Learning Cell-Aware Hierarchical Multi-Modal Representations for Robust Molecular Modeling

Mengran Li[1,2,3*], **Zelin Zang**[2,3,4*], **Wenbin Xing**[1], **Junzhou Chen**[1†], **Ronghui Zhang**[1], **Jiebo Luo**[3], **Stan Z. Li**[2],
[1]School of Intelligent Systems Engineering, Sun Yat-sen University
[2]Westlake University
[3]Center for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science and Innovation (HKISI)
[4]Center for Integrated Circuits and Artificial Intelligence, Tsientang Institute for Advanced Study
limr39@mail2.sysu.edu.cn, zangzelin@westlake.edu.cn, chenjunzhou@mail.sysu.edu.cn

Abstract

Understanding how chemical perturbations propagate through biological systems is essential for robust molecular property prediction. While most existing methods focus on chemical structures alone, recent advances highlight the crucial role of cellular responses such as morphology and gene expression in shaping drug effects. However, current cell-aware approaches face two key limitations: (1) modality incompleteness in external biological data, and (2) insufficient modeling of hierarchical dependencies across molecular, cellular, and genomic levels. We propose **CHMR** (Cell-aware Hierarchical Multi-Modal Representations), a robust framework that jointly models local-global dependencies between molecules and cellular responses and captures latent biological hierarchies via a novel tree-structured vector quantization module. Evaluated on public benchmarks spanning 696 tasks, CHMR outperforms state-of-the-art baselines, yielding average improvements of **3.6%** on classification and **17.2%** on regression tasks. These results demonstrate the advantage of hierarchy-aware, multi-modal learning for reliable and biologically grounded molecular representations, offering a generalizable framework for integrative biomedical modeling.

Code — https://github.com/limengran98/CHMR

1 Introduction

Predicting molecular properties, such as activity (Liu et al. 2023a), toxicity (Deng et al. 2023), and side effects (Zhang et al. 2025a), plays a crucial role in accelerating drug development and ensuring the safety of drug candidates. With the rapid development of deep learning technologies, the prediction of drug molecular properties has increasingly become automated (Liu et al. 2022; Zhou et al. 2023), significantly reducing both development time and resource consumption (Chen et al. 2025). Considering the inherent graph structure of drug molecules, predicting using internal molecular features (such as atomic properties, interatomic bonds, and three-dimensional structures) has become feasible. For example, some methods pre-train on large, unlabeled molecular datasets using perturbation or augmentation strategies (Jiang et al. 2025; Hu et al. 2025; Jia et al. 2025), enabling the transfer of learned knowledge to various downstream tasks.

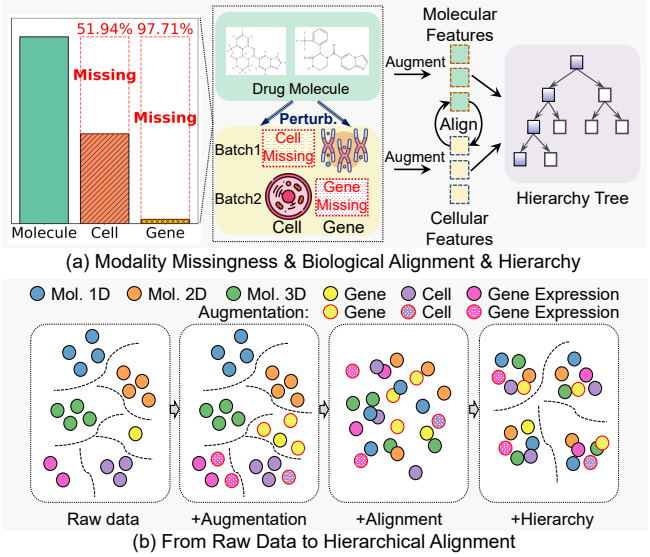


Figure 1: The motivation of this paper. Molecular perturbations trigger cellular or genetic changes, but modality incompleteness is common. Through augmentation, alignment, and hierarchical modeling, multi-modal representations are progressively organized and structured.

While structure-based models have proven effective, they often fail to capture biological responses, including gene expression and cell morphology, that are induced by molecular interactions. To complement this, researchers have started incorporating biological context into molecular representation learning (Wang et al. 2024; Liu et al. 2025). When molecules bind to cellular targets, they trigger signaling cascades that can alter gene expression (Himmelstein et al. 2017; Chandrasekaran et al. 2023) and affect cell morphology (Bray et al. 2016, 2017). By learning cross-modal alignment and building a shared representation space, the differences between biological modalities are bridged, thereby improving the semantic understanding of molecules. Consequently, incorporating

*These authors contributed equally.
†Corresponding author.

cell-aware molecular representation learning is essential for understanding the mechanisms by which molecules exert their effects and for designing safer and more effective compounds (Moshkov et al. 2023).

Despite recent advances, two critical challenges remain (Figure 1(a)). **First, biological modality missingness is pervasive and asymmetric.** While molecular structure data are typically complete, associated cellular phenotypes or gene expression profiles are frequently unavailable due to experimental limitations or cost constraints (Gatto et al. 2023; Zhang et al. 2023). For instance, a compound may have morphological cell readouts but lack transcriptomic data, or vice versa. This leads to distributional shifts and modality imbalance, which undermine model robustness and limit generalization in practical scenarios.

Second, biological modalities exhibit hierarchical dependencies that are difficult to capture. Molecular perturbations initiate cascades of interactions across biological layers—from chemical structures to cellular processes to gene expression programs. Yet many existing models treat these modalities in a flattened latent space and focus on instance-level alignment (Sanchez-Fernandez et al. 2023; Gulati et al. 2025), failing to capture multi-hop semantic relationships and cross-layer dependencies essential for modeling cross-scale biological mechanisms.

To address these challenges, we propose **CHMR** (Cell-aware Hierarchical Multi-Modal Representations), a unified framework for robust and interpretable molecular representation learning across multiple biological modalities. CHMR is designed to address two key limitations in existing cell-aware methods. First, to handle pervasive and asymmetric missingness in biological modalities, CHMR captures both local and global dependencies between molecules and their cellular responses, and enforces semantic consistency between molecular features and available cell- and gene-level information. Second, to model cross-scale biological mechanisms, CHMR incorporates a tree-structured vector quantization module that encodes latent hierarchies across molecules, cells, and genes. This design preserves biologically meaningful structures, mitigates information flattening, and enhances interpretability and generalization. Figure 1(b) illustrates how hierarchical modeling helps align and organize multi-scale biological signals across modalities. Our key contributions are as follows:

- We propose a unified framework that jointly models molecular structures, cellular phenotypes, and gene expression profiles, enabling robust and generalizable representation learning under missing biological modalities.
- We introduce a tree-structured vector quantization module to capture hierarchical dependencies among molecules, cells, and genes, facilitating biologically grounded fusion and improved cross-modal interpretability.
- We validate CHMR on 696 molecular property prediction tasks across benchmark datasets. CHMR consistently outperforms state-of-the-art baselines, achieving average improvements of **3.6%** on classification and **17.2%** on regression tasks.

