Multi-modal Contrastive Learning with Negative Sampling Calibration for Phenotypic Drug Discovery

Supplementary Material

A. Further Implementation Details

A.1. Datasets

To evaluate our model for molecular property prediction, we utilized the ChEMBL2K and Broad6K datasets, following the benchmark settings in InfoAlign [21]. **ChEMBL2K Dataset:** ChEMBL2K is a curated subset of the ChEMBL dataset, overlapping with the JUMP CP datasets. Activity annotations are derived from the "activity comment" field in ChEMBL. To prevent data leakage, molecules included in the pretraining set are excluded. This dataset encompasses 41 tasks related to protein binding affinity, where activity labels are converted to binary values. To ensure sufficient data for each task, we retain only those with at least one positive example and five negative examples. **Broad6K Dataset:** The Broad6K dataset, originally compiled by Moshkov et al., consists of 16,170 molecules assessed across 270 assays, producing a total of 585,439 readouts. However, the dataset contains a substantial number of missing values, with 153 assays having more than 99% missing data. To minimize bias, we focus our analysis on subsets where the proportion of missing values is below 50%.

For molecule-phenotype retrieval, we followed the evaluation strategy used in InfoCORE [41]. We utilize two high-content drug screening datasets for our study: L1000 gene expression profiles (GE) and cell imaging profiles from the Cell Painting assay (CP). For the GE dataset, we focus on data from nine core cell lines, encompassing 17,753 drugs and 82,914 drug-cell line pairs. In the CP dataset, 30,204 small molecules are screened using a single cell line (U2OS). Hand-crafted image features are extracted using the widely adopted CellProfiler method, while chemical structures are represented using Mol2vec embeddings.

Table S1. Dataset Statistics of Molecular Property Predictions

Dataset	Task	Molecules	Gene Expression	Cell Morphology
ChEMBL2K	32	2,355	581	2354
Broad6K	41	6567	3261	6495

A.2. Baselines

- **AttrMask** [15] is a self-supervised pre-training method for Graph Neural Networks designed to learn both local and global representations of graphs. By introducing a self-supervised pre-training strategy that simultaneously learns both local (node-level) and global (graph-level) representations, AttrMask can effectively capture structural and semantic information to provide a generic feature representation for downstream tasks.
- **EdgePred** [11] is a generalization framework designed to efficiently generate node embeddings, especially for dynamic or unseen node scenarios. GraphSAGE achieves the ability to generalize to unseen nodes by learning an aggregation function based on the local neighborhood of a node, rather than training embeddings for each node individually.
- **GROVER** [35] learns rich structural and semantic information from a large amount of unlabeled molecular data at node-level, edge-level, and graph-level, aiming to solve the problems of insufficient molecular data annotation and poor generalization ability to newly synthesized molecules.
- **GraphLoG** [46] is a unified framework for whole-graph self-supervised representation learning, aiming to solve the problem that existing methods fail to capture the global semantic structure of datasets. By introducing hierarchical prototypes, GraphLoG identifies global semantic clusters while maintaining local similarities.
- **JOAO** [48] address the challenge of manual selection for different datasets by automatically and adaptively selecting the appropriate augmentations for specific graph data, significantly enhancing the general applicability of GraphCL, a fairness by minimizing correlation with spurious features or removing sensitive attributes.
- **CLOOME** [36] creates a unified embedding space that allows both bioimages and molecular structures to be encoded together by a multi-modal contrastive learning approach, which enables querying bioimaging databases using chemical structures that can reveal different phenotypic effects.
- **MIGA** [52] is designed to enhance molecular representation learning by leveraging perturbed high-content cell microscopy images at the phenotypic level, utilizing various contrastive loss functions to capture meaningful features.
- **InfoCORE** [41] aims to deal with batch effects and obtain refined molecular representations by adaptively reweighing samples to equalize their implied batch distribution. It has two versions, InfoCORE-CP and InfoCORE-GE, which align molecular graph representation with cell imaging and gene expression, respectively.

