

Learning Molecular Representations from Cellular Phenotypes with Structure Preservation

Xuan Lin^a, Jingyu Sheng^a, Tengfei Ma^{b,*}, Li Sun^{c,*}, Dapeng Xiong^d

^a*School of Computer Science, Xiangtan University, Xiangtan, Hunan, China*

^b*College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China*

^c*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China*

^d*State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu, China*

Abstract

Phenotypic drug discovery enables the discovery of functional relationships between molecular structures and cellular responses. However, existing multimodal representation learning methods often optimize cross-modal alignment without considering the intrinsic organization of chemical space, resulting in distorted molecular representations and loss of structural information. We propose **PhenMol**, a structure-preserving framework for phenotype-aware molecular representation learning. PhenMol disentangles molecular and cellular representations into shared and private components, enabling phenotype-guided alignment while preserving chemical structures through a dedicated molecular branch. This design integrates cellular phenotype information without disrupting molecular neighborhood organization. Experiments on approximately 3.04×10^4 molecule–cell morphology pairs demonstrate that PhenMol improves molecular property prediction across 270 bioactivity tasks, molecule–phenotype retrieval, and clinical trial outcome prediction. Moreover, ECFP4-based structural analysis shows that PhenMol better preserves molecular neighborhoods and reduces embedding distortion compared with existing multimodal alignment methods. These results highlight the importance of structure-aware constraints in multimodal

*Corresponding authors.

Email addresses: `tfma@hnu.edu.cn` (Tengfei Ma), `lsun@bupt.edu.cn` (Li Sun)

molecular representation learning and provide an effective approach for integrating cellular phenotypes with chemical knowledge for drug discovery.

Keywords: Phenotypic drug discovery, Molecular representation learning, Cell Painting, Multimodal learning, Bioactivity prediction

1. Introduction

Molecular representation learning underpins computational drug discovery [1], supporting tasks such as virtual screening [2], property prediction [3], target identification [4], and clinical outcome assessment [5]. Existing approaches, including molecular fingerprints [6], graph neural networks [7], and SMILES-based models [8], primarily learn from intrinsic chemical structure and have been strengthened by self-supervised pretraining, including contextual property and motif prediction in GROVER [9], graph augmentation in GraphCL [10] and JOAO [11], molecular graph contrastive learning in MolCLR [12], and knowledge-guided masked graph modeling in KPGT [13]. Although these representations provide strong chemical priors, they do not directly capture how molecules perturb living cells, motivating the integration of cellular phenotypes into molecular representation learning.

Phenotypic drug discovery offers a complementary route by directly profiling cellular responses to small molecules, enabling the discovery of therapeutically relevant phenotypes and mechanisms that may be difficult to infer from chemical structure alone [14]. High-content imaging platforms such as Cell Painting [15, 16] capture morphological changes [17, 18] across multiple cellular compartments and provide rich functional readouts of compound action. These advances have motivated phenotype-aware molecular representation learning, which integrates cellular phenotypes with molecular structure to produce biologically informed representations for drug discovery [19]. The key challenge is to exploit this biological context without disrupting the chemical organization required for molecular generalization.

However, incorporating cellular phenotypes into molecular representation learning creates a fundamental tension between biological alignment and chemical fidelity. Existing methods typically align molecular and cellular embeddings through global contrastive objectives [20, 21]. Recent work on molecular graph contrastive learning [3] and biomedical image–text fusion [22] further highlights the broader need to learn informative representations while controlling cross-view interference. Yet cellular images are high-dimensional and continuous measurements affected by batch effects, staining variation, cell density, and imaging artifacts [23, 24], whereas molecular representations encode discrete structural priors defined by scaffolds, functional groups, and topology [6, 25]. Forcing these heterogeneous modalities into a shared space can cause molecular embeddings to over-assimilate phenotype-specific variation, disrupting the organization of the chemical representation space. Chemically similar molecules may drift apart, while structurally distinct molecules may become artificially close. We refer to this failure mode as *molecular structure collapse* (Fig.1a), in which a biological context is incorporated at the expense of chemical fidelity.

Existing phenotype-aware methods improve molecule–cell correspondence but do not explicitly prevent this collapse. CLOOME [26] relies primarily on global molecule–image contrastive learning, while MIGA [27] and MINER [28] introduce auxiliary matching, robustness, or noise-mitigation objectives. Related multimodal representation studies also show that cross-view fusion and contrastive alignment can improve biomedical pattern recognition when modality-specific interference is controlled [22, 29]. Although these strategies strengthen cross-modal learning, they mainly optimize alignment quality and do not explicitly separate phenotype-related signals from intrinsic molecular structure (Fig.1b). Consequently, preserving chemically meaningful molecular organization during phenotype integration remains an open challenge.

To address this challenge, we introduce **PhenMol**, a structure-preserving framework for phenotype-aware molecular representation learning. PhenMol decomposes

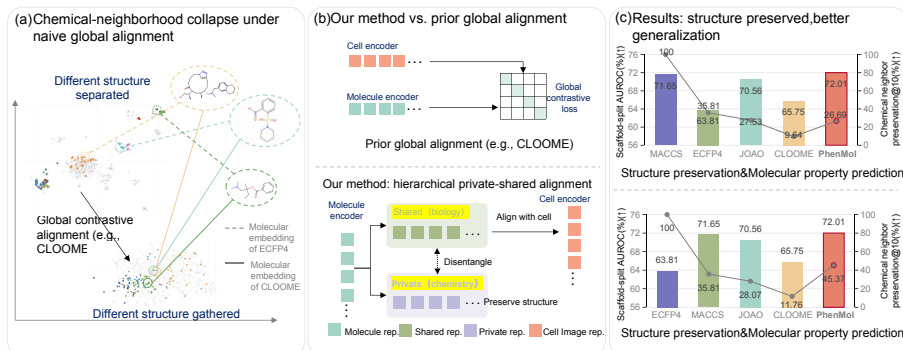


Figure 1: Motivation for structure-preserving phenotype-aware molecular representation learning. (a) Naive global contrastive alignment can cause *molecular structure collapse*, disrupting chemically meaningful local organization under noisy phenotypic supervision. (b) PhenMol separates shared cross-modal signals from modality-private chemical and image factors to preserve molecular structure during alignment. (c) PhenMol quantitatively retains more ECFP4-defined chemical neighbors and improves scaffold-split bioactivity prediction, showing that structural preservation is critical for structure-sensitive generalization.

molecular and cellular representations into modality-shared and modality-private components, discourages batch-specific variation from dominating the shared cellular space, and preserves scaffold-level chemical information in a dedicated molecular private branch. Combined with contrastive alignment and latent-space pair matching, this design integrates cellular phenotypic signals while maintaining chemically meaningful molecular organization. We evaluate PhenMol on approximately 3.04×10^4 paired molecule–cell morphology samples across molecular property prediction, bidirectional molecule–image retrieval, clinical trial outcome prediction, ablation studies, and structural preservation analysis. As shown in Fig. 1c, PhenMol preserves substantially more ECFP4-defined molecular neighbors than prior global alignment methods while achieving the best scaffold-split bioactivity prediction performance. This result shows that maintaining chemical organization during phenotype integration directly benefits structure-sensitive generalization, rather than serving only as an auxiliary representation constraint. The main contributions of this work are:

- We identify **molecular representation collapse under phenotype alignment** as a critical challenge in phenotype-aware molecular representation learning,

where global multimodal alignment can compromise chemical neighborhood preservation.

- We introduce **PhenMol**, a structure-preserving multimodal framework that disentangles shared biological signals from private chemical information to integrate cellular phenotypes without sacrificing molecular identity.
- We demonstrate through molecular prediction, retrieval, clinical transfer, and structural analyses that **PhenMol** learns phenotype-aware molecular representations that are both biologically informative and chemically faithful.

