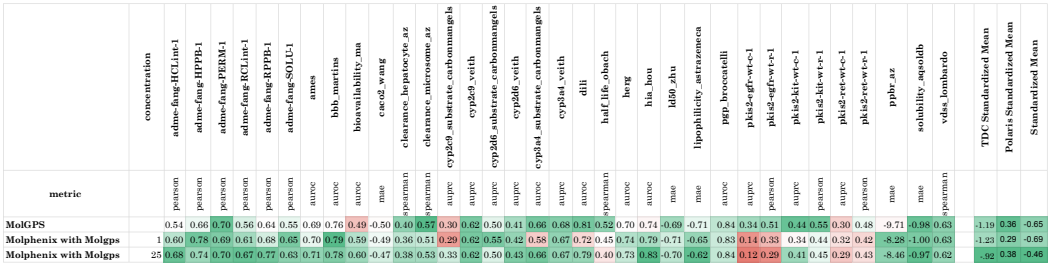
Table 8: Comparison of a KNN applied on MolPhenix molecular embedding with **traditional fingerprints** on different tasks of TDC and Polaris datasets. Mean results for TDC, Polaris and together are available in the last three columns. Binary fingerprints use tanimoto similarity, while floating-point fingerprints use cosine similarity.

	concentration	adme-fang-HClint-1	adme-fang-HPPB-1	adme-fang-PERM-1	adme-fang-RClint-1	adme-fang-RPPB-1	adme-fang-SOLU-1	ames	bbb_martins	bioavailability_ma	caco2_wang	clearance_hepatocyte_az	clearance_microsome_az	cyp2c9_substrate_carbonmangels	cyp2c9_veith	cyp2d6_substrate_carbonmangels	cyp2d6_veith	cyp3a4_substrate_carbonmangels	cyp3a4_veith	dili	half_life_obach	herg	hia_hou	ld50_zhu	lipophilicity_astrazeneca	pgp_broccatelli	pkis2-egfr-wt-c-1	pkis2-egfr-wt-s-1	pkis2-kit-wt-c-1	pkis2-kit-wt-s-1	pkis2-ret-wt-c-1	pkis2-ret-wt-s-1	ppbr_az	solubility_aqsoldb	vdss_lombardo	TDC Standardized Mean	Polaris Standardized Mean	Standardized Mean
metric		pearson	pearson	pearson	pearson	pearson	pearson	auroc	auroc	auroc	mae	spearman	spearman	auprc	auprc	auprc	auprc	auprc	auroc	auprc	auroc	spearman	auroc	auroc	mae	mae	auroc	auprc	pearson	auprc	pearson	auprc	pearson	auprc	mae	mae	spearman			
rdkit		0.32	0.34	0.48	0.23	0.38	0.29	0.69	0.72	0.58	-0.54	0.25	0.45	0.31	0.45	0.45	0.29	0.54	0.59	0.71	0.26	0.61	0.71	-0.70	-0.84	0.76	0.20	0.42	0.33	0.53	0.36	0.45	-13.03	-1.63	0.22		-2.62	-0.51	-1.88
ecfp		0.46	0.60	0.49	0.43	0.60	0.39	0.69	0.75	0.48	-0.43	0.37	0.50	0.32	0.52	0.44	0.33	0.60	0.64	0.67	0.47	0.73	0.65	-0.73	-0.78	0.79	0.41	0.57	0.33	0.51	0.40	0.55	-9.91	-1.27	0.47		-1.96	0.03	-1.26
maccs		0.37	0.56	0.52	0.22	0.43	0.44	0.71	0.77	0.53	-0.47	0.35	0.42	0.32	0.49	0.45	0.32	0.62	0.61	0.75	0.43	0.66	0.70	-0.66	-0.83	0.79	0.21	0.35	0.25	0.32	0.44	0.49	-10.13	-1.47	0.46		-1.91	-0.45	-1.40
Concatnated fps		0.41	0.66	0.58	0.33	0.40	0.37	0.70	0.77	0.58	-0.43	0.38	0.52	0.33	0.54	0.42	0.33	0.57	0.62	0.74	0.45	0.70	0.72	-0.67	-0.80	0.84	0.36	0.56	0.34	0.57	0.44	0.57	-10.94	-1.46	0.48		-1.78	0.00	-1.15
Molphenix fingerprint	1	0.57	0.75	0.57	0.55	0.72	0.57	0.70	0.74	0.54	-0.48	0.29	0.46	0.32	0.57	0.47	0.38	0.59	0.64	0.77	0.55	0.67	0.69	-0.71	-0.70	0.80	0.20	0.41	0.30	0.43	0.31	0.39	-8.93	-1.10	0.55		-1.64	0.14	-1.01
Molphenix fingerprint	25	0.64	0.71	0.65	0.62	0.67	0.58	0.69	0.78	0.54	-0.42	0.30	0.43	0.32	0.56	0.49	0.42	0.60	0.67	0.77	0.38	0.69	0.74	-0.67	-0.66	0.84	0.17	0.42	0.32	0.39	0.37	0.46	-8.43	-1.02	0.50		-1.40	0.26	-0.82