2 Related Work

2.1 Multi-Modal Molecular Representations

In recent years, molecular representation learning has achieved significant progress in drug discovery and chemical property modeling. To better capture molecular properties, recent studies have explored integrating diverse modalities such as 3D conformations, and molecular language into unified frameworks. In 3D geometry, GraphMVP (Liu et al. 2022) enabled 2D-to-3D semantic transfer via contrastive learning and generative reconstruction, without requiring 3D inputs during inference. GEM (Fang et al. 2022) and Uni-Mol (Zhou et al. 2023) introduced geometry-level networks and SE(3)-equivariant Transformers to improve the modeling of 3D conformations. MOLEBLEND (Yu et al. 2024) introduced an atom-relation level fusion mechanism for fine-grained structural alignment. MOL-Mamba (Hu et al. 2025) incorporated electronic semantics at the atomic level, jointly learning with structural information in a graph state space model.

For molecular language models, MolT5 (Edwards et al. 2022) and ChemGPT (Frey et al. 2023) established general-purpose chemical language models, enabling cross-modal translation between molecules and text. MolCA (Liu et al. 2023b) projected structural graphs into large language model (LLM) spaces with a Q-Former, supporting cross-modal generation and retrieval. Atomas (Zhang et al. 2025b) introduced hierarchical alignment from atoms, fragments, and molecules to text for fine-grained structure–text mapping.

While these methods improve multi-modal fusion, they do not explicitly model the hierarchical dependencies between molecules, cells, and genes, making it hard to capture cross-scale biological mechanisms.

2.2 Cell-Responsive Multi-Modal Molecular Representations

With the rise of large-scale cell imaging and multi-omics data, recent studies have explored incorporating cellular-level phenotypic responses into molecular modeling to address the semantic gap in functional biology. MoCoP (Nguyen, Pertusi, and Branson 2023) first used cell phenotype images as functional supervision in molecular pretraining, aligning structure and phenotype through contrastive learning and significantly enhancing biological semantic awareness. CLOOME (Sanchez-Fernandez et al. 2023) extended this to bidirectional retrieval tasks with structure–phenotype paired contrastive loss and a shared embedding space.

Further expanding modality dimensions, InfoCORE (Wang et al. 2024) focused on removing non-biological batch effects by maximizing conditional mutual information for bias-robust structure–phenotype modeling. InfoAlign (Liu et al. 2025) integrated molecular structures, cellular morphology, and transcriptomics into a shared representation space, aligning molecules with their biological responses to enrich both chemical and functional semantics and improve generalization on pharmacological and toxicity prediction tasks.

However, existing approaches still face key limitations. Most methods assume complete multi-modal input and per-

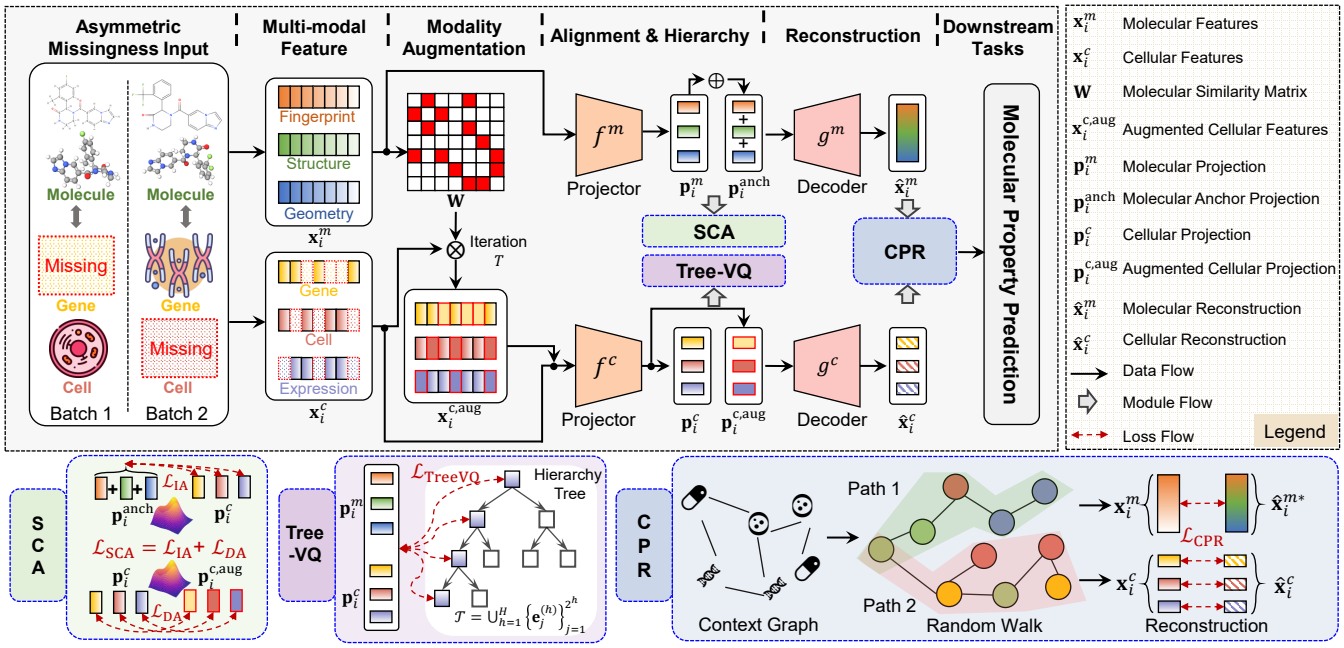


Figure 2: Overview of the CHMR framework for robust molecular property prediction under missing biological modalities. CHMR performs modality augmentation via structure-aware propagation, followed by (1) semantic consistency alignment (SCA) to align molecular and cellular modalities, (2) tree-structured vector quantization (Tree-VQ) to capture hierarchical biological semantics, and (3) context propagation reconstruction (CPR) to enhance generalization through cross-modal context.

form static fusion or flattened alignment, making them vulnerable to modality missingness and distribution shifts.