2. Related Work

2.1. Phenotype-Aware Molecular Representation Learning

Molecular representations based on fingerprints, graph neural networks, and molecular Transformers provide strong chemical priors for drug discovery and have been further improved through self-supervised objectives, including motif prediction in GROVER [9], graph augmentation in GraphCL [10] and JOAO [11], molecular graph contrastive learning in MolCLR [12] and LEMON [3], knowledge-guided masked graph modeling in KPGT [13], structure-enhanced protein–ligand representation learning [30], and multi-view learning with structured or unstructured knowledge [31]. However, these approaches mainly characterize intrinsic molecular structure and do not directly capture how compounds perturb living cells. Phenotypic drug discovery complements structure-based modeling by measuring compound-induced biological responses. High-content imaging and Cell Painting [15, 16] capture morphology changes [17, 18] across cellular compartments, while transcriptomic profiles such as L1000 provide pathway-level response signatures [19]. These readouts can reveal drug effects that are difficult to infer from chemical structure alone, but they also contain batch effects, staining variation, cell-density differences, and other assay-specific

noise. The resulting challenge is therefore not merely to incorporate cellular phenotypes into molecular representations, but to exploit their biological information without distorting the chemical organization required for molecular similarity, activity prediction, and scaffold-level generalization. This motivates PhenMol, which explicitly balances phenotype integration with molecular structure preservation.

2.2. Cell–Molecule Alignment with Structural Preservation

Existing phenotype-aware methods typically align molecular and cellular representations in a shared latent space. CLOOME [26] adopts CLIP-style [21] global image–molecule contrastive learning. MIGA [27] introduces auxiliary matching or reconstruction objectives, and MINER [28] improves robustness to noisy phenotypic supervision through calibrated alignment strategies. Related biomedical multimodal learning studies further show that image–text fusion and attention-guided contrastive views can improve representation quality under heterogeneous inputs [22, 29]. These methods demonstrate the value of cellular morphology for molecular representation learning. Nevertheless, most prior work primarily optimizes cross-modal similarity or robustness, without explicitly preserving the structural geometry of the molecular embedding space. Because cellular images contain batch effects, staining variation, and other phenotype-specific noise, unrestricted alignment may cause chemically similar molecules to drift apart and structurally distinct molecules to become artificially close. PhenMol addresses this limitation by separating modality-shared biological information from modality-private features and preserving scaffold-level chemical priors during alignment. It therefore learns molecular representations that remain chemically faithful while incorporating informative cellular responses.

3. Method

We propose **PhenMol**, a structure-preserving multimodal framework for molecular representation learning in phenotypic drug discovery. Given paired molecular

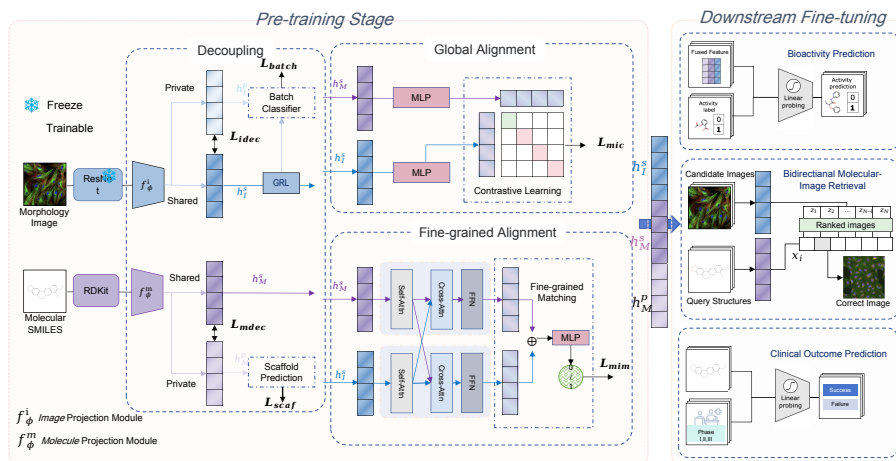


Figure 2: Overview of PhenMol. During pre-training, PhenMol projects molecular fingerprints and cellular phenotype features into a common latent space and decomposes each modality into shared and private representations. The shared branches support global molecule-phenotype contrastive alignment and fine-grained pair matching. In contrast, the private branches retain modality-specific information, including scaffold-level chemical structure and batch-related cellular variation. The resulting representations are transferred to downstream bioactivity prediction, bidirectional molecule-image retrieval, and clinical trial outcome prediction.

inputs and cellular phenotype features, PhenMol integrates drug-induced biological responses into molecular representations while retaining chemically meaningful structural priors. As shown in Fig.2, PhenMol consists of three components: shared-private feature decoupling, global molecule-phenotype alignment, and fine-grained latent pair matching.

3.1. Problem Formulation

Given paired molecular structures and cellular phenotypes, PhenMol aims to learn a molecular representation that is both phenotype-aware and chemically faithful. The paired pre-training dataset is defined as

$$\mathcal{D} = \{(\mathbf{v}_i, \mathbf{m}_i)\}_{i=1}^N, \quad (1)$$

where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ denotes the pre-extracted cellular phenotype feature and $\mathbf{m}_i$ denotes the corresponding molecule. Each molecule $\mathbf{m}_i$ is converted into a 2048-dimensional

ECFP4 fingerprint $\mathbf{x}_i^m \in \{0, 1\}^{2048}$ using RDKit. PhenMol first learns modality-specific encoders

$$\mathbf{z}_i^v = f_v(v_i), \quad \mathbf{z}_i^m = f_m(\mathbf{x}_i^m), \quad (2)$$

where $\mathbf{z}_i^v, \mathbf{z}_i^m \in \mathbb{R}^d$ are cellular and molecular latent embeddings. The first requirement is paired cross-modal alignment: the phenotype embedding and molecular embedding from the same perturbation should be closer than mismatched pairs. This can be written as

$$\text{sim}(\mathbf{z}_i^v, \mathbf{z}_i^m) > \text{sim}(\mathbf{z}_i^v, \mathbf{z}_j^m), \quad i \neq j. \quad (3)$$

The second requirement is molecular structure preservation: if two molecules are structurally similar under ECFP4-defined similarity, their learned molecular embeddings should also remain close. We express this relationship as

$$\text{sim}_{\text{ECFP4}}(\mathbf{x}_i^m, \mathbf{x}_j^m) \uparrow \Rightarrow \text{sim}(\mathbf{z}_i^m, \mathbf{z}_j^m) \uparrow. \quad (4)$$

The overall learning problem therefore balances phenotype alignment and chemical preservation:

$$\min_{\Theta} \mathcal{J} = \mathcal{L}_{\text{align}} + \lambda_{\text{chem}} \mathcal{L}_{\text{chem}}, \quad (5)$$

where Θ denotes all learnable parameters, $\mathcal{L}_{\text{align}}$ optimizes paired molecule–phenotype correspondence, $\mathcal{L}_{\text{chem}}$ constrains the molecular embedding space to retain chemical organization, and λ_{chem} controls the trade-off between the two goals. After pre-training, the learned representations are used for three downstream settings: bioactivity prediction, molecule–phenotype retrieval, and clinical result prediction.

3.2. Framework Overview

As illustrated in Fig.2, PhenMol first projects molecular fingerprints and cellular phenotype features into a common latent space, and then decomposes each modality into shared and private representations. The shared branches capture cross-modal

biological signals, while the private molecular branch preserves chemical structure and the private cellular branch retains modality-specific variation, including batch-related effects. PhenMol performs alignment at two complementary levels. Global contrastive learning aligns paired molecular and cellular representations in the shared space, whereas latent pair matching further models their instance-level compatibility through bidirectional interaction. The resulting representations are used for downstream molecular prediction and molecule-phenotype retrieval.

3.3. Pre-training Stage

The pre-training stage is designed to integrate cellular phenotype information without allowing global alignment to overwrite chemically meaningful molecular organization. The process starts from the projected molecular and cellular embeddings, z^m and z^v , and first decouples each modality into shared and private representations. The shared branches provide the interface for phenotype–molecule correspondence, whereas the private branches retain modality-specific factors, including molecular scaffold information and cellular batch-related variation. Global alignment is then applied to the shared molecular and cellular representations so that paired perturbations are close at the representation level. Finally, fine-grained matching further examines each candidate molecule–phenotype pair after latent interaction, providing instance-level compatibility evidence beyond global similarity. In this way, decoupling defines what information can be aligned, global alignment organizes the shared cross-modal space, and fine-grained matching refines pair-level correspondence while the private branches constrain structure and nuisance variation.