Table 10: **Evaluation on cumulative concentrations for active molecules:** Average Top-1% and Top-5% recall accuracies of methods utilizing different contrastive learning loss functions and concentration encoding information. We evaluate all methods on unseen images, unseen images and unseen molecules and an unseen dataset for zero-shot retrieval. Entries in **bold** denote best performance when the loss function is fixed while entries in highlight denote best performance across all guidelines.

Method	Explicit Concentration (ours)	Modality	top-1%				top-5%			
			Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
CLIP	✗	Phenom1	**.3373**	**.4228**	**.1514**	.3038	**.6162**	**.7182**	**.3660**	.5668
Hopfield-CLIP	✗	Phenom1	**.2578**	**.3559**	**.1256**	.2464	**.5457**	**.6751**	**.3270**	.5159
InfoLOOB	✗	Phenom1	**.3351**	**.4206**	**.1563**	.3040	**.6128**	**.7204**	**.3730**	.5687
CLOOME	✗	Phenom1	.3572	.4348	.1658	.3193	.6330	.7259	.3918	.5836
CLOOME	sigmoid	Phenom1	.5813	.4968	.2360	.4380	.8748	.7658	.4859	.7088
CLOOME	logarithm	Phenom1	**.6057**	**.5255**	**.2445**	**.4586**	**.8858**	.8117	**.4957**	**.7310**
CLOOME	one-hot	Phenom1	.5967	.5255	.2380	.4534	.8800	**.8120**	.4829	.7250
DCL	✗	Phenom1	.6363	.6177	.3184	.5241	.8638	.8180	.5632	.7483
DCL	sigmoid	Phenom1	.8858	.6694	.4527	.6693	**.9600**	.8472	.6845	.8305
DCL	logarithm	Phenom1	.8934	.6952	.4511	.6799	.9581	**.8788**	**.6889**	**.8419**
DCL	one-hot	Phenom1	**.8901**	**.7002**	**.4601**	**.6834**	.9591	.8770	.6873	.8411
CWCL	✗	Phenom1	.7091	.6529	.3556	.5725	.9018	.8368	.6027	.7804
CWCL	sigmoid	Phenom1	.9138	.6985	.4810	.6977	.9681	.8643	.7070	.8464
CWCL	logarithm	Phenom1	.9141	.7248	.4815	.7068	.9651	.8920	**.7131**	**.8567**
CWCL	one-hot	Phenom1	**.9128**	**.7261**	**.4850**	**.7079**	**.9665**	**.8927**	.6998	.8530
SigLip	✗	Phenom1	.7763	.6401	.3396	.5853	.9361	.83038	.5714	.7792
SigLip	sigmoid	Phenom1	.9463	.6931	.4576	.6990	.9816	.8606	.6759	.8393
SigLip	logarithm	Phenom1	**.9493**	.7256	.4868	**.7205**	.9814	**.8926**	.7019	**.8586**
SigLip	one-hot	Phenom1	.9489	**.7210**	**.4859**	.7186	**.9823**	.8868	**.7045**	.8578
MolPhenix (ours)	✗	Phenom1	.9097	.6759	.4181	.6679	.9768	.8539	.6436	.8247
MolPhenix (ours)	sigmoid	Phenom1	.9423	.7155	.4573	.7050	.9808	.8775	.6778	.8453
MolPhenix (ours)	logarithm	Phenom1	.9426	.7451	.4727	.7201	.9808	.8964	.6952	.8574
MolPhenix (ours)	one-hot	Phenom1	**.9430**	**.7490**	**.4850**	**.7256**	**.9816**	**.8984**	**.7040**	**.8613**
MolPhenix (ours)	✗	Phenom1 + MolGPS	.9105	.6710	.4501	.6772	.9755	.8527	.7098	.8460
MolPhenix (ours)	sigmoid	Phenom1 + MolGPS	.9395	.7034	.5252	.7227	.9811	.8729	.7630	.8723
MolPhenix (ours)	logarithm	Phenom1 + MolGPS	.9413	.7505	.5473	.7463	.9811	**.9085**	**.7878**	**.8924**
MolPhenix (ours)	one-hot	Phenom1 + MolGPS	**.9430**	**.7514**	**.5577**	**.7507**	**.9830**	.9043	.7821	.8898

