GRAM-DTI: Adaptive Multimodal Representation Learning for Drug–Target Interaction Prediction

Feng Jiang[1*], **Amina Mollaysa**[2*], **Hehuan Ma**[1], **Tommaso Mansi**[2], **Junzhou Huang**[1], **Mangal Prakash**[2†], **Rui Liao**[2†]

[1]University of Texas at Arlington, [2]Johnson & Johnson Innovative Medicine

Abstract

Drug target interaction (DTI) prediction is central to computational drug discovery. While deep learning has advanced DTI modeling, existing approaches primarily rely on SMILES–protein pairs, failing to exploit rich multimodal information available for molecules and proteins. We introduce GRAM-DTI, a pre-training framework that integrates multimodal molecular and protein inputs into unified representations. GRAM-DTI extends volume-based contrastive learning to four modalities, capturing higher-order semantic alignment beyond pairwise approaches. We propose adaptive modality dropout to dynamically regulate each modality's contribution during pre-training, and incorporate IC_{50} activity measurements, when available, as weak supervision to ground representations in biologically meaningful interaction strengths. Experiments on four datasets demonstrate that GRAM-DTI consistently outperforms state-of-the-art baselines, highlighting the benefits of multimodal alignment and adaptive modality utilization for robust DTI prediction.

1 Introduction

Drug target interaction (DTI) prediction is central to computational drug discovery, enabling rational drug design, drug repurposing, and mechanistic insights [31]. While experimental screening remains reliable, computational methods are increasingly critical for prioritizing candidate drug–protein pairs, accelerating discovery and reducing costs [22, 13]. DTI prediction has evolved from similarity-based heuristics to deep learning approaches [26, 22]. Modern neural models learn directly from SMILES and amino acid sequences [23, 39, 15, 33], but remain largely restricted to SMILES–protein pairs, overlooking richer multimodal information that could yield more robust predictions.

While multimodal pre-training substantially improves molecular property prediction [18, 16], existing approaches rely on pairwise contrastive learning that cannot capture higher-order interdependencies [5]. Besides, they assume equal informativeness of all modalities, ignoring that data sources differ in quality and relevance. Additionally, valuable publicly available IC_{50} activity measurements remain underutilized during pre-training despite their biological relevance for DTI tasks.

We propose GRAM-DTI, a multimodal pre-training framework that integrates diverse small molecule and protein representations while accounting for varying modality informativeness. Our approach (Fig. 1) extends volume-based contrastive learning [5] to integrate multimodal small molecule and protein representations for DTI prediction while introducing an adaptive modality dropout scheme to dynamically regulate modality contributions during pre-training based on gradient-informed informativeness of different modalities. This prevents dominant but less informative modalities from overwhelming complementary signals. Additionally, we mine IC_{50} activity measurements from public databases when available and use them as weak auxiliary supervision to ground learned representations in biologically meaningful interaction strengths. Evaluation on four public datasets demonstrates consistent improvements with GRAM-DTI over state-of-the-art baselines.

[*]These authors contributed equally

[†]These authors contributed equally

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

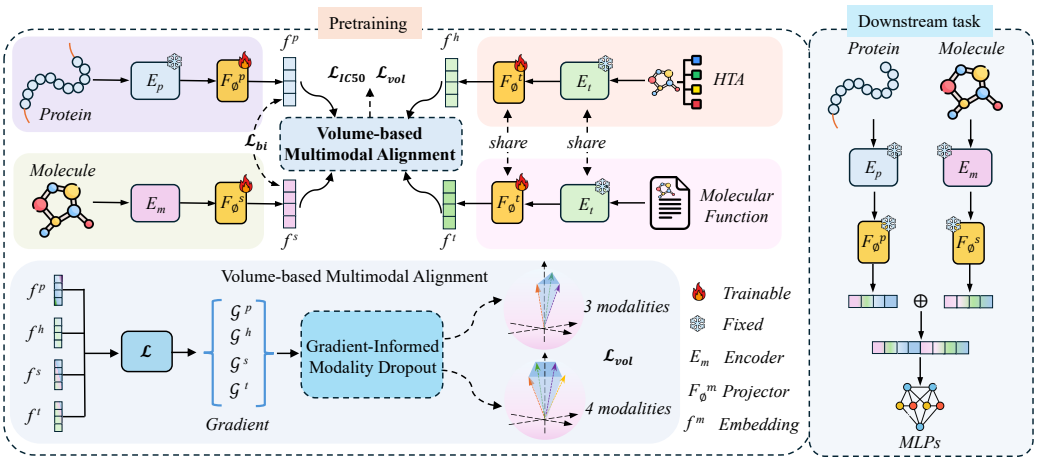

Figure 1: Overview of the GRAM-DTI architecture, including pre-training and downstream setup.

2 Methodology

Building upon recent advances [5, 11] in volume-based modality alignment for effective representation learning, we extend the foundational concept of volume loss [5], which was originally formulated for audio-video-text data, to the domain of protein-small molecule interactions. We aim to learn a unified embedding space that: 1) captures semantic relationships across modalities; 2) remains robust when modalities vary in informativenes; and 3) improves downstream DTI prediction task.

Formally, assume a pretraining dataset $D = \{(x_i^s, x_i^t, x_i^h, x_i^p, \delta_{y_i}^{IC50})\}_{i=1}^N$, where x_i^s, x_i^t, x_i^h, and x_i^p denote the SMILES sequence, textual description, hierarchical taxonomic annotation (HTA) [11], and protein sequence, respectively. The variable $\delta_{y_i}^{IC50}$ indicates the IC50 activity class y_i^{IC50} if a measured IC50 value is available for the protein–molecule pair (x_i^p, x_i^s), and 0 otherwise. As illustrated in Fig. 1, we employ pretrained encoders E_i (MolFormer[25] for SMILES, MolT5[6] for text and HTA, and ESM-2[14] for proteins) to obtain initial modality-specific embeddings. To keep pretraining efficient and scalable, we freeze the backbone encoders and train lightweight neural projectors F_ϕ^m that map each modality embedding into a shared representation space where they are semantically aligned. The resulting projected embeddings are denoted f^m, where $m \in \{\textit{SMILES, text, HTA, protein}\}$.

2.1 Gramian Volume-Based Multimodal Alignment

In contrast to traditional multi-modal representation learning approaches which have been known to fail in capturing the complex interdependencies among three or more modalities [5, 11], volume loss uses Gramian volume-based alignment of modailities ensuring semantic coherence across all modalities simultaneously.