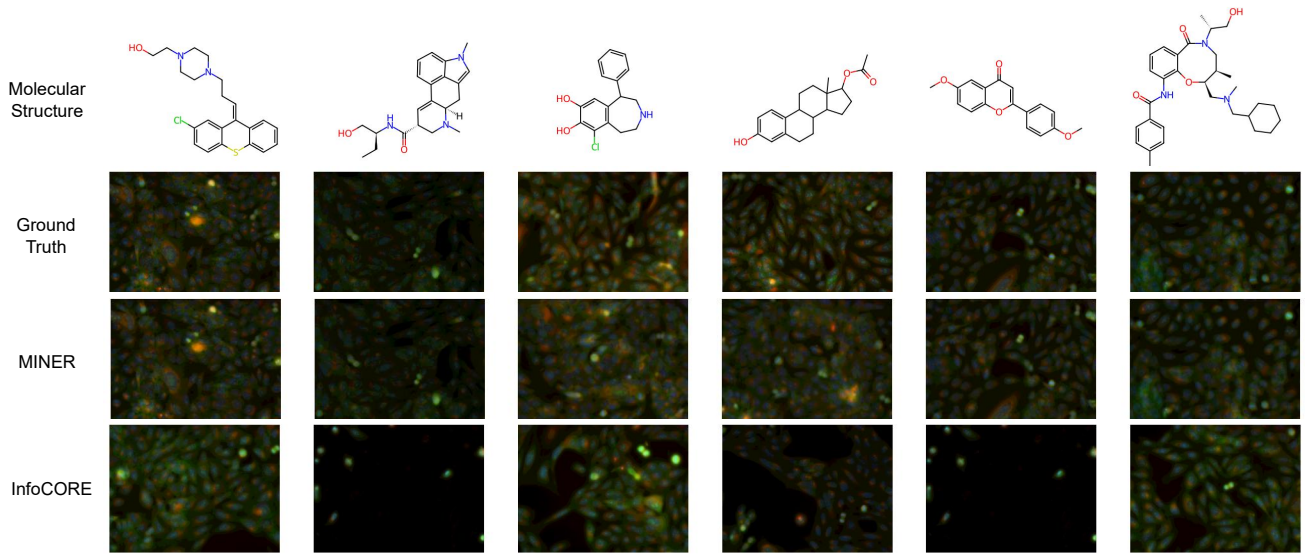

Figure S1. Additional Examples of image retrieval tasks. The query molecules and the top-ranked images being retrieved from different methods and the ground-truth images are shown.

A.3. More Implementation Details

All the experiments were performed five times on an NVIDIA GeForce RTX 4090 GPU. The corresponding Pytorch version and CUDA version are 2.2.2 and 12.1 respectively. Some key parameters are listed as follows:

Table S2. Detailed setting of the hyper-parameter of MINER

Item	ChEMBL2K	Broad6K	GE	CP
λ_1	0.05	0.1	0.001	0.1
λ_2	0.05	0.1	1.0	1.0
λ_3	0.05	0.05	0	0
λ_4	0.05	0.01	0.01	0.01
Beta_1	0.85	0.9	0.5	0.5
Beta_2	0.85	0.999	0.5	0.5
Epoch	120	150	500	500
Dropout	0.5	0.3	0.3	0.25
Batch size	1024	1024	1024	1024
Initial learning rate	0.005	0.0005	0.0005	0.0005

Molecular Property Predictions. We conducted experiments on two molecular property datasets, ChEMBL2K and Broad6K. We employ scaffold splitting for both datasets with a 0.6:0.15:0.25 ratio for training, validation, and test sets. In both datasets, we employ a 3-layer GNNs model to encode molecular graphs, two 5-layer fully connected neural networks to encode the cellular image and gene expression data, and two 5-layer fully connected neural networks to generate corresponding missing data. In ChEMBL2K and Broad6K, we set the hyperparameter $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ as $\{0.05, 0.05, 0.05, 0.05\}$ and $\{0.1, 0.1, 0.05, 0.01\}$, respectively. Besides, temperature coefficient τ is set as 0.07 for contrastive learning. In the training phase, we utilize Adam [7]as the optimizer and the corresponding betas as shown in S2. For the final result, we report means and standard deviations across five runs to ensure consistency.

Molecule-Phenotype Retrieval. In this task, we followed the evaluation strategy used in InfoCORE [41]. Both datasets in molecule-gene(GE) and molecule-cell Painting(CP) were randomly split into a training set with 80% of the molecules and the remaining 20% were held out for testing. As the setup in molecular property predictions, we also use the 3-layer GNNs model and 5-layer fully connected neural networks to encode molecular graphs and gene expression or cellular images. train the classifier to predict the batch number of the sample. Since the dataset does not contain missing modality, we set hyperparameter λ_3 as 0 in both retrieval tasks. The complete $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are set as $\{0.001, 1.0, 0, 0.01\}$ and $\{0.1, 1.0, 0, 0.01\}$ in GE and CP, respectively. We perform 500 training iterations to update the model parameters with Adam [7] as optimizer.

B. Additional Experimental Results

B.1. Image Retrieval

Additional examples for image retrieval tasks, are shown in Fig. S1. The images retrieved by our method and baseline are shown. Our method accurately captures the critical fluorescent areas in images for the first and second molecules. Even for the third and fourth molecules, where MINER

didn't perfectly retrieve the corresponding images, the retrieved images were highly similar to the ground truth. In contrast, the images retrieved by InfoCORE were significantly different from the ground truth, further demonstrating the superior retrieval capability of MINER.