Table 9: Comparison of a KNN applied on MolPhenix molecular embedding with **MolGPS** on different tasks of TDC and Polaris datasets. Mean results for TDC, Polaris and together are available in the last three columns.



E.4 Addressing Challenges in Contrastive Phenomic Retrieval

Table 10 and 12 show the complete Top 1% and 5% results of evaluation on cumulative concentrations on only active and all molecules, respectively. Similarly, Table 11 and 13 presents the full retrieval results of held-out concentrations experiments. In comparison to prior loss functions, S2L loss objective demonstrates consistent high retrieval rate in all tasks and molecular groups (i.e. all or active molecules), while using the same modality (Phenom1) and with or without explicit concentration information.

E.5 Ablation Studies

Figure 10 and Table 15, 16, 17, 18 and 19 present top-1% recall accuracy across for the full ablation study on the variation of MolPhenix key components. We note that compact embedding sizes from pretrained models stabilize training. This indicates that embeddings are expressive and accurately capture intricate aspects of molecules. Larger batch sizes result in a greater number of negative samples, hence improving performance. This is in line with prior contrastive learning methods continuing to improve by increasing the batch size [10]. Increasing the number of parameters leads to more expressive models thereby enhancing retrieval performance. This result is in accordance with

Table 11: **Evaluation on held-out concentration for active molecules:** Average Top-1% and Top-5% recall accuracies of methods utilizing different contrastive learning loss functions and concentration encoding information. We evaluate all methods on unseen images, unseen images and unseen molecules and an unseen dataset for zero-shot retrieval. Entries in **bold** denote highest performance when the loss function is fixed while entries in highlight denote highest performance across all guidelines.