Gramian Volume Given embeddings $f_i^s, f_i^t, f_i^h, f_i^p \in \mathbb{R}^d$ that are learned from the four modalities $x_i^s, x_i^t, x_i^h, x_i^p$ respectively, we first normalize them such that $\|f_i^m\|_2 = 1$. We can then construct the Gram matrix $G \in \mathbb{R}^{4\times4}$ where $G_{kj} = \langle f_i^k, f_i^j \rangle$, $k, j \in \{s, t, h, p\}$. The 4-dimensional volume spanned by these embedded vectors is equal to the square root of the determinat of the Gramian matrix [5]: $V(f_i^s, f_i^t, f_i^h, f_i^p) = \sqrt{\det(G)}$. From multimodal alignment perspective, smaller volume intuitively suggests stronger semantic alignment, as the embeddings occupy a more compact and cohesive subspace and vice-versa.

Volume-Based Contrastive Loss Given the Gramian volume, contrastive objective is cast as volume minimization/maximization. As proposed in [5], to construct negative pairs, we chose an anchor modality $a \in \{s, t, h, p\}$ as one of the four modalities. Therefore, for a batch of B samples, the contrastive loss on their learned embeddings is defined as follows:

$$\mathcal{L}_{\text{vol}}^{\rightarrow} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(-V(a_i, f_i^t, f_i^h, f_i^p)/\tau)}{\sum_{j=1}^{B'} \exp(-V(a_j, f_i^t, f_i^h, f_i^p))/\tau)}, \tag{1}$$

where, for example, the first modality f_i^s is chosen as the anchor a_i, negative pairs are constructed by permuting the anchor, and τ is the temperature parameter. We also add the reverse loss (w.r.t. negative pairs construction) to ensure symmetric alignment: $\mathcal{L}_{\text{vol}}^{\leftarrow} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(-V(a_i,f_i^t,f_i^h,f_i^p)/\tau)}{\sum_{j=1}^{B'}\exp(-V(a_i,f_j^t,f_j^h,f_j^p))/\tau)}$. The combined volume-based loss is $\mathcal{L}_{\text{vol}} = \frac{1}{2}(\mathcal{L}_{\text{vol}}^{\rightarrow} + \mathcal{L}_{\text{vol}})$.

2.2 Gradient-Informed Adaptive Modality Selection

While volume-based contrastive loss treats all modalities equally, different modalities may vary in quality and relevance, with contributions that change during training. Static fusion strategies risk either underutilizing weaker modalities or overfitting to dominant ones. We propose a gradient-informed modality dropout mechanism that dynamically adapts modality usage based on their instantaneous contribution to the loss function.

Gradient Contribution Analysis Assume $\mathcal{L}_{\tilde{t}}$ denotes mini-batch loss at training step $\tilde{t}$. We measure the importance of modality $m \in \{s,t,h,p\}$ by the magnitude of the gradient with respect to its embedding: $g_{\tilde{t}}^m = \left\| \frac{\partial \mathcal{L}_{\tilde{t}}}{\partial f_{\tilde{t}}^m} \right\|_2$, where $f_{\tilde{t}}^m \in \mathbb{R}^d$ is the learned embedding of modality m at gradient step $\tilde{t}$. To avoid noisy decisions, we track the history of gradient contributions over the past K steps: $\bar{g}_{\tilde{t}}^m = \frac{\sum_{k=0}^{K-1}\alpha^k g_{\tilde{t}-k}^m}{\sum_{k=0}^{K-1}\alpha^k}$, where $\alpha \in (0,1)$ is an exponential decay factor which yields a smooth, temporally discounted importance score for each modality.

Adaptive Modality Dropping Strategy We employ a principled adaptive strategy that considers both the magnitude and variance of gradient contributions. Let $\mu_{\tilde{t}} = \frac{1}{4}\sum_m \bar{g}_{\tilde{t}}^m$ and $\sigma_{\tilde{t}} = \sqrt{\frac{1}{4}\sum_m(\bar{g}_{\tilde{t}}^m - \mu_{\tilde{t}})^2}$ denote the mean and standard deviation of weighted gradients across modalities at the current gradient step $\tilde{t}$. We will drop a modality from the volume based contrastive loss calculation with a probability of p_{drop}, which is a hyperparameter. The criteria to drop a modality is defined as follows:

$$m_{\text{drop}}^{(\tilde{t})} = \begin{cases} \arg\max_m \bar{g}_{\tilde{t}}^m & \text{if dominance detected, e.g., } \bar{g}_{\tilde{t}}^m > \mu_{\tilde{t}} + \lambda_\sigma \sigma_{\tilde{t}}, \\ \arg\min_m \bar{g}_{\tilde{t}}^m & \text{otherwise,} \\ \text{none} & \text{with probability } (1 - p_{\text{drop}}). \end{cases}$$

where $\lambda_\sigma = 1.5$ is the threshold multiplier. This means that we adaptively drop modalities based on two criteria: 1) *Dominance prevention*: if a modality's contribution is much larger than others, we drop it to avoid overfitting; 2) *Low-contribution pruning*: Otherwise, we drop the modality with the smallest gradient contribution to encourage use of more informative signals. This dynamic selection balances stability and diversity, ensuring all modalities remain engaged throughout training.

2.3 Weak Supervision Through IC50 Activity Measure

As the IC50 values for wide range of protein-small molecule pairs are availabe on public data sources (BindingDB [8]), we introduce an additional classification task as an auxiliary objective during pre-training. However, IC50 labels are not available for all possible protein-small molecule pairs, this task provides only weak supervisory signal during pre-training when IC50 information is available. We train a classifier F_ϕ^{IC50} to predict the IC50 class from the learned embeddings of all four modalities: $f^{\text{fused}} = [f^s; f^t; f^h; f^p] \in \mathbb{R}^{4d}$. Note that IC50 values are continuous, but given the inherent challenges of IC50 regression, including heterogeneous value distributions, wide dynamic ranges spanning several orders of magnitude, and noisy measurements, we formulate the problem as a three-class classification task by employing discretizations on IC50 values (see Appendix B). However, this discretization comes with class-imbalance described in Appendix B. To address this issue, we employ a weighted cross-entropy loss: $\mathcal{L}_{\text{IC50}} = -\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} w_{y_i}\log p(y_i|f_i^{\text{fused}})$, where $\mathcal{S}$ denotes the set of samples with valid IC50 annotations, and class weights are computed as: $w_c = \frac{N_{\text{total}}}{C\cdot N_c}$, where N_{total} being the total number of samples, C the number of classes, and N_c the number of samples in class c.

Auxiliary Bimodal Contrastive Loss As the downstream task involves protein and molecule embeddings, to emphasize alignment between these two, we also incorporate traditional pairwise contrastive losses between SMILES and protein modalities:$\mathcal{L}_{\text{bi}} = \frac{1}{2}(\mathcal{L}_{s\rightarrow p} + \mathcal{L}_{p\rightarrow s})$ where $\mathcal{L}_{s\rightarrow p}$ and $\mathcal{L}_{p\rightarrow s}$ follow the standard CLIP-style contrastive formulation.