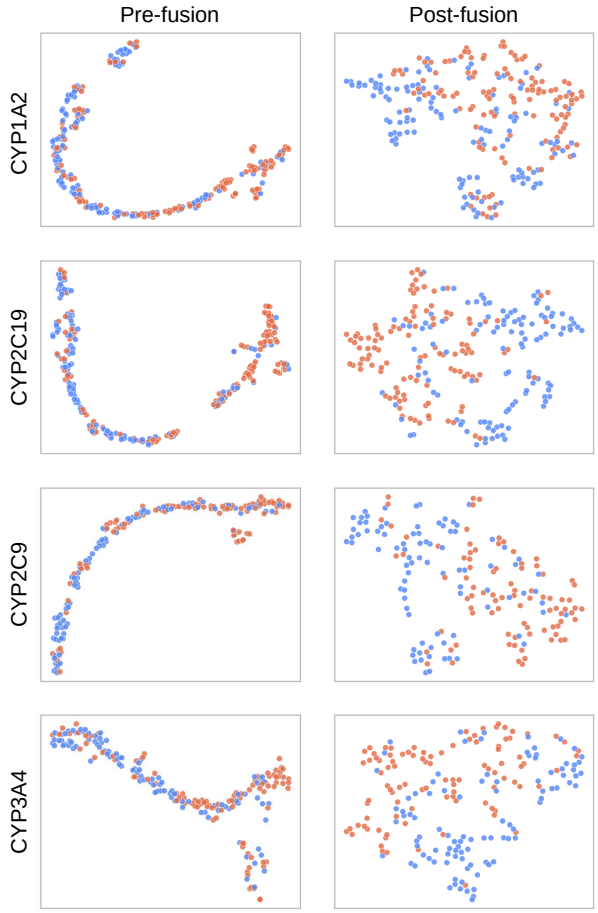

Figure S2. t-SNE projection of features pre-fusion and post-fusion.

B.2. Visualizations

To demonstrate the superiority of our method in integrating multimodal features, we visualized the distribution of representation for pre-fusion cellular image features and the post-fusion multimodal features under different labels. The left column of Fig. S2 presents the pre-fusion cellular image features across four assays, while the right column displays the multimodal features after fusion.

Obviously, the pre-fusion features exhibit significant overlap between positive and negative labels, making them difficult to distinguish. In contrast, the post-fusion multimodal features show significantly enhanced separability across all four assays, effectively distinguishing whether a sample is active in assays. These results highlight the effectiveness of our multimodal fusion strategy in improving feature integration and enhancing the discriminative power of the representations.

C. Drug repurposing

We provide a comprehensive drug recommendation workflow based on the large-scale integration atlas of perturbation profiles. In step 1, given the compounds' structure of the screening library, MINER retrieved the perturbed transcriptional profiles of all candidate compounds. In step 2, the perturbation fold-change and the average fold-change for the transformed expression profile are computed. The gene ranking is then performed based on the fold-change values. In step 3, given the gene signature for a specific disease, we compute the enrichment scores for up- and down-regulated gene sets of screening compounds. Finally, compounds in the screening library are ranked based on these enrichment scores.

After ranking the drugs for breast cancer, literature verification revealed that Sulbactam (rank 1), Norelgestromin (rank 3), Triethylenetetramine (rank 5), Chlorpromazine (rank 7), and Roxithromycin (rank 9) are supported for use in breast cancer treatment (Table S3). A study by Wen et al. [44] on mammalian cells suggested that sulbactam can reduce the expression of ABC transporter proteins in breast cancer cells, thereby decreasing the efflux of doxorubicin and enhancing its efficacy. Another study [27] demonstrated that Norelgestromin acts as a selective estrogen enzyme modulator in human breast cancer cell lines, affecting sulfatase activity in comparison to medroxyprogesterone acetate. Furthermore, Triethylenetetramine [42] has been shown to synergize with pharmacologic Ascorbate (Asc) autoxidation and H2O2 overproduction in breast cancer cells, suppressing the RAS/ERK pathway and inducing apoptosis.

Table S3. Top 10 enrichment scores for drugs against Breast Cancer. Compounds marked in gray have literature support.

Disease	FDA-approved Drug	Enrichment score	Literature
Breast Cancer	Sulbactam	0.3103	[44]
	Troleandomycin	0.2936	
	Norelgestromin	0.2750	[27]
	Perazine	0.2554	
	Triethylenetetramine	0.2546	[42]
	Ethchlorvynol	0.2533	
	Chlorpromazine	0.2481	[17]
	Technetium Tc-99m pyrophosphate	0.2392	
	Roxithromycin	0.2383	[39]
	Boceprevir	0.2377	