Method	Explicit Concentration (ours)	Modality	top-1%				top-5%			
			Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
CLIP	✗	Phenom1	**.2109**	**.2425**	**.1519**	**.2018**	**.4458**	**.4968**	**.3591**	**.4339**
Hopfield-CLIP	✗	Phenom1	**.1581**	**.2034**	**.1198**	**.1604**	**.3783**	**.4413**	**.3045**	**.3747**
InfoLOOB	✗	Phenom1	**.2122**	**.2496**	**.1501**	**.2040**	**.4443**	**.5003**	**.3515**	**.4320**
CLOOME	✗	Phenom1	.2164	.2461	.1479	.2035	.4590	.4956	.3528	.4358
CLOOME	sigmoid	Phenom1	.3338	.2681	.1801	.2606	.6037	.5202	.3879	.5039
CLOOME	logarithm	Phenom1	.3094	.2345	.1665	.2368	.5960	.4874	.3534	.4790
CLOOME	one-hot	Phenom1	.3073	.2040	.1670	.2261	.5997	.4246	.3657	.4633
DCL	✗	Phenom1	.4717	.4027	.2841	.3861	.7352	.6579	.5322	.6417
DCL	sigmoid	Phenom1	**.7282**	**.4100**	**.3560**	**.4980**	**.9226**	**.6561**	**.6015**	**.7267**
DCL	logarithm	Phenom1	.6903	.3558	.3211	.4557	.8869	.6146	.5667	.6894
DCL	one-hot	Phenom1	.6562	.3607	.3272	.4480	.8831	.5983	.5659	.6824
CWCL	✗	Phenom1	.5731	.4403	.3232	.4455	.8218	.6833	.5706	.6919
CWCL	sigmoid	Phenom1	**.7780**	**.4425**	**.3777**	**.5327**	**.9386**	**.6844**	**.6244**	**.7491**
CWCL	logarithm	Phenom1	.7452	.3989	.3523	.4988	.9117	.6482	.5962	.7187
CWCL	one-hot	Phenom1	.7048	.4009	.3593	.4883	.9037	.6284	.6061	.7127
SigLip	✗	Phenom1	.5718	.4217	.3021	.4318	.8104	.6602	.5176	.6627
SigLip	sigmoid	Phenom1	**.8366**	**.4640**	**.3830**	**.5612**	**.9623**	**.7023**	**.6080**	**.7575**
SigLip	logarithm	Phenom1	.8097	.4391	.3747	.5411	.9437	.6746	.6046	.7409
SigLip	one-hot	Phenom1	.7561	.4020	.3345	.4975	.9279	.6248	.5557	.7028
MolPhenix (ours)	✗	Phenom1	**.8334**	.4615	**.3792**	.5580	.9638	.6937	**.6128**	.7567
MolPhenix (ours)	sigmoid	Phenom1	.8256	**.4692**	.3765	**.5571**	**.9638**	**.7068**	.6115	**.7607**
MolPhenix (ours)	logarithm	Phenom1	.7953	.4466	.3664	.5361	.9466	.6889	.5924	.7426
MolPhenix (ours)	one-hot	Phenom1	.7489	.4088	.3379	.4985	.9325	.6465	.5644	.7144
MolPhenix (ours)	✗	Phenom1 & MolGPS	**.8277**	.4739	.4071	.5695	**.9602**	.7041	.6798	.7813
MolPhenix (ours)	sigmoid	Phenom1 & MolGPS	.8218	**.4771**	.4287	**.5758**	.9588	.7117	**.7045**	**.7916**
MolPhenix (ours)	logarithm	Phenom1 & MolGPS	.7836	.4757	**.4297**	.563	.9402	**.7138**	.7011	.7850
MolPhenix (ours)	one-hot	Phenom1 & MolGPS	.7391	.4307	.3940	.5212	.9198	.6724	.6698	.7540

Table 12: **Evaluation on cumulative concentrations for active and inactive perturbations** Average Top-1% and Top-5% Recall accuracy of methods utilizing different contrastive learning methods. Best performing methods are highlighted in **bold**.