Table 1: Performance comparison on DTI and MoA prediction benchmarks.

Data	Metric	Scenario	CPL-GNN	MPNN-CNN	TransformerCPI	KGE-NFM	DTIAM	GRAM-DTI	Data	AI-DTI	DTIAM	GRAM-DTI
Yamanishi_08	AUPR	Warm start	0.431	0.816	0.802	0.817	0.901	**0.904**	Activation	0.583	0.623	**0.642**
		Drug cold start	0.167	0.408	0.410	0.341	0.439	**0.440**		0.550	0.611	**0.628**
		Target cold start	0.380	0.602	0.646	0.761	0.844	**0.849**		0.219	0.391	**0.470**
	AUROC	Warm start	0.821	0.952	0.953	0.948	0.967	**0.977**		0.888	0.903	**0.914**
		Drug cold start	0.629	0.797	0.767	0.779	0.818	**0.828**		0.879	0.907	**0.913**
		Target cold start	0.800	0.856	0.870	0.923	0.941	**0.955**		0.652	0.792	**0.834**
Hetionet	AUPR	Warm start	0.441	0.734	-	0.789	**0.879**	0.864	Inhibition	0.840	**0.845**	0.776
		Drug cold start	0.219	0.453	-	0.391	0.514	**0.535**		**0.830**	0.731	0.721
		Target cold start	0.433	0.470	-	0.612	**0.799**	0.630		0.215	0.397	**0.484**
	AUROC	Warm start	0.810	0.956	-	0.968	0.957	**0.982**		0.952	**0.954**	0.950
		Drug cold start	0.685	0.831	-	0.803	0.752	**0.863**		**0.948**	0.921	0.940
		Target cold start	0.810	0.858	-	0.915	0.917	**0.921**		0.605	0.819	**0.823**

2.4 Unified Training Objective

The complete training objective integrates all components with appropriate weighting: $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{vol}} + \lambda_2 \mathcal{L}_{\text{bi}} + \lambda_3 \mathcal{L}_{\text{IC50}}$ where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. Note that $\mathcal{L}_{\text{vol}}$ and $\mathcal{L}_{\text{bi}}$ are applied on all the training instances while $\mathcal{L}_{\text{IC50}}$ are only applied for pairs of protein and molecule with valid IC50 annotations. For gradient-based dropping of a modality in volume contrastive loss, we use $\mathcal{L} = \lambda_2 \mathcal{L}_{\text{bi}} + \lambda_3 \mathcal{L}_{\text{IC50}}$. See Appendix E for model architecture and parameters.

3 Experiments

For pre-training, we employ the dataset proposed in [11] consisting of 47,269 triplets of SMILES, text descriptions, and HTA annotations, and extend it with BindingDB[8] protein binding data and IC50 information when available. To prevent data leakage between pretraining and downstream evaluation, we removed overlapping $(\text{SMILES}, \text{protein})$ pairs from the pretraining set. After filtering, the dataset contained 6,545 unique molecules and 4,418 proteins, forming a total of 50,968 quadruplets with 16,035 containing IC_{50} measurements.For downstream DTI evaluation, we used four benchmark datasets from DTIAM [17]: Activation (1,913 interactions, 1,426 drugs, 281 targets), Yamanishi_08 (5,127 DTIs, 791 drugs, 989 targets), Hetionet (49,942 DTIs, 1,384 drugs, 5,763 targets), and Inhibition (21,055 interactions, 14,049 drugs, 1,088 targets). See Appendix C for details.

To assess generalization, we follow the splits in [17] comprising 1) warm start: no common protein-molecule pairs in train/test. 2) drug cold start: no molecules shared between sets. 3) target cold start: no proteins shared. These evaluate performance on unseen pairs, molecules, or proteins. We follow the DTIAM framework: 10-fold cross-validation for DTI tasks (Yamanishi_08, Hetionet) and 5-fold for MoA tasks (Activation, Inhibition), reporting mean and std across folds. We compared GRAM-DTI's performance with state-of-the-art models across all benchmark datasets to demonstrate its effectiveness. Table 1 compares GRAM-DTI with five baselines: CPL-GNN [30], MPNN-CNN [7], TransformerCPI [4], and KGE-NFM [35] on the Yamanishi_08 Hetionet dataset and with AI-DTI [12], DTIAM [17]on Activation and Inhibation dataset, showing significant improvements across datasets and evaluation scenarios.

GRAM-DTI demonstrates strong performance across benchmark datasets, particularly excelling in target cold start scenarios. On Yamanishi_08, our method achieves significant improvements in both warm start and target cold start settings. On the larger Hetionet dataset, GRAM-DTI outperforms most baselines in multiple scenarios. For MoA prediction, GRAM-DTI consistently outperforms baselines on the Activation dataset, especially in target cold start scenarios. On the Inhibition dataset, GRAM-DTI demonstrates excellent target cold start performance. These results suggest that activation mechanisms rely more on sequence and semantic features captured by our multimodal approach. The strong performance in target cold start scenarios highlights the robustness of our volume-based multimodal alignment for discovering interactions with novel protein targets. Additional experimental analysis and ablation studies are provided in Appendix D.

4 Conclusion

We presented GRAM-DTI, a multimodal pretraining framework that extends volume-based contrastive learning to four modalities with gradient-informed adaptive modality dropout and IC_{50} auxiliary supervision. Evaluation across four benchmark datasets shows GRAM-DTI consistently outperforms baselines, particularly in cold start scenarios. Ablation studies (Appendix section D) confirm synergistic contributions of each component. These results highlight the potential of multimodal pretraining for drug discovery, where integrating diverse data sources leads to more robust prediction models. Future work could explore additional modalities and extend our adaptive mechanism to other domains.

References

[1] Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20208–20217, 2022.

[2] Faisal Bin Ashraf, Sanjida Akter, Sumona Hoque Mumu, Muhammad Usama Islam, and Jasim Uddin. Bio-activity prediction of drug candidate compounds targeting sars-cov-2 using machine learning approaches. *Plos one*, 18(9):e0288053, 2023.

[3] Rohit Bavi, Raj Kumar, Light Choi, and Keun Woo Lee. Exploration of novel inhibitors for bruton's tyrosine kinase by 3d qsar modeling and molecular dynamics simulation. *PloS one*, 11(1):e0147190, 2016.

[4] Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

[5] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*, 2024.

[6] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[7] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr, 2017.