3 Preliminaries and Notation

Definition 1: Cell-aware molecular representation learning Let $\mathcal{V} = \{v_i\}_{i=1}^{N}$ denote the set of N molecules. For each molecule $v_i \in \mathcal{V}$, we define its multi-modal feature set as $\mathbf{X}_i = \{\mathbf{x}_i^{\xi} \mid \xi \in \mathcal{M} \cup \mathcal{C}\}$, where ξ indexes a generalized modality. The set $\mathcal{M}$ denotes molecular modalities, each $m \in \mathcal{M}$ corresponding to structural molecular features $\mathbf{x}_i^m$ such as molecular fingerprints encoding substructure information (Rogers and Hahn 2010), graph-based representations derived from molecular topology via graph neural networks (Wang et al. 2022), or geometric features obtained from molecular conformations (Zhou et al. 2023). The set $\mathcal{C}$ denotes external cellular modalities, each $c \in \mathcal{C}$ representing a molecular-induced biological response $\mathbf{x}_i^c$, such as cell morphology profiles from imaging (Bray et al. 2016), gene perturbation features (Chandrasekaran et al. 2023), or gene expression data (Himmelstein et al. 2017).

Definition 2: Missing external cellular modality The issue of missing cellular modalities arises due to experimental constraints or cost limitations. Formally, let $\mathcal{C}^{\text{obs}} \subseteq \mathcal{C}$ represent the set of observable external cellular modalities, and let $\mathcal{C}^{\text{miss}} = \mathcal{C} \setminus \mathcal{C}^{\text{obs}}$ denote the missing modalities. In this study, we assume that all molecular modalities $\{\mathbf{x}_i^m \mid m \in \mathcal{M}\}$ are fully available, with the focus being on the scenario where external cellular modality features $\{\mathbf{x}_i^c \mid c \in \mathcal{C}^{\text{miss}}\}$ are missing. Different molecules may exhibit different patterns of missing cellular modalities. For example, some molecules may lack cellular phenotype features, while others may lack gene expression data (Figure 1). Missing cellular modality features are replaced with placeholders, like zero vectors or global averages, to ensure consistent feature matrix dimensions.

Research Objectives Our objective is to learn robust molecular representations that support semantic completion, preserve molecule–cell–gene hierarchical consistency under missing modalities, and generalize to downstream property prediction tasks such as solubility, activity, and toxicity.

4 Methodology

We propose the CHMR framework to handle multi-modal missing data and capture hierarchical dependencies between molecules and cells, ensuring accurate prediction of molecular properties. The architecture is shown in Figure 2.

4.1 Modality Augmentation

To address incomplete and asymmetric biological modalities (Figure 1(a)), we propose a structure-aware propagation strategy for modality augmentation that captures both local and global molecular context, mitigating the limitations of mean or K-NN imputation.

Given a molecule set $\mathcal{V} = \{v_i\}_{i=1}^{N}$, we construct a pairwise similarity matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $\mathbf{W}_{ij}$ denotes the structural similarity between molecules v_i and v_j. For each molecule v_i, only its top-K nearest neighbors $\mathcal{N}_K(v_i)$ are retained to enhance locality.

For each external modality $c \in \mathcal{C}$ of molecule v_i, we perform iterative propagation inspired by Dirichlet energy minimization (Rossi et al. 2022):

$$\mathbf{x}_i^{c,(T)} = \begin{cases} \mathbf{x}_i^c, & c \in \mathcal{C}_i^{\text{obs}}, \\ \sum\limits_{j \in \mathcal{N}_K(v_i)} \mathbf{W}_{ij}\, \mathbf{x}_j^{c,(T-1)}, & c \in \mathcal{C}_i^{\text{miss}}. \end{cases} \quad (1)$$

Observed modalities retain their original features, while missing ones are iteratively estimated via neighbor aggregation. The final augmented feature $\mathbf{x}_i^{c,\text{aug}}$ incorporates local structural context.

4.2 Semantic Consistency Alignment

To mitigate semantic discrepancies between molecular and external cellular modalities, especially those amplified by modality augmentation, we introduce the SCA module.

We begin by projecting each modality into a shared latent space:

$$\mathbf{p}_i^m = f^m(\mathbf{x}_i^m), \quad \mathbf{p}_i^c = f^c(\mathbf{x}_i^c), \quad \mathbf{p}_i^{c,\text{aug}} = f^c(\mathbf{x}_i^{c,\text{aug}}), \quad (2)$$

where $f^m(\cdot)$ and $f^c(\cdot)$ denote the projectors for molecular and cellular modalities, respectively, mapping features into a common representation space.

Sample-level Alignment We enforce alignment between molecular and cellular representations using an InfoNCE-style contrastive loss. Molecular features are aggregated into an anchor vector $\mathbf{p}_i^{\text{anch}} = \sum_{m \in \mathcal{M}} \mathbf{p}_i^m$, and paired with $\mathbf{p}_i^c$ as positives:

$$\mathcal{L}_{\text{IA}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \log \frac{\exp\left(\cos(\mathbf{p}_i^{\text{anch}}, \mathbf{p}_i^c)/\tau\right)}{\sum_{j \in \mathcal{V}} \exp\left(\cos(\mathbf{p}_i^{\text{anch}}, \mathbf{p}_j^c)/\tau\right)}, \quad (3)$$

where $\tau > 0$ is the temperature parameter.

Distribution-level Alignment To counteract potential distributional shifts introduced by propagation-based augmentation, we adopt a VICReg-style loss to align augmented and original cellular features:

$$\mathcal{L}_{\text{DA}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \text{VICReg}\left(\mathbf{p}_i^c, \mathbf{p}_i^{c,\text{aug}}\right), \quad (4)$$

where $\text{VICReg}(\cdot)$ (Bardes, Ponce, and LeCun 2022) enforces semantic and statistical consistency via invariance, variance, and covariance regularization.

The overall SCA objective combines both levels of alignment: $\mathcal{L}_{\text{SCA}} = \mathcal{L}_{\text{IA}} + \mathcal{L}_{\text{DA}}$.

4.3 Tree-Structured Vector Quantization

Biological responses across molecular, cellular, and genomic levels are inherently hierarchical (Gulati et al. 2025). To capture cross-scale semantics between molecular structures and biological signals, we introduce the Tree-VQ module for hierarchical representation learning. Unlike flat alignment, Tree-VQ interprets the layers of the tree as representing biological levels ranging from molecular fingerprints (shallow) to cellular phenotypes and gene expression (deep), aligned with the hierarchical responses illustrated in Figure 1(b).

Multi-Modal Collaborative Tree Structure Let $\mathbf{p}^\xi \in \mathbb{R}^{d_\xi}$ denote the projected feature vector of modality ξ, where $\xi \in \mathcal{M} \cup \mathcal{C}$. Tree-VQ constructs a binary tree $\mathcal{T} = \bigcup_{h=1}^{H} \mathcal{E}^h$ of depth H, where $\mathcal{E}^h = \left\{\mathbf{e}_j^h\right\}_{j=1}^{2^h}$ represents the set of embeddings at level h, and $\mathbf{e}_j^h$ denotes the j-th node embedding at level h. Crucially, this tree is shared across all modalities, enabling heterogeneous features to be jointly embedded and routed through a unified semantic hierarchy.