Decoupling. Given projected cellular and molecular features, z^v and z^m , PhenMol decomposes each modality into shared and private representations:

$$\mathbf{h}_v^s, \mathbf{h}_v^p = g_v(z^v), \quad \mathbf{h}_m^s, \mathbf{h}_m^p = g_m(z^m). \quad (6)$$

The shared representations, $\mathbf{h}_v^s$ and $\mathbf{h}_m^s$, capture biological signals that are comparable across modalities, whereas the private representations, $\mathbf{h}_v^p$ and $\mathbf{h}_m^p$, retain modality-specific information. This separation is necessary because cell morphology contains useful perturbation responses but also batch effects and phenotype-specific nuisance variation. The molecular private branch reserves capacity for scaffold-level chemical information, while the cellular private branch absorbs batch-related variation that should not define molecular similarity. To reduce redundancy between the two spaces, we impose an orthogonality constraint:

$$\mathcal{L}_{\text{ortho}} = \ell(\mathbf{h}_v^s, \mathbf{h}_v^p) + \ell(\mathbf{h}_m^s, \mathbf{h}_m^p), \quad (7)$$

where $\ell(\cdot, \cdot)$ is the average squared inner product between normalized feature vectors. For cellular phenotypes, a private batch classifier encourages batch-related variation to remain in $\mathbf{h}_v^p$, while a gradient-reversal classifier discourages such information from dominating $\mathbf{h}_v^s$ [32]. For molecules, a scaffold prediction objective is applied to $\mathbf{h}_m^p$ to preserve scaffold-level chemical information.

Global alignment. Global molecule–phenotype alignment is performed in the shared space. Paired samples are treated as positives, whereas unpaired samples within the mini-batch serve as negatives [29, 33]. After normalization, the similarity between the i -th phenotype and the j -th molecule is computed as

$$S_{ij} = \frac{(\bar{\mathbf{h}}_{v,i}^s)^\top \bar{\mathbf{h}}_{m,j}^s}{\tau}, \quad (8)$$

where τ is a learnable temperature parameter. The resulting bidirectional contrastive objective is

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \left[\text{CE}(\mathbf{S}, \mathbf{y}) + \text{CE}(\mathbf{S}^\top, \mathbf{y}) \right], \quad (9)$$

where $\mathbf{y}$ contains the diagonal pairing labels. This objective organizes the shared embedding space so that matched molecule–phenotype pairs are globally closer than un-

matched pairs. However, because it compares only global vectors, it does not explicitly assess the compatibility of individual molecule–phenotype pairs after cross-modal interaction.

Fine-grained matching. To complement global alignment, PhenMol introduces a latent pair-matching module. The shared molecular and cellular representations are treated as single-token sequences and processed by self-attention, bidirectional cross-attention, and feed-forward layers. Because each modality is represented by a single latent token, this module performs instance-level interaction rather than atom-level or image-patch-level co-attention. Let $\mathbf{r}_{ij}$ denote the interacted representation of a candidate molecule–phenotype pair. A binary matching head predicts

$$p_{ij} = \sigma(\phi(\mathbf{r}_{ij})), \quad (10)$$

where p_{ij} is the predicted probability that molecule j matches phenotype i , $\phi(\cdot)$ is the pair classifier, and $\sigma(\cdot)$ is the sigmoid function. True molecule–phenotype pairs are used as positives, while shuffled pairs provide negatives. The matching loss is

$$\mathcal{L}_{\text{itm}} = -\frac{1}{B} \sum_{i=1}^B [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad (11)$$

where B is the batch size and $y_i \in \{0, 1\}$ is the pair label. This module checks pair compatibility after latent interaction and therefore complements the global contrastive objective.

3.4. Fine-tuning and downstream prediction

After pre-training, PhenMol is adapted to downstream tasks using task-specific combinations of the learned shared and private representations.

Bioactivity prediction. For molecular bioactivity prediction, the model is fine-tuned on a dataset containing more than 16,100 compounds and 270 binary assay tasks. The prediction head concatenates the molecular shared representation, molecular private

representation, and paired cellular shared representation:

$$\hat{\mathbf{y}}_i = \psi_{\text{bio}}([\mathbf{h}_{m,i}^s; \mathbf{h}_{m,i}^p; \mathbf{h}_{v,i}^s]), \quad (12)$$

where $\psi_{\text{bio}}(\cdot)$ is the bioactivity classifier and $\hat{\mathbf{y}}_i$ contains assay-specific activity probabilities. This task evaluates whether phenotype-aware molecular representations improve activity prediction under both random and scaffold splits.

Modal retrieval. For molecule–phenotype retrieval, no additional classifier is introduced. Molecular and cellular embeddings are extracted from the held-out 10% paired samples and compared through the full cross-modal similarity matrix:

$$\mathbf{R}_{ij} = \text{sim}(\mathbf{h}_{v,i}^s, \mathbf{h}_{m,j}^s). \quad (13)$$

Image-to-molecule retrieval ranks candidate molecules for each phenotype query, whereas molecule-to-image retrieval ranks cellular phenotype profiles for each molecular query. This task directly evaluates cross-modal correspondence learned during pre-training.

Clinical result prediction. For clinical result prediction, Phase I, Phase II, and Phase III are treated as independent binary classification tasks. Because cellular images are not available at downstream clinical inference, the classifier uses only molecular shared and private representations:

$$\hat{\mathbf{y}}_i^{(q)} = \psi_{\text{clin}}^{(q)}([\mathbf{h}_{m,i}^s; \mathbf{h}_{m,i}^p]), \quad q \in \{\text{I, II, III}\}. \quad (14)$$

This setting tests whether the learned molecular representation transfers beyond paired Cell Painting readouts to clinically relevant outcome prediction. Together, the three downstream tasks evaluate bioactivity prediction, cross-modal retrieval, and clinical transfer.

3.5. Model Training Objective

PhenMol is trained by jointly optimizing cross-modal alignment, pair-level matching, shared-private decoupling, and modality-specific preservation:

$$\begin{aligned}\mathcal{L}_{\text{total}} = & w_{\text{itc}}\mathcal{L}_{\text{itc}} + w_{\text{itm}}\mathcal{L}_{\text{itm}} + w_{\text{ortho}}\mathcal{L}_{\text{ortho}} \\ & + w_{\text{batch}}^p\mathcal{L}_{\text{batch}}^p + w_{\text{batch}}^s\mathcal{L}_{\text{batch}}^s + w_{\text{scaf}}\mathcal{L}_{\text{scaf}}.\end{aligned}\tag{15}$$

Here, $\mathcal{L}_{\text{itc}}$ and $\mathcal{L}_{\text{itm}}$ promote global and pair-level molecule–phenotype correspondence, $\mathcal{L}_{\text{ortho}}$ reduces redundancy between shared and private representations, the batch-related objectives encourage nuisance variation to remain outside the shared cellular space, and $\mathcal{L}_{\text{scaf}}$ preserves scaffold-level molecular information. All loss weights are fixed during pre-training, while the gradient-reversal coefficient may be progressively increased to stabilize adversarial optimization. The contribution of each objective is examined through ablations on feature decoupling, pair matching, and private-branch supervision.

4. Experiments

4.1. Experimental Setup

We evaluate PhenMol along five complementary axes: molecular property prediction, component ablation, molecule–image bidirectional retrieval, clinical trial outcome prediction, and structural preservation. These evaluations test whether the learned representation improves downstream prediction, supports cross-modal matching, transfers to a clinically relevant task, and retains chemically meaningful neighborhoods.

Pre-training and downstream data. PhenMol is pre-trained on approximately 3.04×10^4 valid molecule–cell morphology pairs from JUMP-CP and the Broad Drug Repurposing Hub [34, 35]. For evaluation, we use: (i) 270 bioactivity-assay tasks for property prediction [19]; (ii) a held-out test library for molecule–image retrieval; and (iii)

Phase I–III clinical trial outcomes as independent binary classification tasks following TOP/HINT [27, 36].

Implementation and training. SMILES are featurized as 2048-bit ECFP4 fingerprints via RDKit [37], while cellular images are encoded offline via a frozen ResNet-50 into 512-dimensional vectors.