[8] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

[9] Daniel S Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

[10] Feng Jiang, Mangal Prakash, Hehuan Ma, Jianyuan Deng, Yuzhi Gao, Amina Mollaysa, Tommaso Mansi, Rui Liao, and Junzhou Huang. TRIDENT: Tri-modal molecular representation learning with taxonomic annotations and local correspondence. In *ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

[11] Feng Jiang, Mangal Prakash, Hehuan Ma, Jianyuan Deng, Yuzhi Guo, Amina Mollaysa, Tommaso Mansi, Rui Liao, and Junzhou Huang. Trident: Tri-modal molecular representation learning with taxonomic annotations and local correspondence. *arXiv preprint arXiv:2506.21028*, 2025.

[12] Won-Yung Lee, Choong-Yeol Lee, and Chang-Eop Kim. Predicting activatory and inhibitory drug–target interactions based on structural compound representations and genetically perturbed transcriptomes. *PLoS One*, 18(4):e0282042, 2023.

[13] Qian Liao, Yu Zhang, Ying Chu, Yi Ding, Zhen Liu, Xianyi Zhao, Yizheng Wang, Jie Wan, Yijie Ding, Prayag Tiwari, et al. Application of artificial intelligence in drug-target interactions prediction: a review. *npj biomedical innovations*, 2(1):1, 2025.

[14] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[15] Sizhe Liu, Yuchen Liu, Haofeng Xu, Jun Xia, and Stan Z Li. Sp-dti: subpocket-informed transformer for drug–target interaction prediction. *Bioinformatics*, 41(3):btaf011, 2025.

[16] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023.

[17] Zhangli Lu, Guoqiang Song, Huimin Zhu, Chuqi Lei, Xinliang Sun, Kaili Wang, Libo Qin, Yafei Chen, Jing Tang, and Min Li. Dtiam: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nature Communications*, 16(1):2548, 2025.

[18] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

[19] Nicholas Magal, Minh Tran, Riku Arakawa, and Suzanne Nie. Negative to positive co-learning with aggressive modality dropout. *arXiv preprint arXiv:2501.00865*, 2025.

[20] Tin Nguyen, Hien Le, Timothy P Quinn, Thin Nguyen, Trung Le, and Svetha Venkatesh. Graphdta: prediction of drug–target binding affinity using graph convolutional networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[21] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[22] Fatemeh Panahandeh and Najme Mansouri. A comprehensive review of neural network-based approaches for drug–target interaction prediction. *Molecular Diversity*, pages 1–48, 2025.

[23] Lihong Peng, Xin Liu, Min Chen, Wen Liao, Jiale Mao, and Liqian Zhou. Mgndti: a drug-target interaction prediction framework based on multimodal representation learning and the gating mechanism. *Journal of Chemical Information and Modeling*, 64(16):6684–6698, 2024.

[24] Abid Qureshi, Himani Tandon, and Manoj Kumar. Avp-ic50pred: multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (ic50). *Peptide Science*, 104(6):753–763, 2015.

[25] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[26] Wen Shi, Hong Yang, Linhai Xie, Xiao-Xia Yin, and Yanchun Zhang. A review of machine learning-based methods for predicting drug–target interactions. *Health Information Science and Systems*, 12(1):30, 2024.

[27] Bonggun Shin, Sanghyun Park, Kyungsook Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug–target interaction. *Proceedings of Machine Learning Research*, 106:955–970, 2019.

[28] Hannes Stärk, Octavian-Eugen Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

[29] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[30] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

[31] Ali Vefghi, Zahed Rahmati, and Mohammad Akbari. Drug-target interaction/affinity prediction: Deep learning models and advances review. *Computers in Biology and Medicine*, 196:110438, 2025.

[32] Feng Wan, Lin Hong, An Xiao, Tong Jiang, and Jianyang Zeng. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1237–1245, 2019.

[33] Xiaoqiong Xia, Chaoyu Zhu, Fan Zhong, and Lei Liu. Mdtips: a multimodal-data-based drug–target interaction prediction system fusing knowledge, gene expression profile, and structural data. *Bioinformatics*, 39(7):btad411, 2023.

[34] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[35] Qing Ye, Chang-Yu Hsieh, Ziyi Yang, Yu Kang, Jiming Chen, Dongsheng Cao, Shibo He, and Tingjun Hou. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications*, 12(1):6775, 2021.

[36] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[37] Qingyu Zhang, Qing Yang, Yue Zhang, and Buzhou Tang. Biot5+: Towards general-purpose large language models for the life sciences. *arXiv preprint arXiv:2306.02275*, 2023.

[38] Siqin Zhang, Kuo Yang, Zhenhong Liu, Xinxing Lai, Zhen Yang, Jianyang Zeng, and Shao Li. Drugai: a multi-view deep learning model for predicting drug–target activating/inhibiting mechanisms. *Briefings in bioinformatics*, 24(1), 2023.

[39] Yanpeng Zhao, Yuting Xing, Yixin Zhang, Yifei Wang, Mengxuan Wan, Duoyun Yi, Chengkun Wu, Shangze Li, Huiyan Xu, Hongyang Zhang, et al. Evidential deep learning-based drug-target interaction prediction. *Nature Communications*, 16(1):6915, 2025.

[40] Ying Zhou, Yintao Zhang, Xichen Lian, Fengcheng Li, Chaoxin Wang, Feng Zhu, Yunqing Qiu, and Yuzong Chen. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic acids research*, 50(D1):D1398–D1407, 2022.

A Related Works

Multimodal Molecular Representation Learning Recent advancements in molecular representation learning have shifted towards integrating multiple data modalities to enhance predictive performance. For instance, frameworks like TRIDENT [10] combine SMILES strings, hierarchical taxonomic annotations, and functional text of small molecules to capture richer molecular semantics. These approaches leverage contrastive learning to align diverse data sources, even in the absence of fully paired datasets, thereby improving generalization across various molecular tasks. Beyond TRIDENT, several molecular foundation models have been introduced, including MolFM [18] and MolCA [16], which integrate molecular graphs, textual descriptions, and domain-specific annotations into unified representations. Similarly, cross-modal pretraining strategies such as BioT5+ [37] demonstrate that incorporating protein, molecular, and textual modalities at scale improves transferability to downstream tasks such as property prediction and binding affinity estimation. These works highlight the broader trend of leveraging multimodal pretraining to construct general-purpose molecular representations.

Drug–Target Interaction (DTI) Prediction DTI prediction has traditionally relied on unimodal representations, such as SMILES strings for drugs and amino acid sequences for proteins. Early deep learning models such as DeepDTA [21], MT-DTI [27], and TransformerCPI [4] demonstrated the effectiveness of sequence-based architectures for interaction prediction. Beyond sequence-based methods, more recent work has explored graph neural networks and SE(3)-equivariant geometric deep learning models, such as GraphDTA [20] and EquiBind [28], which leverage spatial and structural information of drugs and proteins to enhance binding affinity prediction. In parallel, knowledge graph–based methods such as NeoDTI [32] and Hetionet-based repurposing frameworks [9] exploit biomedical networks to capture higher-order relations among drugs, targets, and diseases. More recently, multimodal approaches have been proposed to better capture the complexity of drug–target interactions. For example, MDTips [33] integrates knowledge graphs, gene expression profiles, and structural information, while MGNDTI [23] employs a multimodal graph neural network to improve robustness and generalization.