At each level h, we compute the cosine distance between $\mathbf{p}^\xi$ and candidate tree nodes:

$$\delta_j^{(\xi,h)} = 1 - \cos\left(\mathbf{p}^\xi, \mathbf{e}_j^h\right). \quad (5)$$

To ensure that the quantization path follows the tree hierarchy, a routing mask is applied. If the parent node selected at level $h-1$ has index $j_{\text{par}}^{\xi,h-1}$, only its two child nodes $\{2j_{\text{par}}^{\xi,h-1},\ 2j_{\text{par}}^{\xi,h-1}+1\}$ are considered:

$$\tilde{\delta}_j^{\xi,h} = \begin{cases} \delta_j^{\xi,h}, & j \in \{2j_{\text{par}}^{\xi,h-1}, 2j_{\text{par}}^{\xi,h-1}+1\}, \\ +\infty, & \text{otherwise}. \end{cases} \quad (6)$$

Hierarchical Quantization and Symmetric VQ Loss Tree-VQ performs hierarchical vector quantization by assigning each projected feature vector to the closest tree node:

$$j^{*\xi,h} = \arg\min_j \tilde{\delta}_j^{\xi,h}, \quad \mathbf{q}^{\xi,h} = \mathbf{e}_{j^{*\xi,h}}^h. \quad (7)$$

To enforce bidirectional consistency between the encoder and tree nodes, we define a symmetric VQ loss:

$$\begin{aligned} \mathcal{A}\left(\mathbf{p}^\xi, \mathbf{q}^{\xi,h}\right) = 1 - \cos\left(\text{sg}\left[\mathbf{q}^{\xi,h}\right], \mathbf{p}^\xi\right) \\ + \eta\left(1 - \cos\left(\mathbf{q}^{\xi,h}, \text{sg}\left[\mathbf{p}^\xi\right]\right)\right), \end{aligned} \quad (8)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation to ensure differentiability, and η controls the reverse commitment weight.

Finally, the Tree-VQ loss is defined by averaging across modalities and tree levels:

$$\mathcal{L}_{\text{TreeVQ}} = \frac{1}{|\{\xi\}|} \sum_{\xi} \frac{1}{H} \sum_{h=1}^{H} \mathcal{A}\left(\mathbf{p}^\xi, \mathbf{q}^{\xi,h}\right). \quad (9)$$

4.4 Context-Propagation Reconstruction

To compensate for the lack of explicit supervision in modality augmentation, the CPR module introduces structure-aware reconstruction guided by cross-modal contextual signals. Following (Liu et al. 2025), we utilize a contextual graph $\mathcal{H} = (\mathcal{U}, \mathcal{R})$, where $\mathcal{U}$ includes molecules and external biological modalities, and each relation $(u_i, u_j) \in \mathcal{R}$ encodes a potential association. Importantly, these relations incorporate biological priors, such as known molecular perturbation–response pairs, functional associations, and shared regulatory pathways (Himmelstein et al. 2017; Chandrasekaran et al. 2023). Each relation is assigned a normalized weight $\beta \in [0, 1]$, indicating semantic strength derived from experimental, structural, or statistical sources.

On graph $\mathcal{H}$, we perform a random walk of length L starting from each node u_i, yielding a path $(u_i, u_{i_1}, ..., u_{i_L})$. The

	Dataset (Molecule / Task)		ChEMBL (2355/41)	ToxCast (8576/617)	Broad (6567/32)	Biogen (MAE $\times 100 \downarrow$) (3521 / 6)						
	Method	**Venue**	**AUC% ↑**	**AUC% ↑**	**AUC% ↑**	**Avg. ↓**	**HLM ↓**	**RLM ↓**	**ER ↓**	**Solubility ↓**	**hPPB ↓**	**rPPB ↓**
Single-Modal	MLP	JCIM'10	76.8±2.2	57.6±1.0	63.3±0.3	66.2±2.4	66.1±2.6	69.5±3.0	56.8±2.3	56.5±4.2	74.6±6.2	73.7±7.3
	RF	JCIM'10	54.7±0.7	52.3±0.1	55.5±0.1	52.8±0.2	44.2±0.1	51.6±0.1	44.2±0.1	42.0±0.2	67.7±0.7	66.9±0.9
	GP	JCIM'10	51.0±0.0	OOM	50.6±0.0	60.0±0.0	51.3±0.0	61.6±0.0	59.5±0.0	49.7±0.0	68.8±0.0	69.3±0.0
	AttrMask	ICLR'20	73.9±0.5	63.1±0.8	59.8±0.2	67.3±0.3	82.4±1.1	99.1±1.2	49.8±0.7	51.7±1.0	<u>57.9±0.6</u>	62.6±0.5
	ContextPred	ICLR'20	77.0±0.5	63.0±0.6	60.0±0.2	68.5±0.9	85.0±7.9	96.5±3.7	49.7±0.4	55.1±2.7	61.4±1.8	63.1±0.5
	EdgePred	ICLR'20	75.6±0.5	63.5±1.1	59.9±0.2	67.8±0.9	81.2±10.2	99.1±6.9	48.0±0.5	53.5±2.8	62.2±1.8	62.9±0.7
	GraphCL	NeurIPS'20	75.6±1.6	52.2±0.2	67.2±0.5	53.9±0.6	43.8±0.3	49.6±0.3	45.4±0.6	40.6±0.5	76.7±1.0	67.1±2.2
	GROVER	NeurIPS'20	73.3±1.4	53.1±0.4	66.2±0.1	54.9±1.6	44.5±0.4	52.6±0.3	46.5±0.7	41.7±0.6	73.2±5.7	71.0±4.3
	JOAO	ICML'21	75.1±1.0	52.3±0.2	67.3±0.4	55.0±0.8	44.5±0.5	51.4±0.6	47.6±0.5	40.6±0.2	74.3±2.8	71.5±2.6
	MGSSL	NeurIPS'21	75.1±1.1	64.2±0.2	66.9±0.5	53.2±0.3	44.8±0.6	41.5±0.2	65.6±1.8	64.6±0.5	52.7±0.5	49.7±0.3
	GraphLoG	ICML'21	73.5±0.7	58.6±0.4	62.9±0.4	56.9±0.4	49.3±0.3	58.8±0.5	54.8±0.5	42.6±0.3	66.8±1.7	69.0±1.3
	GraphMAE	KDD'22	74.7±0.1	53.3±0.1	66.8±0.3	52.8±0.8	43.3±0.9	50.9±1.4	51.2±0.8	40.9±0.3	64.4±2.7	65.9±3.8
	DSLA	NeurIPS'22	69.3±1.0	57.8±0.5	63.3±0.3	57.9±0.7	50.4±0.7	60.9±0.6	53.6±1.7	43.3±0.9	68.6±1.2	70.8±2.0
Multi-Modal	Roberta-102M	-	74.7±1.9	64.2±0.8	59.8±0.7	69.0±2.6	71.4±14.5	76.7±13.2	65.1±19.2	63.7±24.6	67.5±5.2	69.9±4.9
	GPT2-87M	-	71.0±3.4	61.5±1.1	60.6±0.3	74.0±8.5	65.4±12.9	81.8±25.5	73.1±20.8	54.1±12.9	83.2±21.5	86.1±19.8
	MolT5	EMNLP'22	69.9±0.8	64.7±0.9	55.1±0.9	65.1±0.5	76.7±2.1	65.3±1.7	55.9±1.1	49.2±1.0	70.3±0.8	73.1±1.0
	UniMol	ICLR'23	76.8±0.4	64.6±0.2	65.4±0.1	55.8±2.8	50.1±5.2	59.9±6.6	49.9±5.6	43.6±1.1	65.4±4.9	65.8±1.2
	CLOOME	NC'23	66.7±1.8	54.2±0.9	61.7±0.4	64.3±0.4	65.2±1.5	75.0±2.1	56.9±0.8	44.2±0.8	70.7±0.4	73.6±0.8
	InfoCORE (GE)	ICLR'24	79.3±0.9	65.3±0.2	60.2±0.2	69.9±1.2	79.9±3.6	80.3±0.9	51.6±1.8	51.3±2.1	78.6±0.3	77.8±1.9
	InfoCORE (CP)	ICLR'24	73.8±2.0	62.4±0.4	61.1±0.2	71.0±0.6	74.5±4.9	84.4±1.0	53.5±0.7	53.6±2.1	80.8±1.5	79.4±3.4
	InfoAlign	ICLR'25	<u>81.3±0.6</u>	<u>66.4±1.1</u>	<u>70.0±0.1</u>	<u>49.4±0.2</u>	<u>39.7±0.4</u>	<u>48.4±0.6</u>	<u>39.2±0.3</u>	<u>40.5±0.6</u>	66.7±1.7	<u>62.0±1.5</u>
	CHMR	**Ours**	**84.7±0.2**	**69.3±0.3**	**71.4±0.2**	**40.9±0.3**	**33.7±0.4**	**39.8±0.3**	**35.2±0.2**	**34.9±0.5**	**53.1±1.3**	**48.5±0.9**