Evaluation protocols and baselines. We adopt both random and scaffold splits with a target 8:1:1 train/validation/test ratio. The random split yields 24,323 training, 3,040 validation, and 3,041 test samples. The scaffold split yields 24,985 training, 2,699 validation, and 2,720 test samples; the slight deviation from the target ratio is due to grouping by molecular scaffold. Reporting protocols are task-dependent. For property prediction, we report mean $\pm$ standard deviation over five seeds under random and scaffold splits. Clinical tasks use stratified five-fold cross-validation, and retrieval is evaluated once on the full test library. We compare against a comprehensive set of baselines, including (1) single-modality models trained on Cell Painting profiles [15, 16, 19], ResNet features [38], MACCS keys [39], and RDKit/Morgan fingerprints [6, 37]; (2) multimodal alignment models such as CLOOME [26], MIGA [27], and MINER [28]; and (3) LR, XGBoost, and HINT for clinical tasks [27, 36]. All methods are evaluated under identical splits and protocols, and modality differences in baseline implementations are controlled to ensure fair comparison. For reproducibility, the five property-prediction seeds are 0, 10, 20, 30, and 40; all other task-specific reporting units follow the protocols described above.

Hardware and main hyperparameters. Experiments were conducted with PyTorch 2.8.0 and Python 3.12 on Ubuntu 22.04 with CUDA 12.8, using one NVIDIA RTX 4090 GPU with 24GB memory, a 16-vCPU Intel(R) Xeon(R) Gold 6430 CPU, and 120GB RAM. PhenMol uses latent dimension $d = 512$ and batch size 64, and is trained for at most 250 epochs. Pre-training uses AdamW with learning rate 4×10^{-4} and weight decay 0.01, while activity-prediction fine-tuning uses AdamW with learning rate 1×10^{-4} ; retrieval is evaluated without additional optimization. The main

loss weights are $w_{\text{itc}} = 1.0$, $w_{\text{itm}} = 2.0$, $w_{\text{ortho}} = 0.5$, $w_{\text{batch}}^p = w_{\text{batch}}^s = 0.2$, and $w_{\text{scaf}} = 0.2$. The GRL coefficient is fixed to $\alpha = 1.0$ by default, and the learnable temperature τ is initialized to 0.07 and clipped to [0.001, 0.5]. The downstream classifier takes a $3d = 1536$ -dimensional input and uses one 256-unit hidden layer with LayerNorm, GELU, and dropout 0.1. Early stopping uses patience 10, monitoring training ITC loss during pre-training and validation mean ROC-AUC during activity-prediction fine-tuning.

Evaluation metrics. Molecular property prediction is evaluated by AUC, AUP, and $\text{AUC} > 0.8/0.85/0.9$ thresholds. Retrieval is evaluated by Hit@1/5/10 and mean reciprocal rank (MRR). Clinical trial outcome prediction is evaluated by AUC and PR-AUC. Structural preservation is evaluated by R-Precision, mean absolute rank shift, and top- K Jaccard using ECFP4-defined molecular neighborhoods.

4.2. Molecular Property Prediction

We conducted experiments on 270 bioactivity-assay tasks to test the model generalization under both random and scaffold splits. As shown in Table 1, PhenMol achieves the best performance across all reported metrics under both random and scaffold splits. Under the random split, it improves AUC from 69.57 to 73.88 (**+4.31**) and AUP from 38.42 to 42.09 (**+3.67**) compared with the strongest multimodal baseline (i.e., CLOOME). The gains are statistically significant with $p < 0.05$ for AUC and AUP, respectively. Under the more challenging scaffold split, PhenMol maintains an AUC of 72.01 and AUP of 40.00, outperforming the strongest baseline by 5.86% and 8.94%, respectively, and this advantage grows substantially under scaffold split. This widening gap suggests that structure-preserving alignment confers particular benefits when the model must generalize beyond its training chemical distribution.

The differential improvements in AUC and AUP can be attributed to our model’s enhanced integration of multimodal data, while scaffold-split results reveal the importance of preserving chemical structure. The AUC gain reflects improved overall rank-

Table 1: Performance comparison on molecular property prediction tasks. Results are reported as mean $\pm$ standard deviation over five seeds. The best results are highlighted in bold, and the second-best results are underlined. The p-value row reports two-sided Welch’s t-test results comparing PhenMol with the strongest baseline for each metric within the same split.

Type	Method	AUC	AUP	AUC>0.8	AUC>0.85	AUC>0.9
Single mode	CellProfiler	61.99 $\pm$ 2.3	29.07 $\pm$ 0.5	13.63 $\pm$ 3.4	11.63 $\pm$ 3.5	10.07 $\pm$ 2.9
	ResNet	68.43 $\pm$ 1.9	31.70 $\pm$ 2.0	22.96 $\pm$ 0.6	18.67 $\pm$ 0.8	14.89 $\pm$ 0.4
	MACCS Keys	63.70 $\pm$ 1.1	26.52 $\pm$ 1.9	18.44 $\pm$ 1.8	13.11 $\pm$ 1.4	9.26 $\pm$ 1.2
	RDKit FP	64.69 $\pm$ 1.2	27.78 $\pm$ 2.6	23.93 $\pm$ 1.1	19.70 $\pm$ 1.1	14.96 $\pm$ 1.2
	Morgan FP	63.51 $\pm$ 1.7	29.67 $\pm$ 2.1	24.22 $\pm$ 1.2	19.26 $\pm$ 0.0	14.81 $\pm$ 0.5
Random split	Moshkov	68.35 $\pm$ 1.9	31.92 $\pm$ 2.7	27.07 $\pm$ 2.2	18.46 $\pm$ 2.1	11.55 $\pm$ 1.4
	CLOOME	<u>69.57 $\pm$ 2.1</u>	<u>38.42 $\pm$ 2.1</u>	26.00 $\pm$ 1.6	<u>21.63 $\pm$ 2.9</u>	<u>15.15 $\pm$ 1.1</u>
	MIGA	63.38 $\pm$ 3.0	27.93 $\pm$ 4.0	<u>27.57 $\pm$ 3.9</u>	20.18 $\pm$ 3.2	13.92 $\pm$ 2.5
	MINER	66.45 $\pm$ 4.7	31.81 $\pm$ 6.1	20.00 $\pm$ 5.2	14.74 $\pm$ 5.0	10.29 $\pm$ 5.1
	PhenMol	73.88 $\pm$ 0.6	42.09 $\pm$ 0.4	28.44 $\pm$ 0.7	22.59 $\pm$ 0.8	17.63 $\pm$ 1.1
	<i>p-value</i>	0.0082	0.0162	0.0231	0.5100	0.0074
Scaffold split	Moshkov	<u>66.15 $\pm$ 2.3</u>	<u>31.06 $\pm$ 2.9</u>	22.81 $\pm$ 2.0	<u>17.26 $\pm$ 1.9</u>	<u>10.89 $\pm$ 1.7</u>
	CLOOME	65.75 $\pm$ 0.2	28.72 $\pm$ 0.2	<u>25.11 $\pm$ 0.7</u>	13.85 $\pm$ 0.6	6.67 $\pm$ 0.3
	MIGA	55.68 $\pm$ 1.1	20.62 $\pm$ 0.6	10.28 $\pm$ 1.0	7.71 $\pm$ 1.1	4.90 $\pm$ 0.7
	MINER	64.33 $\pm$ 1.4	25.09 $\pm$ 1.2	18.74 $\pm$ 1.3	12.00 $\pm$ 1.6	8.81 $\pm$ 0.9
	PhenMol	72.01 $\pm$ 1.0	40.00 $\pm$ 0.6	26.81 $\pm$ 1.1	22.00 $\pm$ 0.8	17.70 $\pm$ 0.8
	<i>p-value</i>	0.0324	0.0291	0.0536	0.0346	0.0095

ing across active and inactive compounds, consistent with the use of shared phenotype information and latent pair matching to inject biological response signals. The larger AUP gain is diagnostically significant for drug discovery, as AUP emphasizes performance on rare positive examples, and this capability is critical for hit identification. The scaffold split offers a more rigorous test because test molecules are chemically distinct from training molecules at the scaffold level. PhenMol’s larger margin in this setting underscores a key design principle. For phenotype-aware molecular representation learning, the critical challenge is not simply whether modalities are aligned, but whether the alignment preserves chemically transferable structure under scaffold shift. The scaffold-preserving private branch appears to serve this function by maintaining chemical identity that might otherwise be overwritten by phenotype-specific noise during alignment. At the same time, the observation that several phenotype-aware baselines do not consistently exceed single-modality performance suggests that