Another emerging direction is pretraining with large-scale unlabeled data to mitigate the scarcity of labeled DTI pairs. For instance, DTIAM [17] introduces separate pretraining for drug and target modalities before merging the learned representations for DTI prediction. In contrast, our framework performs *joint multimodal pretraining* that integrates multiple drug and protein modalities into a unified space from the outset. This allows us to capture higher-order semantic relationships beyond pairwise fusion.

Modality Dropout Modality dropout techniques have been proposed to enhance the robustness of multimodal models by preventing over-reliance on any single modality. For instance, the Learnable Irrelevant Modality Dropout (IMD) method [1] selectively drops irrelevant modalities during training, improving performance in multimodal action recognition tasks. Additionally, approaches like aggressive modality dropout have been shown to mitigate negative co-learning effects and enhance model accuracy in multimodal settings [19]. Beyond dropout, adaptive fusion mechanisms have also been investigated. Cross-attention and gating strategies [29, 23] dynamically regulate modality contributions, while tensor fusion methods [36] capture higher-order interactions across modalities. These ideas inform the design of adaptive strategies in molecular contexts, where modality informativeness often varies across data sources and training stages.

B IC50 Values discretizations

Given the inherent challenges of IC50 regression—including heterogeneous value distributions, wide dynamic ranges spanning several orders of magnitude, and noisy measurements—we formulate the problem as a three-class classification task. The IC50 values are discretized based on pharmaceutical relevance thresholds:

$$\text{IC50 class} = \begin{cases} 0 & \text{if IC50} < 10\mu\text{M (effective)} \\ 1 & \text{if } 10\mu\text{M} \leq \text{IC50} \leq 1000\mu\text{M (moderate)} \\ 2 & \text{if IC50} > 1000\mu\text{M (ineffective)} \end{cases} \tag{2}$$

This discretization strategy aligns with established drug discovery practices [24, 3, 2] where compounds with IC50 $< 10\mu$M are considered highly active, those between $10 - 1000\mu$M show moderate activity, and those $> 1000\mu$M are typically considered inactive.

C Dataset

Pretraining Data Our pretraining dataset builds upon the high-quality multimodal molecular dataset from TRIDENT [11], which provides comprehensive molecular representations through the integration of SMILES strings, natural language descriptions, and Hierarchical Taxonomic Annotations (HTA). The original TRIDENT dataset contains 47,269 carefully curated ⟨SMILES, Text, HTA⟩ triplets sourced from PubChem, where each molecule is annotated across 32 diverse taxonomic classification systems.

To enable protein-molecule interaction modeling, we extended this dataset by incorporating binding affinity information from BindingDB, a comprehensive database of measured binding affinities for protein-molecule interactions. We mapped molecules from the TRIDENT dataset to BindingDB entries using molecular identifiers, creating 5-tuples of the form ⟨SMILES, Text, HTA, Protein, IC_{50}⟩. This integration combines the rich semantic and structural information from TRIDENT with quantitative binding affinity measurements, providing a unified multimodal representation that captures both molecular properties and protein-molecule interactions. Following standard practices in molecular property prediction, we implemented careful data filtering to prevent information leakage between pretraining and downstream evaluation. Specifically, we removed all SMILES-protein binding pairs that appear in our downstream task datasets to ensure fair evaluation and prevent overfitting to specific molecular-protein combinations seen during pretraining.

After filtering, 6,545 unique molecules have associated protein binding information. Considering that each molecule can interact with multiple proteins, this results in a total of 50,968 quadruplets ⟨Protein, SMILES, Text, HTA⟩, covering 4,418 unique proteins. Among these quadruplets, 16,035 entries include quantitative IC_{50} measurements, providing high-quality binding affinity annotations for modeling.

Downstream Task Datasets We evaluated our approach on four benchmark datasets (see Table 2) from the DTIAM framework [17], covering drug-target interaction (DTI) prediction and mechanism of action (MoA) prediction tasks. 1) **Activation dataset** obtained from the Therapeutic Target Database (TTD) [40], containing 1,426 drugs, 281 targets, and 1,913 known activation interactions. 2) **Yamanishi_08** originally introduced by [34] and consists of four sub-datasets: G-Protein Coupled Receptors (GPCR), Ion Channels (IC), Nuclear Receptors (NR), and Enzymes (E). We use the combined dataset constructed by [35], containing 791 drugs, 989 targets, and 5,127 known DTIs. 3) **Hetionet dataset** constructed by [9], which integrated biomedical data from 29 public resources, comprising 1,384 drugs, 5,763 targets, and 49,942 DTIs. 4) **Inhibition dataset** also derived from TTD [40], containing 14,049 drugs, 1,088 targets, and 21,055 known inhibition interactions.

Table 2: Statistics of downstream task datasets for binary classification. Known Interactions represents the number of positive drug-target binding pairs, while Total Samples includes both positive samples and 10 times negative samples generated following standard practice.

Dataset	Task Type	Drugs	Targets	Known Interactions	Total Samples
Yamanishi_08	DTI	791	989	5,127	56,397
Hetionet	DTI	1,384	5,763	49,942	549,362
Activation	MoA	1,426	281	1,913	21,043
Inhibition	MoA	14,049	1,088	21,055	231,605

The MoA refers to how a drug works on its target to produce the desired effects, which involve two major roles: activation and inhibition mechanisms. Distinguishing the activation and inhibition MoA between drugs and targets is critical and challenging in the drug discovery and development process, as well as their clinical applications [38].

C.1 Experimental Setup

In the DTI and MoA prediction task, the objective is to determine whether a given drug-target pair interacts, which constitutes a binary classification problem. Note that dataset only includes those pairs that interacts (positive class). Following standard practice, we generated negative samples using a 1:10 ratio with positive samples for all datasets. To evaluate the model's generalization performance, we employed three different data splitting strategies for train-test division: 1) *warm start*: The data is split based on protein-molecule pairs, ensuring that no common pairs appear in both the training and test sets. 2) *drug cold start*: This split is performed at the molecule level, guaranteeing that no drug in the test set is present in the training set. 3) *target cold start*: Similar to the above, but split at the protein level, meaning no protein in the test set is seen during training. These three settings allow us to assess how well the model performs when faced with unseen molecule-protein pairs, unseen molecules, or unseen proteins, respectively. For evaluation, we followed the cross-validation protocols established in the original DTIAM framework: 10-fold cross-validation for DTI prediction tasks (Yamanishi_08 and Hetionet datasets) and 5-fold cross-validation for MoA prediction tasks (Activation and Inhibition datasets). All results report the mean and standard deviation across folds.