Loss	Explicit Concentration	Modality	top-1%				top-5%			
			Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
CLIP	✗	Phenom1	.1761	.1867	.0734	.1454	.3710	.3769	.2065	.3181
Hopfield-CLIP	✗	Phenom1	.1531	.1709	.0673	.1304	.3464	.3637	.1942	.3014
InfoLOOB	✗	Phenom1	.1746	.1860	.0745	.1450	.3697	.3756	.2058	.3170
CLOOME	✗	Phenom1	.1968	.2005	.0911	.1628	.3938	.3888	.2321	.3383
CLOOME	sigmoid	Phenom1	.3875	.2592	.1415	.2627	.5662	.4601	.2940	.4401
CLOOME	logarithm	Phenom1	**.4088**	.3046	**.1503**	.2879	.5730	.5166	.3053	.4650
CLOOME	one-hot	Phenom1	.4080	**.3123**	.1496	**.2900**	**.5801**	**.5306**	**.3054**	**.4720**
DCL	✗	Phenom1	.3277	.2562	.1364	.2401	.4856	.4170	.2768	.3931
DCL	sigmoid	Phenom1	.4881	.3380	.2009	.3423	.6222	.5186	.3381	.4930
DCL	logarithm	Phenom1	.4983	.3615	.2122	.3573	.6311	.5581	.3587	.5160
DCL	one-hot	Phenom1	**.5226**	**.3790**	**.2288**	**.3768**	**.6791**	**.5870**	**.3968**	**.5543**
CWCL	✗	Phenom1	.3635	.2696	.1526	.2619	.5122	.4267	.2933	.4107
CWCL	sigmoid	Phenom1	.5070	.3457	.2101	.3542	.6378	.5272	.3462	.5037
CWCL	logarithm	Phenom1	.5146	.3725	.2246	.3706	.6437	.5733	.3660	.5277
CWCL	one-hot	Phenom1	**.5401**	**.3849**	**.2336**	**.3862**	**.6882**	**.5991**	**.4001**	**.5625**
SigLip	✗	Phenom1	.3729	.2544	.1470	.2581	.5200	.4179	.2838	.4072
SigLip	sigmoid	Phenom1	.5021	.3275	.2072	.3456	.6360	.5231	.3430	.5007
SigLip	logarithm	Phenom1	.5156	.3636	.2233	.3675	.6452	.5689	.3653	.5265
SigLip	one-hot	Phenom1	**.5354**	**.3745**	**.2317**	**.3805**	**.6858**	**.5928**	**.3945**	**.5577**
S2L (ours)	✗	Phenom1	.4688	.2852	.1838	.3126	.5970	.4519	.3171	.4554
S2L (ours)	sigmoid	Phenom1	.5071	.3441	.2144	.3552	.6428	.5315	.3554	.5099
S2L (ours)	logarithm	Phenom1	.5183	.3700	.2275	.3720	.6492	.5650	.3756	.5300
S2L (ours)	one-hot	Phenom1	**.5433**	**.3819**	**.2384**	**.3879**	**.6954**	**.5895**	**.4030**	**.5626**
S2L (ours)	✗	Phenom1 & MolGPS	.4688	.2729	.2001	.3139	.5956	.4374	.3430	.4587
S2L (ours)	sigmoid	Phenom1 & MolGPS	.4983	.3230	.2397	.3537	.6343	.5035	.3790	.5056
S2L (ours)	logarithm	Phenom1 & MolGPS	.5101	.3589	.2535	.3742	.6398	.5660	.3992	.5350
S2L (ours)	one-hot	Phenom1 & MolGPS	**.5370**	**.3720**	**.2676**	**.3922**	**.6870**	**.5888**	**.4326**	**.5695**

Table 13: **Evaluation on held-out concentrations for active and inactive perturbations** Average Top-1% and Top-5% Recall accuracy of methods utilizing different contrastive learning methods. Best performing methods are highlighted in **bold**.