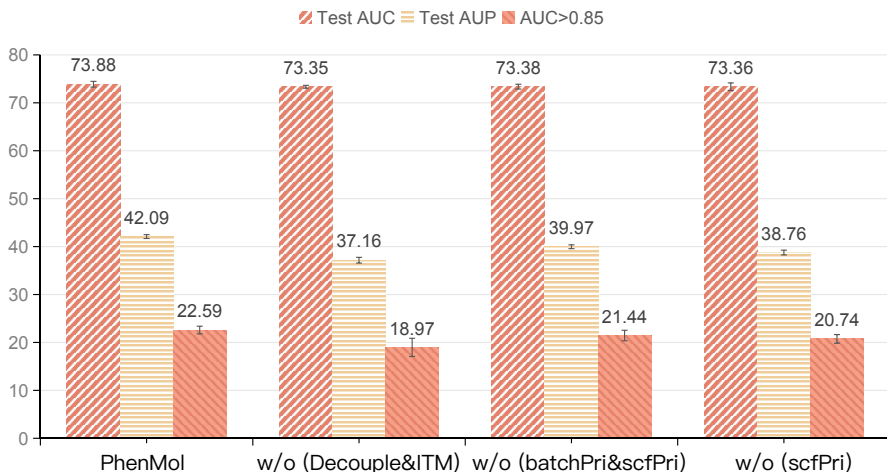


Figure 3: Ablation study of key modules in PhenMol.

global cross-modal alignment can import noise or weaken molecular structure when two modalities are forced into a single geometry without appropriate structural constraints.

4.3. Ablation Study

We conducted ablation studies on the 270 bioactivity-assay tasks to investigate factors influencing performance. Fig.3 reports results for different model variants: **PhenMol w/o Decouple&ITM** removes both feature decoupling and image–molecule matching; **PhenMol w/o batchPri&scfPri** removes the private-branch auxiliary supervision for image batch prediction and molecular scaffold prediction; **PhenMol w/o scfPri** removes only the molecular scaffold prediction constraint; and PhenMol denotes the complete model.

The ablation results show that PhenMol’s gain comes from the interaction among decoupling, pair-level matching, and private-branch constraints rather than from a single auxiliary loss. Removing feature decoupling slightly increases average AUC, from 73.88 to 74.99, but reduces AUP from 42.09 to 39.31 and lowers high-AUC task coverage. This discrepancy indicates that average AUC alone can overstate performance

when positive compounds are sparse: the model may rank negatives well overall while becoming less reliable at identifying rare active compounds. Removing ITM produces a similar but milder pattern, showing that global contrastive alignment alone is not sufficient for stable pair-level discrimination.

The strongest degradation appears when decoupling and ITM are removed together, reducing AUP to 37.16 and lowering the proportion of tasks with AUC > 0.8 to 24.79%. This result supports the design in Fig.2: global alignment and fine-grained matching are complementary, but they need a decoupled representation space so that phenotype information does not overwrite molecular structure. Removing private-branch constraints also reduces AUP, especially when scaffold prediction is removed. This confirms the role of the molecular private branch as a structure-preserving path rather than a redundant auxiliary head.

4.4. *Molecule-Image Retrieval*

We further evaluate cross-modal alignment through bidirectional retrieval. Image-to-molecule retrieval asks whether a cellular phenotype can recover its inducing molecule, whereas molecule-to-image retrieval asks whether a molecule can recover its paired cellular response. This task directly tests whether the learned representation places paired molecules and phenotypes close to one another, with results summarized in Table 2.

To complement the aggregate retrieval metrics, we visualize a representative scaffold-split retrieval case in Fig.4. The example compares the Top-1 retrieval result of each method and reports where the ground-truth pair appears in each method’s ranked list.

Under the random split, PhenMol achieves the best performance in both retrieval directions. For image-to-molecule retrieval, it obtains 11.74% Hit@1, 33.08% Hit@5, 45.38% Hit@10, and 0.2227 MRR, consistently outperforming InfoNCE, InfoLOOB, MIGA, and MINER. For molecule-to-image retrieval, PhenMol also achieves the highest Hit@1, Hit@5, Hit@10, and MRR scores. The comparison with InfoNCE and

Table 2: Molecule-image bidirectional retrieval results under random and scaffold splits. Hit@K is reported in percentage. MRR denotes mean reciprocal rank. Retrieval is evaluated over the complete test candidate library rather than mini-batches. The best results are highlighted in bold, and the second-best results are underlined.

Split	Direction	Method	Hit@1	Hit@5	Hit@10	MRR
Random	Image-to-molecule	Random	0.03	0.16	0.33	0.0000
	Image-to-molecule	InfoNCE	10.03	29.10	38.84	0.1967
	Image-to-molecule	InfoLOOB	<u>10.70</u>	<u>30.94</u>	<u>41.85</u>	<u>0.2054</u>
	Image-to-molecule	MIGA	8.85	25.72	36.80	0.1756
	Image-to-molecule	MINER	8.11	24.05	35.13	0.1447
	Image-to-molecule	PhenMol	11.74	33.08	45.38	0.2227
	Molecule-to-image	Random	0.03	0.16	0.33	0.0000
	Molecule-to-image	InfoNCE	10.75	29.43	38.01	0.2011
	Molecule-to-image	InfoLOOB	<u>11.06</u>	<u>30.22</u>	<u>42.76</u>	<u>0.2059</u>
	Molecule-to-image	MIGA	8.68	26.37	36.67	0.1755
	Molecule-to-image	MINER	8.21	24.59	35.36	0.1645
	Molecule-to-image	PhenMol	12.56	33.74	44.33	0.2259
Scaffold	Image-to-molecule	Random	0.04	0.18	0.37	0.0000
	Image-to-molecule	InfoNCE	0.95	4.26	<u>9.59</u>	0.0400
	Image-to-molecule	InfoLOOB	0.77	4.78	9.41	0.0406
	Image-to-molecule	MIGA	0.97	4.68	9.11	0.0426
	Image-to-molecule	MINER	<u>1.06</u>	<u>4.98</u>	9.51	<u>0.0446</u>
	Image-to-molecule	PhenMol	1.14	6.10	11.40	0.0482
	Molecule-to-image	Random	0.03	0.16	0.33	0.0000
	Molecule-to-image	InfoNCE	0.94	4.04	8.33	0.0426
	Molecule-to-image	InfoLOOB	1.10	4.74	9.04	0.0424
	Molecule-to-image	MIGA	0.99	4.50	8.70	0.0421
	Molecule-to-image	MINER	<u>1.18</u>	<u>5.04</u>	<u>9.41</u>	<u>0.0458</u>
	Molecule-to-image	PhenMol	1.36	5.33	9.82	0.0481

InfoLOOB is informative: these conventional contrastive objectives are competitive under the random split, indicating that global alignment can work when test molecules are close to training molecules. However, they only optimize vector-level similarity and do not explicitly separate phenotype-relevant information from chemical identity or nuisance variation.

Under the scaffold split, retrieval becomes substantially more challenging for all methods, reflecting the difficulty of cross-modal matching on unseen chemical scaffolds. The absolute Hit@1 values drop sharply, showing that cellular phenotypes

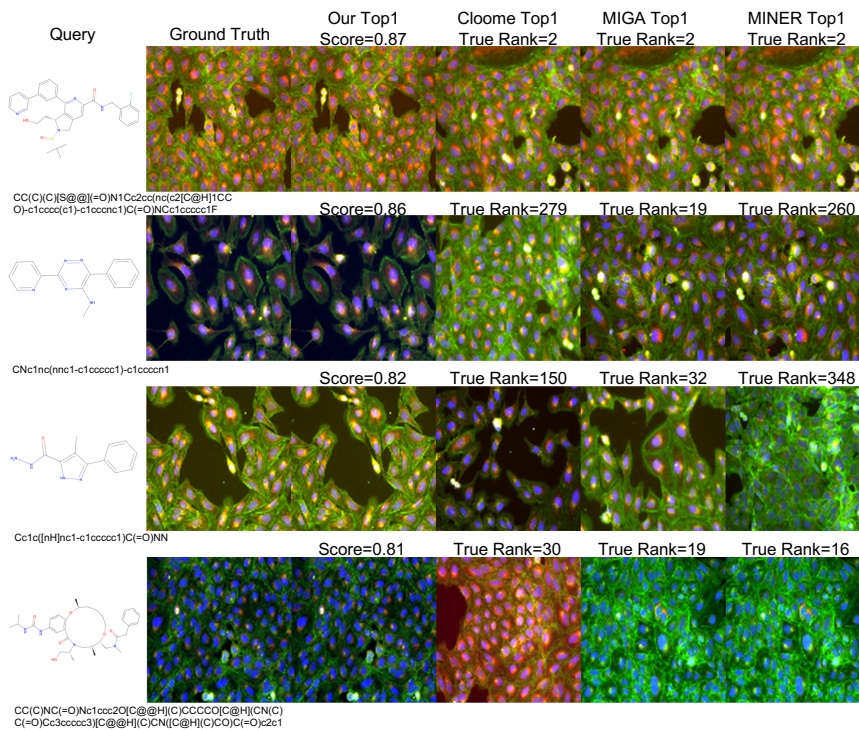


Figure 4: Qualitative scaffold-split example of molecule–image retrieval under competing multimodal methods. Given a query sample and its ground-truth match, the figure shows the Top-1 retrieval result produced by PhenMol , CLOOME, MIGA, and MINER. The retrieval score for PhenMol and the ground-truth rank assigned by each baseline are reported to illustrate whether each method retrieves the correct paired sample near the top of the full test-set ranking.