Table 1: The performance comparison on four datasets is reported as mean±standard deviation. The **best** and <u>second-best</u> average scores are highlighted.

propagation weights $\beta_{i,l}$ are accumulated along the path to quantify the influence of neighboring nodes.

We then use the features projected into the shared latent space as inputs to decoders for reconstructing both molecular and external modalities:

$$\hat{\mathbf{x}}_i^{\text{anchor}} = g^m\big(\mathbf{p}_i^{\text{anch}}\big), \quad \hat{\mathbf{x}}_i^c = g^c\big(\mathbf{p}_i^c\big), \tag{10}$$

where $g^m(\cdot)$ and $g^c(\cdot)$ are the decoders for molecular and biological modalities, respectively.

To enhance reconstruction fidelity under cross-modal and hierarchical contexts, we define a unified reconstruction loss that adapts to feature types:

$$\mathcal{L}_{\text{CPR}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\sum_{l=0}^{L}\beta_{i,l}\,\mathcal{D}\Big(\hat{\mathbf{x}}_{u_{i_l}}, \mathbf{x}_{u_{i_l}}\Big), \tag{11}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes the reconstruction discrepancy, using binary cross-entropy (BCE) for discrete features and mean-squared error (MSE) for continuous ones.

4.5 Overall Objective

Pretraining To leverage the synergy of all components, we jointly optimize the loss functions of each module as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CPR}} + \lambda_1\mathcal{L}_{\text{SCA}} + \lambda_2\mathcal{L}_{\text{TreeVQ}}, \tag{12}$$

where λ_1 and λ_2 are balancing hyperparameters that control the relative importance of each module's contribution.

Downstream Evaluation For molecular property prediction, we freeze the pre-trained CHMR backbone and train a lightweight, task-specific prediction head $g_\theta(\cdot)$ to perform supervised prediction.

5 Experiments and Analysis

To systematically evaluate the effectiveness and generalization ability of the proposed method, we conducted experiments addressing the following key research questions:

RQ1. Does our method outperform baselines on property prediction?

RQ2. Are its key modules effective?

RQ3. Does hyperparameter tuning confirm effectiveness?

RQ4. Are alignment and hierarchy modeling effective?

5.1 Experimental Setup

Dataset We pretrain our model on molecular and cellular datasets from multiple sources, constructed from DrugBank (Wishart et al. 2018), Cell Painting images (Bray et al. 2016; Chandrasekaran et al. 2023), the JUMP-CP multi-omics platform (Chandrasekaran et al. 2023), and L1000 gene expression profiles (Subramanian et al. 2017). It contains 129,592 molecules with complete structural modalities and highly incomplete external biological modalities, where some modalities are missing for over 90% of molecules.

We evaluate our framework on benchmark datasets: ChEMBL (Gaulton et al. 2012), ToxCast (Richard et al. 2016), Broad (Moshkov et al. 2023), and Biogen (Fang et al. 2023), covering 696 prediction tasks across classification and regression settings.

Baselines To validate the effectiveness of our method, we compare it with over 20 baseline models: Single-Modality Models: Molecular fingerprint models (Rogers and Hahn 2010), AttentiveFP (Xiong et al. 2019), AttrMask/ContextPred/EdgePred (Hu et al. 2020b), GraphCL (You et al.

2020), GROVER (Rong et al. 2020), JOAO (You et al. 2021), MGSSL (Zhang et al. 2021), GraphLoG (Xu et al. 2021), GraphMAE (Hou et al. 2022), and DSLA (Kim, Baek, and Hwang 2022). Multi-Modality Models: RoBERTa-102M/GPT2-87M (Mary et al. 2024), MolT5 (Edwards et al. 2022), UniMol (Zhou et al. 2023), CLOOME (Sanchez-Fernandez et al. 2023), InfoCORE (Wang et al. 2024), and InfoAlign (Liu et al. 2025).