			top-1%				top-5%			
Loss	Explicit Concentration	Modality	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
CLIP	✗	Phenom1	.1684	.1111	.0964	.1253	.3916	.2545	.2356	.2476
Hopfield-CLIP	✗	Phenom1	.1290	.0921	.0756	.0989	.3485	.2287	.2095	.2217
InfoLOOB	✗	Phenom1	.1715	.1114	.0948	.1259	.3944	.2578	.2349	.2495
CLOOME	✗	Phenom1	.1745	.1088	.0910	.1248	.4093	.2487	.2355	.2439
CLOOME	sigmoid	Phenom1	**.2573**	**.1208**	**.1062**	**.1614**	**.5169**	**.2638**	**.2513**	**.3440**
CLOOME	logarithm	Phenom1	.2379	.1081	.0992	.1484	.4958	.2444	.2324	.3242
CLOOME	one-hot	Phenom1	.2346	.0970	.0974	.1430	.5014	.2224	.2348	.3195
DCL	✗	Phenom1	.3516	.1655	.1533	.2235	.5693	.3125	.3006	.3082
DCL	sigmoid	Phenom1	**.4741**	**.1725**	**.1726**	**.2731**	**.6637**	**.3261**	**.3105**	**.3204**
DCL	logarithm	Phenom1	.4286	.1596	.1581	.2488	.6244	.3071	.3032	.3056
DCL	one-hot	Phenom1	.4308	.1495	.1600	.2468	.6244	.2938	.3015	.2966
CWCL	✗	Phenom1	.4126	.1801	.1667	.2531	.6128	.3266	.3066	.3194
CWCL	sigmoid	Phenom1	**.5112**	**.1856**	**.1811**	**.2926**	**.6901**	**.3384**	**.3190**	**.3314**
CWCL	logarithm	Phenom1	.4664	.1696	.1709	.2690	.6502	.3195	.3066	.3148
CWCL	one-hot	Phenom1	.4681	.1612	.1734	.2676	.6465	.3019	.3104	.3050
SigLip	✗	Phenom1	.3942	.1578	.1390	.2303	.5931	.3015	.2737	.2914
SigLip	sigmoid	Phenom1	**.5392**	**.1828**	**.1710**	**.2977**	**.7102**	**.3399**	**.3121**	**.3298**
SigLip	logarithm	Phenom1	.5022	.1698	.1669	.2796	.6841	.3240	.3068	.3177
SigLip	one-hot	Phenom1	.4657	.1443	.1451	.2517	.6544	.2879	.2790	.2847
S2L (ours)	✗	Phenom1	.5336	.1842	.1713	.2963	.6961	.3322	.3045	.3221
S2L (ours)	sigmoid	Phenom1	**.5409**	**.1899**	**.1753**	**.3020**	**.7178**	**.3469**	**.3201**	**.3372**
S2L (ours)	logarithm	Phenom1	.5036	.1791	.1727	.2851	.6925	.3342	.3157	.3275
S2L (ours)	one-hot	Phenom1	.4726	.1537	.1521	.2595	.6696	.2998	.2887	.2958
S2L (ours)	✗	Phenom1 & MolGPS	.5248	.1829	.1910	.2996	.6904	.3268	.3305	.3281
S2L (ours)	sigmoid	Phenom1 & MolGPS	**.5338**	**.1897**	**.2029**	**.3088**	**.7098**	**.3427**	**.3495**	**.3452**
S2L (ours)	logarithm	Phenom1 & MolGPS	.4900	.1839	.2031	.2923	.6776	.3354	.3511	.3411
S2L (ours)	one-hot	Phenom1 & MolGPS	.4622	.1569	.1762	.2651	.6578	.3030	.3187	.3087

recent advances in language modelling and scaling laws across different data and compute budgets [21].

Model size	Depth	Width	Unseen images	Unseen images + Unseen molecules	Unseen dataset (0-shot)
Tiny - 2.7m	4 ResBlocks	256	.8337	.7186	.4030
Small - 9.4m	6 ResBlocks	512	.9174	.7352	.4562
Medium - 38.7m	8 ResBlocks	1024	.9430	.7490	.485

Table 14: Ablations across different model sizes. Larger capacity models are found to be more expressive.

Batch size	Unseen images	Unseen images + Unseen molecules	Unseen dataset (0-shot)
128	.8600	.7163	.4044
512	.9252	.7511	.4657
2048	.9450	.7616	.4940
8192	.9489	.7563	.4966

Table 15: Ablation across different batch sizes. Larger batch sizes benefit contrastive learning.

E.6 Investigating Other Pre-trained Phenomic Encoders

To investigate the impact of pre-trained encoders, we perform additional experiments evaluating a supervised phenomic image encoder (Table 20). Instead of Phenom1, we trained Molphenix framework using AdaBN, a CNN-based supervised phenomic encoder, with an analogous implementation discussed in [46]. We find that the general trends between Phenom1 and AdaBN are consistent with a slight decrease in overall performance. These findings provide additional support to the generality of the proposed guidelines.