alone do not make scaffold-level retrieval easy. This is the setting where naive multimodal alignment is most vulnerable: images may capture similar morphology for chemically different perturbations, while scaffold changes can move the true molecular match far from the training distribution. PhenMol improves this setting by coupling global contrastive alignment with fine-grained latent pair matching and structure-preserving decoupling. The contrastive objective organizes the coarse image–molecule space, the matching branch checks pair compatibility after latent interaction, and the private molecular branch prevents the shared alignment objective from erasing scaffold-level information. This experiment emphasizes the importance of combining multimodal alignment with chemical-structure preservation rather than treating re-

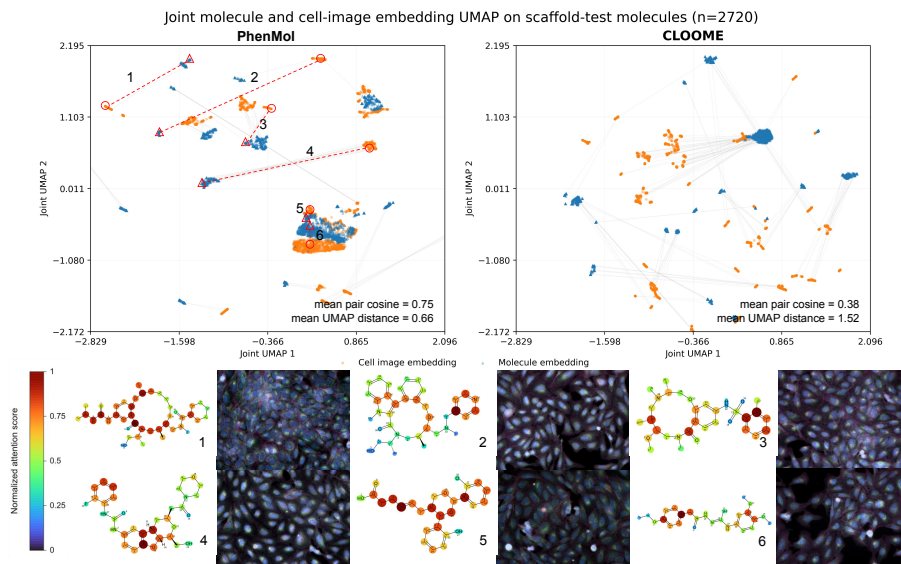


Figure 5: Visualization of image–molecule alignment and molecular attention on scaffold-test retrieval examples. Orange points denote cellular phenotype embeddings, blue triangles denote molecular embeddings, and gray lines connect paired observations. The lower panels show selected molecule–image pairs, with atom color indicating normalized attention score.

trieval as a pure contrastive matching problem.

We further inspect the paired molecule–cell embedding space on the scaffold-test molecules used for retrieval evaluation in Fig.5. PhenMol places paired cell-image and molecule embeddings closer than CLOOME in the displayed joint projection, with a mean pair cosine similarity of 0.75 and a mean UMAP distance of 0.66, compared with 0.38 and 1.52 for CLOOME. These values are descriptive summaries of the visualization rather than separate evaluation metrics. The lower panels show selected retrieved molecule–image pairs with normalized atom-level attention scores, which should be interpreted as model-derived saliency over the molecular graph rather than direct evidence of a biochemical mechanism.

4.5. Clinical Trial Outcome Prediction

We evaluate whether the learned molecular representations transfer beyond activity assays to clinical trial outcome prediction. This task tests whether phenotype-

Table 3: Performance comparison on clinical trial outcome prediction. Results are reported as mean $\pm$ standard deviation across stratified five-fold cross-validation within each phase. The best results are highlighted in bold, and the second-best results are underlined.

Method	Phase	AUC	PR-AUC
LR	Phase I	0.487 ± 0.006	0.634 ± 0.007
	Phase II	0.534 ± 0.017	0.509 ± 0.014
	Phase III	0.528 ± 0.003	0.675 ± 0.010
XGBoost	Phase I	0.508 ± 0.006	0.646 ± 0.003
	Phase II	0.516 ± 0.007	0.481 ± 0.004
	Phase III	0.597 ± 0.015	0.712 ± 0.009
HINT	Phase I	0.516 ± 0.005	<u>0.673 ± 0.015</u>
	Phase II	0.584 ± 0.003	0.537 ± 0.004
	Phase III	0.621 ± 0.006	0.689 ± 0.003
CLOOME	Phase I	0.514 ± 0.024	0.605 ± 0.005
	Phase II	0.563 ± 0.014	0.491 ± 0.021
	Phase III	0.571 ± 0.033	0.724 ± 0.036
MIGA	Phase I	<u>0.569 ± 0.050</u>	0.659 ± 0.040
	Phase II	0.593 ± 0.012	0.563 ± 0.017
	Phase III	0.603 ± 0.039	<u>0.757 ± 0.024</u>
MINER	Phase I	0.564 ± 0.033	0.633 ± 0.037
	Phase II	0.581 ± 0.012	0.515 ± 0.016
	Phase III	<u>0.631 ± 0.043</u>	0.751 ± 0.038
PhenMol	Phase I	0.616 ± 0.010	0.675 ± 0.040
	Phase II	0.593 ± 0.020	<u>0.521 ± 0.026</u>
	Phase III	0.669 ± 0.037	0.778 ± 0.029

aware molecular representations contain information relevant to real-world drug development outcomes, with results summarized in Table 3.

PhenMol achieves the best AUC and PR-AUC in Phase I, with an AUC of 0.616 and a PR-AUC of 0.675. It also achieves the best Phase III performance, with an AUC of 0.669 and a PR-AUC of 0.778. These two phases represent different clinical settings: early-stage trials emphasize safety and tolerability signals, whereas Phase III trials are closer to efficacy confirmation in larger populations. The improvement across both phases indicates that the learned molecular representation contains transferable information beyond the original cell-imaging pre-training task.

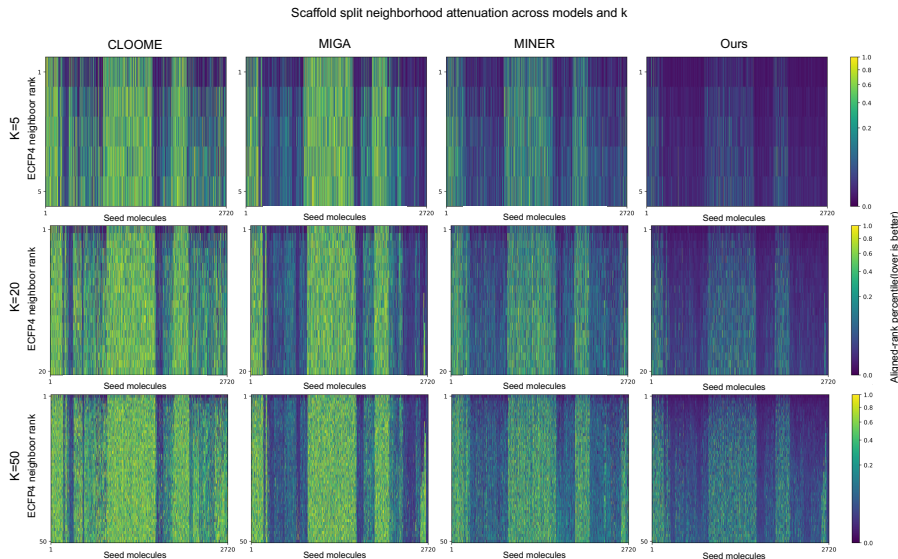


Figure 6: Visualization of molecular structural neighborhood preservation in different multimodal embedding spaces. Compared with global alignment methods such as CLOOME and robustness-enhanced methods such as MIGA and MINER, PhenMol better maintains ECFP4-defined structural neighbors after multimodal learning.