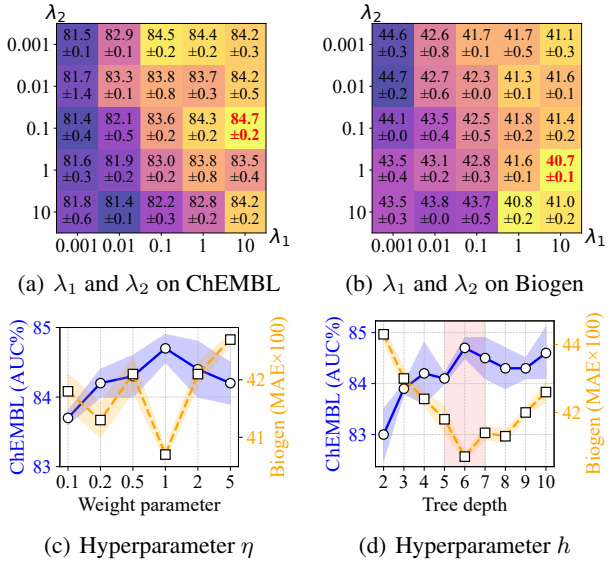
Implementation Details In our experiments, we carefully tune the hyperparameters for both pretraining and downstream evaluation to ensure stable convergence across diverse molecular property prediction tasks. We follow the training split strategy from (Liu et al. 2025), using a 0.6:0.25:0.15 ratio for training, validation, and testing sets on the ChEMBL, Broad, and Biogen datasets. For the ToxCast dataset, we adopt the scaffold-based splitting strategy, following the default 0.8:0.1:0.1 ratio used in OGB (Hu et al. 2020a). For classification tasks, we evaluate performance using AUC, while for regression tasks, we use MAE as the evaluation metric. All experiments are conducted using different random seeds (ranging from 0 to 4) and run five times, reporting the mean and standard deviation of the results. We perform experiments on an NVIDIA RTX 3090 GPU.

Model Variant	ChEMBL ↑	ToxCast ↑	Broad ↑	Biogen ↓	Δ %
Ours					
a. Full Model	**84.7±0.2**	**69.3±0.3**	**71.4±0.2**	**40.9±0.3**	–
Modality Augmentation					
b. Zero	81.6±0.4	66.4±0.4	68.7±0.3	44.8±0.4	−5.3
c. Random	81.9±0.5	66.9±0.5	69.0±0.3	44.1±0.5	−4.5
d. Neighbor	83.1±0.3	67.2±0.3	70.0±0.2	42.8±0.3	−2.9
Semantic Consistency Alignment					
e. w/o SCA	82.4±0.3	66.7±0.3	69.5±0.3	43.1±0.4	−3.6
f. w/o DA	83.5±0.3	68.2±0.3	70.5±0.2	41.9±0.3	−1.7
g. w/o IA	82.8±0.4	67.0±0.4	70.1±0.3	42.6±0.4	−2.9
Tree-VQ					
h. w/o Tree-VQ	82.3±0.3	66.9±0.3	69.0±0.3	43.4±0.4	−3.9
i. Flat VQ	83.2±0.3	67.9±0.3	70.4±0.2	42.0±0.3	−2.0
Context-Propagation Reconstruction					
j. w/o CPR	82.6±0.3	66.8±0.3	69.4±0.3	43.0±0.4	−3.5
k. w/o Walk	83.3±0.3	68.0±0.3	70.3±0.2	42.1±0.3	−2.0
Multi-Modal Synergy Verification					
l. Mol-Only	81.5±0.4	66.8±0.4	68.7±0.4	44.3±0.4	−4.9
m. Mol+Gene	82.7±0.3	67.5±0.3	70.0±0.3	43.6±0.5	−3.4
n. Mol+Cell	82.6±0.3	67.1±0.3	69.8±0.3	43.1±0.4	−3.3
o. Mol+Express	82.4±0.3	66.9±0.3	69.7±0.3	43.3±0.4	−3.6
SOTA Baseline					
p. InfoAlign	81.3±0.6	66.4±1.1	70.0±0.1	49.4±0.2	−7.7

Table 2: Ablation study of our framework. The Δ reports the average relative change (%) vs. Full Model.

5.2 Performance Evaluation

Molecular Property Prediction Comparison Results (RQ1) To validate the effectiveness of our method, we conduct comparative experiments on multiple molecular property and function prediction tasks. Table 1 presents the performance of our method on four representative drug molecule prediction datasets (ChEMBL, ToxCast, Broad, Biogen), where the first three are multi-task classification tasks used for drug activity prediction, biological response, or toxicity prediction, and the last one is a multi-task regression task for in vivo absorption, distribution, metabolism, and excretion (ADME) properties. The comparison methods are divided into two categories: single-modality and multi-modal methods. The experimental results show that our method achieves the best performance on all metrics. For instance, on Biogen, the average MAE is reduced by 17.2% compared to the second-best method, InfoAlign. Additionally, for classification tasks, we achieve an average improvement of approximately 2.0–4.4%. Notably, these gains are observed under biological multi-modal settings where missing modalities are common, demonstrating that our framework remains robust and effective in handling incomplete data.



(a) λ_1 and λ_2 on ChEMBL (b) λ_1 and λ_2 on Biogen

(c) Hyperparameter η (d) Hyperparameter h

Figure 3: Sensitivity analysis of hyperparameters on ChEMBL (AUC% ↑) and Biogen (MAE ×100 ↓).

Effectiveness of Key Modules (RQ2) To rigorously assess the contribution of each component in our framework, we conduct ablation studies (Table 2). For MA, naive zero imputation (variant b) and random imputation (c) significantly reduce performance by 5.3% and 4.5%, respectively, while neighborhood-based imputation (d) results in a smaller drop by 2.9%, indicating that leveraging neighborhood information through graph propagation yields more semantically coherent reconstructions than simple imputation strategies. Removing SCA (e) leads to a notable performance decrease by 3.6%, highlighting the importance of aligning augmented and original modality distributions. Ablating either distribution-level alignment (f) or instance-level alignment (g) alone also harms performance by 1.7% and 2.9%, respectively, demonstrating their complementary roles in mitigating inconsistencies during imputation. Similarly, excluding Tree-

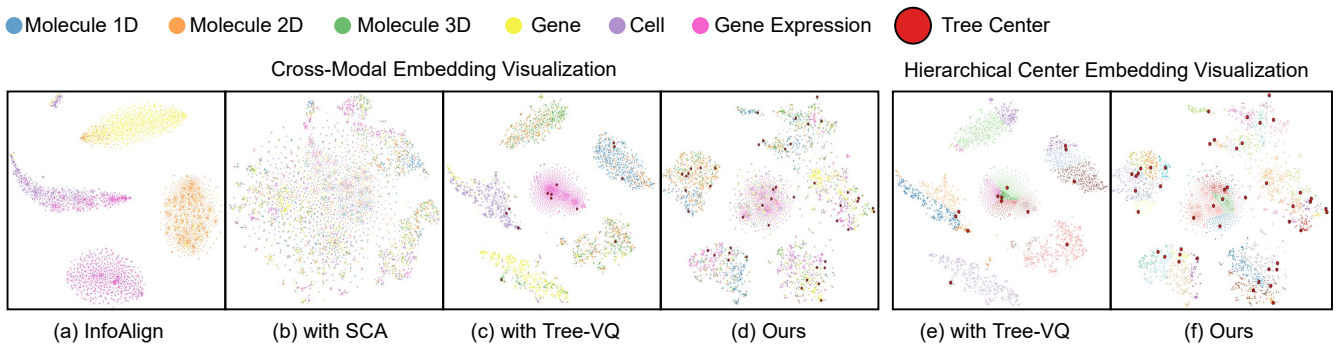


Figure 4: Visualization of cross-modal alignment and hierarchy. (a–d) show embeddings colored by modality; (e–f) display Tree codes with color-coded assignments and red dots indicating active hierarchical centers.