E.7 Integrating MolGPS Embeddings With Other Fingerprints

Molphenix architecture is flexible, allowing that the proposed components be replaced by other phenomic or molecular pretrained models. We leveraged from MolGPS, which is a MPNN based

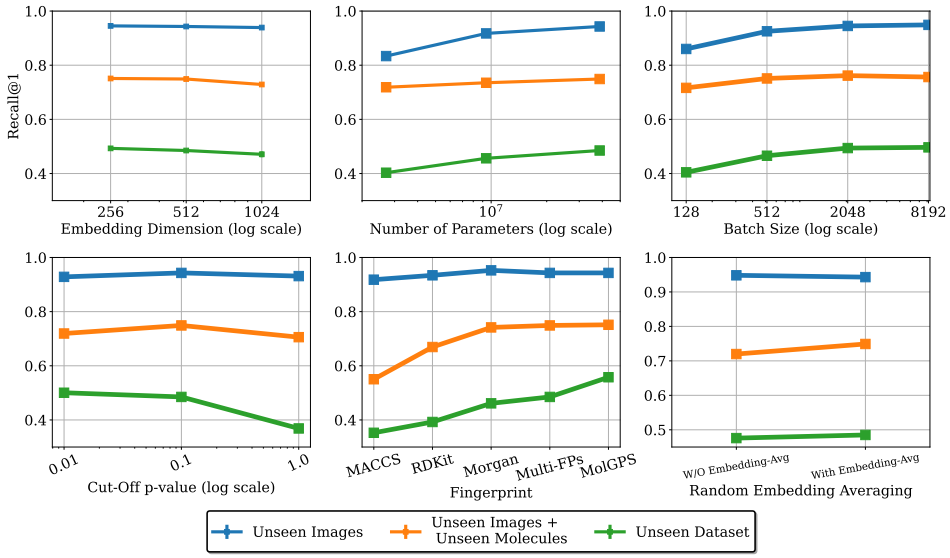


Figure 10: Ablations of top-1 % recall accuracy with **(top-left)** the size of embedding dimension, **(top-center)** number of parameters, **(top-right)** batch size, **(bottom-left)** cutoff p value, **(bottom-center)** fingerprint type, and **(bottom-right)** random batch averaging. Compact embedding sizes from pretrained models, larger number of parameters, larger batch sizes, lower cutoff p-values, pretrained MolGPS fingerprints and presence of random batch averagin improving retrieval of our MolPhenix framework.

Dim size	Unseen images	Unseen images + Unseen molecules	Unseen dataset (0-shot)
256	.9452	.7510	.4929
512	.9430	.7490	.4850
1024	.9392	.7288	.4710

Table 16: Ablation across different embedding dimensions. Compact embedding sizes capture more molecular information.

GNN model with 1B parameters which allows us to maximize architecture expressivity while minimizing the risk of overfitting [29, 47]. For additional investigation, we combine MolGPS molecular embeddings with RDKIT, MACCS, and Morgan fingerprints and show that they can provide Molphenix with richer molecular information and yields overall higher performance of MolPhenix in both cumulative and held-out concentration scenarios. Results for active and all molecules retrieval of Molphenix trained on the discussed combinational molecular embeddings are available in table 21 and 22.

cut-off	Unseen images	Unseen images + Unseen molecules	Unseen dataset (0-shot)
p < 1.0	.9312	.7057	.3686
p < .1	.9430	.7490	.4850
p < .01	.9284	.7192	.5005

Table 17: Ablation across different p-value cutoff threhsolds. p values < .1 benefit retrieval of active molecules.

fingerprint	unseen images	unseen images + unseen molecule	unseen dataset
MACCS	.9180	.5503	.3526
RDKit	.9341	.6693	.3925
Morgan	.9524	.7417	.4613
Multi-FPs	.9430	.7490	.485
Phenom1 + MolGPS	.9430	.7514	.5577

Table 18: Ablation across different fingerprint types. A combination of embeddings bootstrapped from Phenom1 and MolGPS significantly benefit retrieval.

	Unseen images	Unseen images + Unseen molecules	Unseen dataset (0-shot)
W/O Random Embedding Avg.	.9482	.7198	.4759
With Random Embedding Avg.	.9430	.7490	.485

Table 19: Ablation across random embedding averaging. Utilizing random batch averaging stabilizes training and benefits retrieval.