Table 3: Performance of GRAM-DTI under different training objectives across data splits on Activation dataset.

Experiment	Split Type	AUROC ↑	AUPRC↑	Sensitivity↑	F1 ↑	Accuracy ↑
Exp 1: $\mathcal{L}_{\text{total}}$	warm start	**0.914±0.008**	**0.642±0.022**	**0.516±0.024**	**0.600±0.007**	**0.936±0.002**
	drug cold start	**0.913±0.007**	**0.628±0.022**	**0.514±0.035**	**0.588±0.019**	**0.935±0.003**
	target cold start	**0.834±0.026**	**0.470±0.047**	**0.331±0.058**	**0.463±0.053**	**0.922±0.006**
Exp 2: $\lambda_2\mathcal{L}_{\text{bi}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.903±0.009	0.625±0.022	0.504±0.026	0.580±0.020	0.934±0.004
	drug cold start	0.901±0.011	0.628±0.018	0.457±0.016	0.553±0.011	0.934±0.002
	target cold start	0.813±0.020	0.448±0.044	0.269±0.033	0.381±0.035	0.921±0.008
Exp 3: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.876±0.008	0.506±0.012	0.320±0.074	0.420±0.061	0.919±0.006
	drug cold start	0.884±0.010	0.514±0.027	0.353±0.054	0.448±0.043	0.922±0.001
	target cold start	0.805±0.020	0.385±0.032	0.253±0.042	0.335±0.034	0.909±0.013
Exp 4: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_2\mathcal{L}_{\text{bi}}$	warm start	0.903±0.005	0.605±0.014	0.476±0.037	0.556±0.024	0.931±0.005
	drug cold start	0.904±0.010	0.605±0.027	0.444±0.047	0.538±0.047	0.932±0.001
	target cold start	0.833±0.021	0.450±0.026	0.308±0.035	0.405±0.028	0.918±0.009
Exp 5: w/o Modality Dropout	warm start	0.886±0.007	0.590±0.009	0.464±0.012	0.545±0.010	0.930±0.005
	drug cold start	0.884±0.006	0.575±0.012	0.455±0.037	0.538±0.020	0.929±0.001
	target cold start	0.815±0.021	0.444±0.018	0.292±0.047	0.394±0.047	0.919±0.009

Table 4: Performance of GRAM-DTI under different training objectives across different data split strategies on Yamanishi_08 dataset.

Experiment	Split Type	AUROC↑	AUPRC↑	Sensitivity ↑	F1 ↑	Accuracy↑
Exp 1: $\mathcal{L}_{\text{total}}$	warm start	**0.977±0.005**	0.904±0.016	0.802±0.035	0.844±0.011	0.973±0.003
	drug cold start	**0.828±0.027**	**0.440±0.059**	**0.196±0.065**	**0.302±0.076**	**0.919±0.014**
	target cold start	**0.955±0.013**	**0.849±0.034**	**0.727±0.048**	**0.790±0.030**	**0.965±0.004**
Exp 2: $\lambda_2\mathcal{L}_{\text{bi}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.971±0.005	**0.906±0.012**	**0.804±0.021**	**0.848±0.015**	**0.974±0.002**
	drug cold start	0.801±0.046	0.392±0.060	0.153±0.084	0.235±0.117	0.916±0.013
	target cold start	0.949±0.017	0.843±0.039	0.721±0.054	0.784±0.040	0.964±0.005
Exp 3: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.959±0.006	0.844±0.012	0.694±0.034	0.769±0.016	0.962±0.002
	drug cold start	0.793±0.045	0.373±0.062	0.112±0.066	0.182±0.098	0.913±0.014
	target cold start	0.941±0.015	0.805±0.028	0.641±0.059	0.731±0.036	0.958±0.004
Exp 4: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_2\mathcal{L}_{\text{bi}}$	warm start	0.961±0.005	0.851±0.016	0.704±0.032	0.781±0.020	0.964±0.003
	drug cold start	0.775±0.065	0.378±0.088	0.150±0.071	0.237±0.103	0.916±0.012
	target cold start	0.942±0.015	0.814±0.035	0.675±0.046	0.754±0.035	0.960±0.005
Exp 5: w/o Modality Dropout	warm start	0.966±0.007	0.871±0.017	0.750±0.019	0.807±0.017	0.967±0.003
	drug cold start	0.798±0.036	0.381±0.072	0.147±0.050	0.237±0.071	0.916±0.013
	target cold start	0.949±0.010	0.827±0.028	0.689±0.048	0.763±0.025	0.961±0.005

D Ablation Study

Note that our main objective consists of three components (Eq.2.4). To evaluate the contribution of each component, we conducted a comprehensive ablation study, comparing the performance of our model with each component systematically removed. This results in five experimental setups:

Table 5: Performance of GRAM-DTI under different training objectives across different data split strategies on Inhibition dataset.

Experiment	Split Type	AUROC↑	AUPRC↑	Sensitivity ↑	F1 ↑	Accuracy↑
Exp 1: $\mathcal{L}_{\text{total}}$	warm start	**0.950±0.002**	0.785±0.006	0.659±0.011	0.720±0.006	0.954±0.001
	drug cold start	**0.940±0.002**	**0.756±0.003**	0.595±0.018	0.680±0.008	0.949±0.001
	target cold start	**0.823±0.021**	**0.464±0.056**	0.258±0.087	0.369±0.087	**0.922±0.009**
Exp 2: $\lambda_2\mathcal{L}_{\text{bi}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.947±0.001	**0.787±0.005**	**0.660±0.001**	**0.725±0.004**	**0.955±0.001**
	drug cold start	0.934±0.002	0.747±0.007	**0.598±0.011**	**0.681±0.007**	**0.949±0.001**
	target cold start	0.818±0.042	0.463±0.058	**0.298±0.080**	**0.408±0.079**	**0.923±0.008**
Exp 3: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.925±0.002	0.697±0.006	0.507±0.013	0.613±0.007	0.942±0.001
	drug cold start	0.917±0.002	0.675±0.007	0.483±0.039	0.592±0.024	0.940±0.001
	target cold start	0.821±0.025	0.441±0.064	0.218±0.089	0.320±0.105	0.920±0.008
Exp 4: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_2\mathcal{L}_{\text{bi}}$	warm start	0.923±0.002	0.688±0.005	0.472±0.032	0.590±0.019	0.940±0.001
	drug cold start	0.914±0.003	0.667±0.005	0.461±0.017	0.577±0.012	0.939±0.001
	target cold start	0.824±0.025	0.436±0.062	0.213±0.074	0.320±0.088	0.920±0.008
Exp 5: w/o Modality Dropout	warm start	0.933±0.002	0.725±0.006	0.558±0.013	0.652±0.007	0.946±0.001
	drug cold start	0.922±0.002	0.692±0.004	0.511±0.015	0.612±0.008	0.941±0.001
	target cold start	0.827±0.024	0.442±0.063	0.227±0.096	0.330±0.110	0.920±0.010

Table 6: Performance of GRAM-DTI under different training objectives across different data split strategies on Hetionet dataset.