VQ (h) or replacing Tree-VQ with flat vector quantization (i) leads to performance degradation by 3.9% and 2.0%, respectively, indicating that hierarchical quantization better captures modality semantics and cross-modal relationships. For CPR, excluding the reconstruction loss (j) or disabling random walks in the biological graph (k) results in a performance decrease by 3.5% and 2.0%, respectively, confirming the benefit of contextual supervision in regularizing missing modalities. Finally, testing multi-modal synergy, using molecular features alone (l) leads to substantial performance drops by 4.9%, while combining modalities such as Mol+Gene (m), Mol+Cell (n), or Mol+Gene Expression (o) partially recovers performance, validating the benefit of integrating complementary biological information. Overall, our full model consistently outperforms all ablations and achieves superior results compared to the SOTA baseline, InfoAlign (p).

Robustness across Hyperparameters (RQ3) We evaluate the impact of key hyperparameters on model performance, including the loss balancing coefficients λ_1, λ_2, the weight parameter η, and the tree depth h. Figures 3(a)–(b) show grid search results over $\lambda_1, \lambda_2 \in \{0.001, 0.01, 0.1, 1, 10\}$. λ_1, which controls the weight of the SCA module, shows that small values (*e.g.*, 0.001 or 0.01) fail to enforce sufficient semantic consistency across modalities, leading to performance degradation. In contrast, larger values (*e.g.*, $\lambda_1{=}10$) yield the best results on both ChEMBL and Biogen datasets. λ_2 governs the strength of the Tree-VQ module. Moderate values (*e.g.*, 0.1 or 1) achieve optimal performance, indicating that appropriate hierarchical structural constraints help capture the latent dependencies among molecules, cells, and genes. However, overly large λ_2 values may impose excessive constraints and limit flexibility. Figure 3(c) investigates $\eta \in \{0.1, 0.2, 0.5, 1, 2, 5\}$, which controls the commitment weight between the encoder and tree nodes in Tree-VQ. The model achieves peak performance at $\eta{=}1$, suggesting that balanced bidirectional commitment effectively aligns the projections with the discrete semantic paths. Finally, Figure 3(d) examines the effect of tree depth $h \in \{2, 3, \ldots, 10\}$. Shallow trees ($h{\leq}4$) exhibit limited capacity and underperform, while very deep trees ($h{\geq}8$) may cause overfitting and semantic fragmentation. A moderate depth ($h{=}6$) strikes a good trade-off between expressiveness and generalization.

Cross-Modal Alignment and Hierarchical Representation Analysis (RQ4) Figure 4 presents two sets of t-SNE visualizations to evaluate the effectiveness of CHMR in aligning multi-modal features and capturing hierarchical dependencies. In the Cross-Modal Embedding Visualization (Figure 4(a–d)), we compare four methods in terms of modality alignment. Figure 4(a) (InfoAlign) shows that the four modalities remain largely separated, indicating poor alignment. Figure 4(b) (SCA only) achieves better alignment across modalities, but the resulting clusters are flat and lack hierarchical organization. Figure 4(c) (Tree-VQ only) captures hierarchical clusters but fails to align modalities, with each modality still forming disjoint clusters. In contrast, Figure 4(d) (CHMR) demonstrates both strong cross-modal alignment and clear hierarchical structuring, where modalities of the same sample overlap in the latent space and clusters are organized in a multi-level hierarchy. In the Hierarchical Center Embedding Visualization (Figure 4(e–f)), we further examine the Tree codes learned by different methods. Figure 4(e) shows that Tree-VQ alone utilizes only a limited number of Tree codes, resulting in poorly differentiated clusters. Figure 4(f) (CHMR) achieves full utilization of Tree codes and generates compact clusters with clear hierarchical branching, demonstrating its ability to capture fine-grained semantic dependencies across modalities.

6 Conclusion

We propose CHMR, a unified framework for cell-aware hierarchical multi-modal representation learning that integrates molecular structures with cellular and genomic responses. To address modality incompleteness and hierarchical dependencies, CHMR incorporates structure-aware propagation, semantic consistency alignment, context-guided reconstruction, and tree-structured vector quantization for robust cross-modal representation learning. Extensive experiments on benchmark datasets covering 696 molecular property prediction tasks demonstrate consistent and significant improvements over state-of-the-art baselines in both classification and regression. These results underscore CHMR's robustness to missing modalities and its ability to capture cross-scale biological mechanisms, offering a promising direction for biologically grounded molecular modeling and drug discovery.

Acknowledgments

This work was supported by the Scientific Research Innovation Capability Support Project for Young Faculty (No. ZYGXQNJSKYCXNLZCXM-I28), the Tongchuang Intelligent Medical Inter-disciplinary Talent Training Fund of Sun Yat-sen University (No. 76160-54990001), the National Natural Science Foundation of China (No. U21A20427), the National Key R&D Program of China (No. 2022ZD0115100), and the Center of Synthetic Biology and Integrated Bioengineering of Westlake University (No. WU2022A009), the Zhejiang Province Selected Funding for Postdoctoral Research Projects (No. ZJ2025113).

We thank the Westlake University HPC Center for providing computational resources. This work was supported by the InnoHK program. We thank Prof. Zhen Lei from the Center for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science and Innovation (HKISI) for his valuable suggestions and comments. We additionally thank Bo Li, a Ph.D. candidate at the University of Macau, for his helpful assistance during this work.

References

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In *Proceedings of the International Conference on Learning Representations*.

Bray, M.-A.; Gustafsdottir, S. M.; Rohban, M. H.; Singh, S.; Ljosa, V.; Sokolnicki, K. L.; Bittker, J. A.; Bodycombe, N. E.; Dančík, V.; Hasaka, T. P.; et al. 2017. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience*, 6(12): giw014.