In Phase II, PhenMol matches the best AUC but has lower PR-AUC than MIGA. This mixed Phase II result provides an important boundary. Mid-stage clinical outcomes are strongly influenced by indication, patient selection, dose, endpoint design, and protocol context, which are not fully captured by molecule–phenotype representation learning. Thus, PhenMol improves the molecular component of clinical transfer, but the result also shows that molecule and cellular phenotype information alone cannot replace richer trial-level modeling.

4.6. Structural Preservation Analysis

Finally, we examine whether PhenMol preserves molecular structural neighborhoods after multimodal learning. This analysis directly tests the central concern that cross-modal alignment can distort chemically meaningful molecular organization. Fig.6 provides a qualitative comparison of learned embedding spaces, and Table 4 quantifies whether ECFP4-defined neighbors remain close after multimodal training.

Table 4: Structural preservation analysis based on ECFP4-defined molecular neighborhoods.

K	Model	R-Prec.↑	Rank Shift↓	Jaccard↑
5	MINER	0.1766	298.2	0.1137
	MIGA	0.1435	562.4	0.0958
	CLOOME	0.1427	708.9	0.0950
	PhenMol	0.4480	61.8	0.3313
20	MINER	0.1914	390.5	0.1166
	MIGA	0.1442	606.9	0.0886
	CLOOME	0.1150	853.2	0.0669
	PhenMol	0.4552	131.3	0.3230
50	MINER	0.2205	459.2	0.1432
	MIGA	0.2211	649.5	0.1331
	CLOOME	0.1307	917.8	0.0758
	PhenMol	0.4563	209.8	0.3195

PhenMol consistently achieves the best R-Precision and top- K Jaccard, together with the lowest mean absolute rank shift across all neighborhood sizes. At $K = 5$, PhenMol improves R-Precision from 0.1766 to 0.4480 compared with the strongest baseline and reduces mean rank shift from 298.2 to 61.8. This means that close ECFP4 neighbors remain much closer in the learned representation, instead of being displaced by phenotype-driven alignment. The advantage remains at larger neighborhood sizes, with R-Precision of 0.4552 at $K = 20$ and 0.4563 at $K = 50$. This scale consistency is important because local neighborhood preservation and broader scaffold-level organization are different requirements: a method can preserve a few nearest analogs while still distorting the larger chemical neighborhood. PhenMol improves both, indicating that the private scaffold-preserving branch constrains molecular geometry beyond the immediate nearest neighbors. Together with the property prediction and retrieval results, this analysis supports the central mechanism of the paper: phenotype-aware learning is most effective when cross-modal alignment is coupled with explicit molecular structure preservation.

The UMAP visualization in Fig.7 provides a complementary view of the scaffold-test molecular embedding space. Gray points denote the full molecule set, and col-

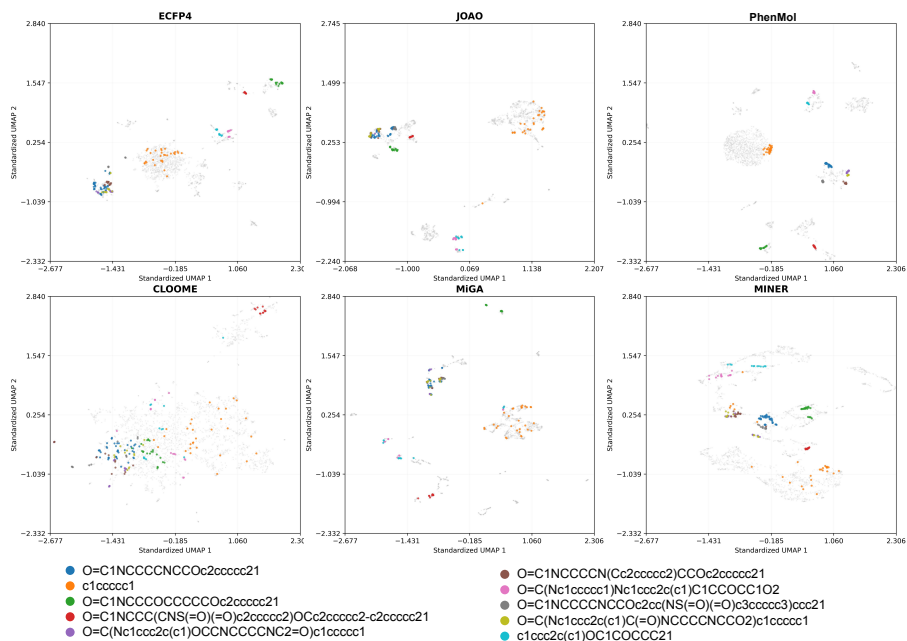


Figure 7: UMAP visualization of molecular embedding spaces on the scaffold-test molecules. Gray points denote the full molecule set, and colored points highlight representative molecules from related chemical neighborhoods.

ored points highlight representative molecules from related chemical neighborhoods. Compared with multimodal alignment baselines whose local organization becomes more diffuse, PhenMol maintains compact highlighted neighborhoods while still being trained with paired cellular phenotype information, indicating that structure preservation is visible in the learned geometry rather than only in the rank-based metrics.

5. Conclusion

We introduced PhenMol, a novel approach designed to address a key failure mode in phenotype-aware molecular representation learning: cellular image supervision can improve biological alignment but may also distort chemically meaningful molecular neighborhoods. PhenMol incorporates structure-preserving multimodal learning that constrains cross-modal alignment by chemical structure rather than optimizing solely

for global image–molecule similarity. Extensive experiments show that PhenMol consistently outperforms state-of-the-art methods under both random and scaffold splits, achieving superior accuracy and robustness in molecular property prediction, bidirectional molecule–phenotype retrieval, and clinical trial outcome prediction. PhenMol also preserves ECFP4-defined structural neighborhoods more effectively than existing cross-modal methods and demonstrates strong Phase I and Phase III clinical trial results, offering new insights into the role of chemical structure preservation in phenotype-aware molecular learning and facilitating the discovery of novel therapeutic solutions.

PhenMol learns from molecular fingerprints and pre-extracted cellular phenotype features, but does not explicitly model targets, pathways, transcriptomic responses, disease context, or clinical trial variables. As a result, although it improves structure-preserving molecule–cell alignment, its predictions cannot fully explain the underlying biological mechanisms or capture all factors that determine clinical outcomes, especially in Phase II trials where indication, dosage, patient population, endpoints, and protocol design are often critical. Therefore, PhenMol should be used as a research-oriented representation-learning and candidate-prioritization tool rather than a standalone system for clinical decision-making. Its predictions should be treated as hypotheses requiring experimental validation and, where relevant, independent clinical assessment. Future work should incorporate richer biomedical and clinical modalities and improve interpretability by linking molecular substructures to cellular phenotypic patterns and downstream outcomes.

CRedit Authorship Contribution Statement

Xuan Lin: Resources, Methodology, Investigation, Formal analysis, Funding acquisition, Conceptualization. Jingyu Sheng: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. Tengfei

Ma: Visualization, Supervision, Project administration, Investigation, Conceptualization, Funding acquisition. Li Sun: Writing – review & editing, Resources, Methodology, Investigation, Formal analysis. Dapeng Xiong: Project administration, Investigation, Data curation, Funding acquisition.

Acknowledgements

We thank our colleagues for their helpful discussions and constructive suggestions. This research was supported by the National Natural Science Foundation of China (No. 62573372 to Xuan Lin and No. 625B2067 to Tengfei Ma), and by the Southeast University Startup Grant (No. 4007032510 to D.X.).