Experiment	Split Type	AUROC ↑	AUPRC ↑	Sensitivity ↑	F1↑	Accuracy↑
Exp 1: $\mathcal{L}_{\text{total}}$	warm start	**0.982±0.001**	**0.864±0.004**	**0.767±0.010**	**0.793±0.005**	**0.964±0.001**
	drug cold start	0.863±0.040	0.535±0.069	**0.298±0.062**	**0.420±0.064**	**0.927±0.012**
	target cold start	0.921±0.007	**0.630±0.022**	0.430±0.030	0.530±0.022	0.932±0.004
Exp 2: $\lambda_2\mathcal{L}_{\text{bi}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.981±0.001	0.863±0.004	0.760±0.008	0.792±0.005	**0.964±0.001**
	drug cold start	**0.880±0.030**	**0.553±0.059**	0.266±0.044	0.392±0.049	0.926±0.013
	target cold start	0.921±0.009	0.620±0.025	0.433±0.042	0.535±0.035	0.932±0.004
Exp 3: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_3\mathcal{L}_{\text{IC50}}$	warm start	0.973±0.001	0.819±0.005	0.700±0.028	0.741±0.010	0.956±0.001
	drug cold start	0.768±0.046	0.353±0.054	0.143±0.060	0.229±0.080	0.916±0.016
	target cold start	0.921±0.008	0.617±0.025	**0.461±0.051**	0.549±0.036	0.932±0.004
Exp 4: $\lambda_1\mathcal{L}_{\text{vol}} + \lambda_2\mathcal{L}_{\text{bi}}$	warm start	0.975±0.001	0.829±0.006	0.717±0.015	0.752±0.006	0.957±0.001
	drug cold start	0.843±0.040	0.489±0.065	0.242±0.064	0.359±0.077	0.923±0.016
	target cold start	**0.923±0.006**	0.627±0.019	0.457±0.043	**0.551±0.030**	**0.933±0.004**
Exp 5: w/o Modality Dropout	warm start	0.978±0.001	0.844±0.006	0.730±0.016	0.768±0.005	0.960±0.001
	drug cold start	0.838±0.052	0.491±0.091	0.247±0.098	0.360±0.117	0.924±0.016
	target cold start	0.919±0.008	0.616±0.025	0.443±0.049	0.538±0.035	0.932±0.004

1. **Exp 1**: Training using full objective, i.e., $\mathcal{L} = \mathcal{L}_{\text{total}}$
2. **Exp 2**: Training without volume loss, i.e., $\mathcal{L} = \lambda_2\mathcal{L}_{\text{bi}} + \lambda_3\mathcal{L}_{\text{IC50}}$
3. **Exp 3**: Training without traditional pairwise contrastive loss, i.e., $\mathcal{L} = \lambda_1\mathcal{L}_{\text{vol}} + \lambda_3\mathcal{L}_{\text{IC50}}$
4. **Exp 4**: Training without IC_{50} supervision, i.e., $\mathcal{L} = \lambda_1\mathcal{L}_{\text{vol}} + \lambda_2\mathcal{L}_{\text{bi}}$
5. **Exp 5**: Without adaptive modality dropout (using all four modalities consistently)

The ablation study results are presented in Tables 3, 4, 5, and 6. Several key insights emerge from these comprehensive experiments:

Volume-based Loss Contribution: The Gramian volume-based alignment demonstrates clear benefits across multiple datasets. On the Activation dataset, comparing the full model (Exp 1) with the version excluding volume loss (Exp 2) shows consistent improvements: warm start AUPR increases from 0.625 to 0.642 and target cold start from 0.448 to 0.470. Similar patterns are observed on Yamanishi_08, where the volume-based component contributes to improved performance across different evaluation scenarios.

Bimodal Contrastive Loss Importance: The comparison between Exp 1 and Exp 3 reveals that the auxiliary SMILES-protein contrastive loss is critical for maintaining model performance. Removing this component leads to substantial performance degradation across datasets. On Yamanishi_08, the drug cold start AUPR drops from 0.440 (Exp 1) to 0.373 (Exp 3), representing a notable 15.2% performance loss. This validates our design choice to maintain traditional pairwise alignment alongside higher-order volume-based objectives, as the two approaches provide complementary benefits.

IC_{50} Supervision Benefits: The incorporation of IC50 auxiliary supervision consistently improves performance across most evaluation scenarios. Comparing Exp 1 and Exp 4 (without IC50 supervision) shows that the additional biological supervision enhances model generalization. On the Inhibition dataset, incorporating IC50 supervision improves drug cold start AUPR from 0.667 to 0.756, representing a substantial 13.4% improvement. The benefits are particularly pronounced in cold start scenarios, where the biological grounding helps the model handle novel compounds and targets.

Adaptive Modality Dropout Impact: The adaptive modality dropout mechanism shows positive effects across several datasets. Comparing the full model with Exp 5 (without adaptive dropout) reveals improvements in multiple scenarios. On the Activation dataset, the adaptive mechanism enhances warm start AUPR from 0.590 to 0.642 and target cold start from 0.444 to 0.470. The benefits appear most pronounced in challenging scenarios involving novel targets, where dynamic modality selection helps prevent overfitting to dominant but potentially less informative modalities.

Synergistic Effects: The full model (Exp 1) achieves the best overall performance across the majority of evaluation scenarios, demonstrating that the combination of all components provides synergistic benefits. Each component addresses different aspects of the learning challenge: volume-based alignment captures higher-order multimodal relationships, bimodal contrastive loss ensures stable SMILES-protein alignment, IC50 supervision provides biological grounding, and adaptive dropout prevents modality dominance.

E Pre-training Setup and Architectural Details

E.1 Pre-training Infrastructure

Our four-modal contrastive learning framework employs a two-stage training pipeline. First, we extract embeddings from domain-specific pre-trained models (MoLFormer-XL [25] for SMILES, MolT5[6] for text/HTA, ESM2 [14] for proteins). Second, we train projection networks and the GRAM4Modal loss using distributed training across multiple GPUs. The complete training procedure is detailed in Algorithm 1, which incorporates our gradient-based modality dropping strategy (Algorithm 2).