Table 20: Evaluation on **cumulative concentrations** while using **AdaBN**. Molphenix is trained on combination of RDKIT, MACCS, and Morgan fingerprints in this experiment

Method	Explicit Concentration	Modality	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
			top-1% active molecules				**top-5% active molecules**			
MolPhenix	-	AdaBN	.8568	.5336	.3525	.581	.9562	.7603	.5772	.7646
MolPhenix	sigmoid	AdaBN	.911	.5858	.4	.6323	.971	.7997	.6203	.797
MolPhenix	logarithm	AdaBN	.9155	.6106	.4242	.6501	.9729	.8332	.6503	.8188
MolPhenix	one-hot	AdaBN	.9187	.6125	.4225	.6512	.9744	.8302	.6419	.8155
			top-1% all molecules				**top-5% all molecules**			
MolPhenix	-	AdaBN	.4593	.2409	.1599	.2867	.5983	.4081	.285	.4305
MolPhenix	sigmoid	AdaBN	.5104	.3142	.1957	.3401	.6496	.5165	.331	.499
MolPhenix	logarithm	AdaBN	.5379	.3393	.2071	.3614	.6867	.5561	.3606	.5345
MolPhenix	one-hot	AdaBN	.5476	.3425	.2082	.3661	.7007	.5641	.3603	.5417

Table 21: Evaluation on **cumulative concentrations** while **combining MolGPS, RDKIT, MACCS, and Morgan fingerprints**.

Method	Explicit Concentration	Modality	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
			top-1% active molecules				**top-5% active molecules**			
MolPhenix	-	Phenom1 & MolGPS & 3 fps	.9185	.7212	.4717	.7038	.9784	.8805	.718	.859
MolPhenix	sigmoid	Phenom1 & MolGPS & 3 fps	.9395	.7408	.5119	.7307	.9817	.8932	.7458	.8736
MolPhenix	logarithm	Phenom1 & MolGPS & 3 fps	.9454	.7798	.5658	.7637	.9815	.9163	.7849	.8942
MolPhenix	one-hot	Phenom1 & MolGPS & 3 fps	.9419	.7687	.5526	.7544	.9807	.9113	.7681	.8867
			top-1% all molecules				**top-5% all molecules**			
MolPhenix	-	Phenom1 & MolGPS & 3 fps	.4764	.3011	.2068	.3281	.604	.4647	.3415	.4701
MolPhenix	sigmoid	Phenom1 & MolGPS & 3 fps	.5076	.342	.2382	.3626	.6383	.521	.3769	.512
MolPhenix	logarithm	Phenom1 & MolGPS & 3 fps	.525	.379	.2648	.3896	.658	.5743	.411	.5478
MolPhenix	one-hot	Phenom1 & MolGPS & 3 fps	.5355	.3845	.265	.395	.6862	.5916	.4233	.567

Table 22: Evaluation on **heldout concentrations** while **combining MolGPS, RDKIT, MACCS, and Morgan fingerprints**.

Method	Explicit Concentration	Modality	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.	Unseen Images	Unseen Images + Unseen Molecules	Unseen Dataset (0-shot)	Avg.
			top-1% active molecules				**top-5% active molecules**			
MolPhenix	-	Phenom1 & MolGPS & 3 fps	.8364	.5115	.4263	.5914	.9640	.7363	.6850	.7951
MolPhenix	sigmoid	Phenom1 & MolGPS & 3 fps	.8300	.5021	.4363	.5895	.9640	.7409	.6931	.7993
MolPhenix	logarithm	Phenom1 & MolGPS & 3 fps	.8112	.5107	.4376	.5865	.9544	.7406	.6866	.7939
MolPhenix	one-hot	Phenom1 & MolGPS & 3 fps	.7467	.4409	.3830	.5235	.9320	.6827	.6520	.7556
			top-1% all molecules				**top-5% all molecules**			
MolPhenix	-	Phenom1 & MolGPS & 3 fps	.5339	.1980	.1966	.3095	.6968	.2909	.4274	.4717
MolPhenix	sigmoid	Phenom1 & MolGPS & 3 fps	.5463	.2026	.2066	.3185	.7179	.3116	.4359	.4885
MolPhenix	logarithm	Phenom1 & MolGPS & 3 fps	.5247	.2009	.2078	.3111	.7067	.3133	.4319	.4840
MolPhenix	one-hot	Phenom1 & MolGPS & 3 fps	.4690	.1653	.1756	.2700	.6635	.2592	.4118	.4448