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

Generative AI tools were used to assist with language editing, manuscript organization, and clarity checks during the preparation of this manuscript. They were not used to generate experimental data, run analyses, create evaluation results, or make autonomous scientific conclusions.

Data Availability

The code and data used in this study are available at <https://github.com/JacklinGroup/Phenmol>.

References

- [1] A. V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, *Nature* 616 (7958) (2023) 673–685.

- [2] E. H. B. Maia, L. C. Assis, T. A. De Oliveira, A. M. Da Silva, A. G. Taranto, Structure-based virtual screening: from classical to artificial intelligence, *Front. Chem.* 8 (2020).
- [3] X. Chen, S. Li, R. Liu, B. Shi, J. Liu, J. Wu, K. Xu, Molecular graph contrastive learning with line graph, *Pattern Recognit.* 162 (2025) 111380.
- [4] Y. Tabana, D. Babu, R. Fahlman, A. G. Siraki, K. Barakat, Target identification of small molecules: an overview of the current applications in drug discovery, *BMC Biotechnol.* 23 (1) (2023).
- [5] A. D. Schachter, M. F. Ramoni, Clinical forecasting in drug development, *Nat. Rev. Drug Discov.* 6 (2) (2007) 107–108.
- [6] D. Rogers, M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.* 50 (5) (2010) 742–754.
- [7] E. Hwang, V. Thost, S. S. Dasgupta, T. Ma, An analysis of virtual nodes in graph neural networks for link prediction, in: *Proceedings of the First Learning on Graphs Conference, LoG '22, PMLR, Virtual Event, 2022*, p. 98.
- [8] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 28 (1) (1988) 31–36.
- [9] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, J. Huang, Self-supervised graph transformer on large-scale molecular data, in: *Advances in Neural Information Processing Systems, Vol. 33 of NeurIPS '20, Curran Associates, Inc., Virtual Event, 2020*, pp. 12559–12571.
- [10] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, in: *Advances in Neural Information Processing Systems*,

Vol. 33 of NeurIPS '20, Curran Associates, Inc., Virtual Event, 2020, pp. 5812–5823.

- [11] Y. You, T. Chen, Y. Shen, Z. Wang, Graph contrastive learning automated, in: Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of ICML '21, PMLR, Virtual Event, 2021, pp. 12121–12132.
- [12] Y. Wang, J. Wang, Z. Cao, A. Barati Farimani, Molecular contrastive learning of representations via graph neural networks, *Nat. Mach. Intell.* 4 (3) (2022) 279–287.
- [13] H. Li, D. Zhao, J. Zeng, KPGT: Knowledge-guided pre-training of graph transformer for molecular property prediction, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 857–867.
- [14] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto, M. Mercola, Phenotypic drug discovery: recent successes, lessons learned and new directions, *Nat. Rev. Drug Discov.* 21 (12) (2022) 899–914.
- [15] S. Seal, M.-A. Trapotsi, O. Spjuth, S. Singh, J. Carreras-Puigvert, N. Greene, A. Bender, A. E. Carpenter, Cell painting: a decade of discovery and innovation in cellular imaging, *Nat. Methods* 22 (2) (2025) 254–268.
- [16] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, A. E. Carpenter, Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes, *Nat. Protoc.* 11 (9) (2016) 1757–1774.
- [17] M. Akbarzadeh, I. Deipenwisch, B. Schoelermann, A. Pahl, S. Sievers, S. Ziegler, H. Waldmann, Morphological profiling by means of the cell painting

assay enables identification of tubulin-targeting compounds, *Cell Chem. Biol.* 29 (6) (2022) 1053–1064.

- [18] A. Pahl, B. Schölermann, P. Lampe, M. Rusch, M. Dow, C. Hedberg, A. Nelson, S. Sievers, H. Waldmann, S. Ziegler, Morphological subprofile analysis for bioactivity annotation of small molecules, *Cell Chem. Biol.* 30 (7) (2023) 839–853.
- [19] N. Moshkov, T. Becker, K. Yang, P. Horvath, V. Dancik, B. K. Wagner, P. A. Clemons, S. Singh, A. E. Carpenter, J. C. Caicedo, Predicting compound activity from phenotypic profiles and chemical structures, *Nat. Commun.* 14 (1) (2023).
- [20] Y. Gui, C. Ma, Z. Ma, Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables, in: *Advances in Neural Information Processing Systems*, Vol. 38 of *NeurIPS '25*, Curran Associates, Inc., San Diego, CA, USA, 2025, pp. 122364–122409.
- [21] 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: *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *ICML '21*, PMLR, Virtual Event, 2021, pp. 8748–8763.
- [22] J. Zhang, K. He, Z. Feng, S. Sun, X. Sun, Z. Yuan, J. Yu, Self-adaptive image-text fusion for medical image classification, *Pattern Recognit.* 167 (2025) 111715.
- [23] J. Arevalo, E. Su, J. D. Ewald, R. Van Dijk, A. E. Carpenter, S. Singh, Evaluating batch correction methods for image-based cell profiling, *Nat. Commun.* 15 (1) (2024).
- [24] M. Syetkowski, M. Rezanejad, S. Saberian, O. Kraus, J. Urbanik, J. Taylor,

- B. Mabey, M. Vectors, 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, CVPR '23, IEEE/CVF, Vancouver, BC, Canada, 2023, pp. 4285–4294.
- [25] R. Abe, T. Miyao, J. Bajorath, Accurate prediction of activity cliff compounds based on bioactivity profiles depends on assay nearest neighbor relationships, *J. Cheminform.* 18 (1) (2026).
- [26] A. Sanchez-Fernandez, E. Rumetshofer, S. Hochreiter, G. Klambauer, Cloome: Contrastive learning unlocks bioimaging databases for queries with chemical structures, *Nat. Commun.* 14 (1) (2023).
- [27] S. Zheng, J. Rao, J. Zhang, L. Zhou, J. Xie, E. Cohen, W. Lu, C. Li, Y. Yang, Cross-modal graph contrastive learning with cellular images, *Adv. Sci.* 11 (32) (2024).
- [28] J. Rao, H. Lin, L. Chen, J. Xie, S. Zheng, Y. Yang, Multi-modal contrastive learning with negative sampling calibration for phenotypic drug discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '25, Computer Vision Foundation / IEEE, Nashville, TN, USA, 2025, pp. 30752–30762.
- [29] S. Wang, Z. Zhao, Z. Zhuang, X. Ouyang, L. Zhang, Z. Li, C. Ma, T. Liu, D. Shen, Q. Wang, Learning better contrastive view from radiologist’s gaze, *Pattern Recognit.* 162 (2025) 111350.
- [30] M. Li, Y. Cao, X. Liu, H. Ji, Knowledge-enhanced and structure-enhanced representation learning for protein–ligand binding affinity prediction, *Pattern Recognit.* 166 (2025) 111701.

- [31] Y. Luo, K. Yang, M. Hong, X. Y. Liu, Z. Nie, H. Zhou, Z. Nie, Learning multi-view molecular representations with structured and unstructured knowledge, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 2082–2093.
- [32] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, *J. Mach. Learn. Res.* 17 (59) (2016) 1–35.
- [33] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of ICML '20, PMLR, Virtual Event, 2020, pp. 1597–1607.
- [34] S. N. Chandrasekaran, B. A. Cimini, A. Goodale, L. Miller, M. Kost-Alimova, N. Jamali, J. G. Doench, B. Fritchman, A. Skepner, M. Melanson, et al., Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations, *Nat. Methods* 21 (6) (2024) 1114–1121.
- [35] S. M. Corsello, J. A. Bittker, Z. Liu, J. Gould, P. McCarren, J. E. Hirschman, S. E. Johnston, A. Vrcic, B. Wong, M. Khan, et al., The drug repurposing hub: a next-generation drug library and information resource, *Nat. Med.* 23 (4) (2017) 405–408.
- [36] T. Fu, K. Huang, C. Xiao, L. M. Glass, J. Sun, Hint: Hierarchical interaction network for clinical-trial-outcome predictions, *Patterns* 3 (4) (2022).
- [37] R. contributors, Rdkit: Open-source cheminformatics (2024).
- [38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition,

in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16, IEEE, 2016, pp. 770–778.

- [39] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, Reoptimization of MDL keys for use in drug discovery, *J. Chem. Inf. Comput. Sci.* 42 (6) (2002) 1273–1280.