Bray, M.-A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; and Carpenter, A. E. 2016. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9): 1757–1774.

Chandrasekaran, S. N.; Ackerman, J.; Alix, E.; Ando, D. M.; Arevalo, J.; Bennion, M.; Boisseau, N.; Borowa, A.; Boyd, J. D.; Brino, L.; et al. 2023. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023–03.

Chen, R.; Li, C.; Wang, L.; Liu, M.; Chen, S.; Yang, J.; and Zeng, X. 2025. Pretraining graph transformer for molecular representation with fusion of multimodal information. *Information Fusion*, 115: 102784.

Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; and Wang, F. 2023. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1): 6395.

Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; and Ji, H. 2022. Translation between Molecules and Natural Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 375–413.

Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; and Sciabola, S. 2023. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11): 3263–3274.

Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; and Wang, H. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2): 127–134.

Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C. W.; and Gadepally, V. 2023. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11): 1297–1305.

Gatto, L.; Aebersold, R.; Cox, J.; Demichev, V.; Derks, J.; Emmott, E.; Franks, A. M.; Ivanov, A. R.; Kelly, R. T.; Khoury, L.; et al. 2023. Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. *Nature Methods*, 20(3): 375–386.

Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1): D1100–D1107.

Gulati, G. S.; D'Silva, J. P.; Liu, Y.; Wang, L.; and Newman, A. M. 2025. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nature Reviews Molecular Cell Biology*, 26(1): 11–31.

Himmelstein, D. S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S. L.; Hadley, D.; Green, A.; Khankhanian, P.; and Baranzini, S. E. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6: e26726.

Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 594–604.

Hu, J.; Guo, D.; Si, Z.; Liu, D.; Diao, Y.; Zhang, J.; Zhou, J.; and Wang, M. 2025. MOL-Mamba: Enhancing Molecular Representation with Structural & Electronic Insights. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 317–325.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020a. Open graph benchmark: Datasets for machine learning on graphs. In *Proceedings of the Neural Information Processing Systems*, volume 33, 22118–22133.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020b. Strategies For Pre-training Graph Neural Networks. In *Proceedings of the International Conference on Learning Representations*.

Jia, L.; Ying, Y.; Qiu, T.; Yao, S.; Xue, L.; Lei, J.; Song, J.; Song, M.; and Feng, Z. 2025. Association Pattern-enhanced Molecular Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17599–17607.

Jiang, T.; Wang, Z.; Yu, S.; and Xuan, Q. 2025. Adaptive Substructure-Aware Expert Model for Molecular Property Prediction. *arXiv preprint arXiv:2504.05844*.

Kim, D.; Baek, J.; and Hwang, S. J. 2022. Graph self-supervised learning with accurate discrepancy learning. In *Proceedings of the Neural Information Processing Systems*, volume 35, 14085–14098.

Liu, A.; Seal, S.; Yang, H.; and Bender, A. 2023a. Using chemical and biological data to predict drug toxicity. *SLAS Discovery*, 28(3): 53–64.

Liu, G.; Seal, S.; Arevalo, J.; Liang, Z.; Carpenter, A. E.; Jiang, M.; and Singh, S. 2025. Learning molecular representation in a cell. *Proceedings of the International Conference on Learning Representations*.

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

Liu, Z.; Li, S.; Luo, Y.; Fei, H.; Cao, Y.; Kawaguchi, K.; Wang, X.; and Chua, T.-S. 2023b. MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mary, H.; Noutahi, E.; DomInvivo; Zhu, L.; Moreau, M.; Pak, S.; Gilmour, D.; Whitfield, S.; t; Valence-JonnyHsu; Hounwanou, H.; Kumar, I.; Maheshkar, S.; Nakata, S.; Kovary, K. M.; Wognum, C.; Craig, M.; and Bot, D. 2024. datamol-io/datamol: 0.12.3. *Zenodo, 0.12.3*.

Moshkov, N.; Becker, T.; Yang, K.; Horvath, P.; Dancik, V.; Wagner, B. K.; Clemons, P. A.; Singh, S.; Carpenter, A. E.; and Caicedo, J. C. 2023. Predicting compound activity from phenotypic profiles and chemical structures. *Nature Communications*, 14(1): 1967.

Nguyen, C. Q.; Pertusi, D.; and Branson, K. M. 2023. Molecule-Morphology Contrastive Pretraining for Transferable Molecular Representation. *bioRxiv*, 2023–05.

Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8): 1225–1251.

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

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33: 12559–12571.

Rossi, E.; Kenlay, H.; Gorinova, M. I.; Chamberlain, B. P.; Dong, X.; and Bronstein, M. M. 2022. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Proceedings of the Learning on Graphs Conference*, 11–1. PMLR.

Sanchez-Fernandez, A.; Rumetshofer, E.; Hochreiter, S.; and Klambauer, G. 2023. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1): 7339.

Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; et al. 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6): 1437–1452.

Wang, C.; Gupta, S.; Uhler, C.; and Jaakkola, T. S. 2024. Removing Biases from Molecular Representations via Information Maximization. In *Proceedings of the International Conference on Learning Representations*.

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1): D1074–D1082.

Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16): 8749–8760.

Xu, M.; Wang, H.; Ni, B.; Guo, H.; and Tang, J. 2021. Self-supervised graph-level representation learning with local and global structure. In *Proceedings of the International Conference on Machine Learning*, 11548–11558.

You, Y.; Chen, T.; Shen, Y.; and Wang, Z. 2021. Graph contrastive learning automated. In *Proceedings of the International Conference on Machine Learning*, 12121–12132.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. In *Proceedings of the Neural Information Processing Systems*, volume 33, 5812–5823.

Yu, Q.; Zhang, Y.; Ni, Y.; Feng, S.; Lan, Y.; Zhou, H.; and Liu, J. 2024. Multimodal Molecular Pretraining via Modality Blending. In *Proceedings of the International Conference on Learning Representations*.

Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J. C.; and Yang, S. 2025a. Artificial intelligence in drug development. *Nature Medicine*, 31(1): 45–59.

Zhang, Y.; Wang, H.; Butler, D.; To, M.-S.; Avery, J.; Hull, M. L.; and Carneiro, G. 2023. Distilling missing modality knowledge from ultrasound for endometriosis diagnosis with magnetic resonance images. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 1–5.

Zhang, Y.; Ye, G.; Yuan, C.; Han, B.; Huang, L.-K.; Yao, J.; Liu, W.; and Rong, Y. 2025b. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *Proceedings of the International Conference on Learning Representations*.

Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; and Lee, C.-K. 2021. Motif-based graph self-supervised learning for molecular property prediction. In *Proceedings of the Neural Information Processing Systems*, volume 34, 15870–15882.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *Proceedings of the International Conference on Learning Representations*.