Notably, we deliberately exclude $\mathcal{L}_{\text{vol}}$ from the gradient computation for modality dropping to avoid circular dependency, where the volume loss computation would depend on gradients derived from that same computation. Instead, we use $\mathcal{L} = \lambda_2 \mathcal{L}_{\text{bi}} + \lambda_3 \mathcal{L}_{\text{IC50}}$ to assess modality importance for two key reasons: 1) *Avoiding circular dependency*: The bimodal contrastive loss and IC_{50} loss provide stable, interpretable signals about each modality's contribution without creating computational circularity; 2) *Leveraging weak supervision*: IC_{50} values, though sparsely available, offer biologically meaningful supervision that directly reflects protein-molecule interaction strength. The gradients from $\mathcal{L}_{\text{IC50}}$ thus provide valuable information about which modalities are most important for predicting drug-target activity, making them suitable signals for adaptive modality selection. Table 7 provides comprehensive training configuration details.

E.2 Model Architecture

The projection networks F_ϕ^m map pre-computed embeddings to a unified 512-dimensional space. Each projection consists of three linear layers with GELU activations, layer normalization, and dropout (rate=0.1). The IC50 classification head F_ϕ^{IC50} concatenates all four modality features $f^{\text{fused}} = [f^s; f^t; f^h; f^p]$ and predicts binding affinity classes through a two-layer MLP with dropout (rate=0.3). The pre-trained encoder specifications are detailed in Table 8. All encoders E_m are frozen during training to leverage their pre-trained representations while only fine-tuning the projection networks F_ϕ^m for computational efficiency.

E.3 Volume Computation Details

The GRAM4Modal and GRAM3Modal functions compute volumes using Gram matrix determinants. For anchor features f^a and target features $\{f^{t_1}, f^{t_2}, f^{t_3}\}$, the 4×4 Gram matrix G has entries

$G_{kj} = \langle f^k, f^j \rangle$. The volume is computed as $V = \sqrt{|\det(G)|}$, then converted to similarity via negative volume scaling: $S = -V/\tau$.

Algorithm 2 implements our gradient-informed adaptive modality selection strategy, which maintains consistency between forward $\mathcal{L}_{\text{vol}}^{\rightarrow}$ and reverse $\mathcal{L}_{\text{vol}}^{\leftarrow}$ contrastive computations by using a single drop decision per forward pass.

Algorithm 1 Four-Modal Contrastive Learning with Gradient-based Modality Dropping

Require: Pre-computed embeddings $\{x_i^s, x_i^t, x_i^h, x_i^p\}$
Require: Drop probability p_{drop}, temperature τ
Ensure: Projected features $\{f^s, f^t, f^h, f^p\}$
$f^m \leftarrow F_\phi^m(E_m(x^m))$ for $m \in \{s, t, h, p\}$
$f^m \leftarrow \|f^m\|_2 = 1$ for all modalities
$d \leftarrow$ GradientBasedDrop$(\{f^m\}, \mathcal{L}, p_{\text{drop}})$
if d.should_drop = False **then**
    $V_f \leftarrow$ GRAM4Modal$(f^p, \{f^s_{\text{all}}, f^t_{\text{all}}, f^h_{\text{all}}\})$
    $V_r \leftarrow$ GRAM4Modal$(f^p_{\text{all}}, \{f^s, f^t, f^h\})^T$
else
    $m_a \leftarrow d$.anchor_modality
    $\{m_1, m_2\} \leftarrow$ remaining_modalities $\setminus \{m_a\}$
    $V_f \leftarrow$ GRAM3Modal$(f^{m_a}, \{f^{m_1}_{\text{all}}, f^{m_2}_{\text{all}}\})$
    $V_r \leftarrow$ GRAM3Modal$(f^{m_a}_{\text{all}}, \{f^{m_1}, f^{m_2}\})^T$
end if
$S_f \leftarrow -V_f/\tau$, $S_r \leftarrow -V_r/\tau$
$\mathcal{L}_{\text{vol}} \leftarrow \frac{1}{2}[\mathcal{L}_{\text{vol}}^{\rightarrow} + \mathcal{L}_{\text{vol}}^{\leftarrow}]$
return $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{vol}} + \lambda_2 \mathcal{L}_{\text{bi}} + \lambda_3 \mathcal{L}_{\text{IC50}}$

Algorithm 2 Gradient-based Adaptive Modality Dropping

Require: Features $\{f^m\}_{m \in \{s,t,h,p\}}$, current loss $\mathcal{L}_{\tilde{t}}$, drop probability p_{drop}
Require: Gradient history length K, decay factor α, threshold $\lambda_\sigma = 1.5$
Ensure: Drop decision {should_drop, m_{drop}, anchor_modality}
if random() $> p_{\text{drop}}$ or not training **then**
    return {False, none, protein}
end if
for $m \in \{s, t, h, p\}$ **do**
    $g_{\tilde{t}}^m \leftarrow \left\| \frac{\partial \mathcal{L}_{\tilde{t}}}{\partial f_{\tilde{t}}^m} \right\|_2$
    Update gradient history for modality m
end for
for $m \in \{s, t, h, p\}$ **do**
    $\bar{g}_{\tilde{t}}^m \leftarrow \frac{\sum_{k=0}^{K-1} \alpha^k g_{\tilde{t}-k}^m}{\sum_{k=0}^{K-1} \alpha^k}$
end for
$\mu_{\tilde{t}} \leftarrow \frac{1}{4} \sum_m \bar{g}_{\tilde{t}}^m$, $\sigma_{\tilde{t}} \leftarrow \sqrt{\frac{1}{4} \sum_m (\bar{g}_{\tilde{t}}^m - \mu_{\tilde{t}})^2}$
for $m \in \{s, t, h, p\}$ **do**
    if $\bar{g}_{\tilde{t}}^m > \mu_{\tilde{t}} + \lambda_\sigma \sigma_{\tilde{t}}$ **then**
        $m_{\text{drop}}^{(\tilde{t})} \leftarrow m$; **break**
    end if
end for
if $m_{\text{drop}}^{(\tilde{t})}$ not found **then**
    $m_{\text{drop}}^{(\tilde{t})} \leftarrow \arg\min_m \bar{g}_{\tilde{t}}^m$
end if
$m_{\text{anchor}} \leftarrow$ random_choice$(\{s, t, h, p\} \setminus \{m_{\text{drop}}^{(\tilde{t})}\})$
return {True, $m_{\text{drop}}^{(\tilde{t})}$, m_{anchor}}

Table 7: Training Configuration Parameters

Parameter	Configuration
Hardware	Multi-GPU NVIDIA (CUDA)
Training framework	PyTorch DDP, NCCL
Batch size	1280 per GPU
Learning rate	1×10^{-4} (Adam)
Epochs	40
Temperature τ	0.07
Drop probability p_{drop}	0.8
Gradient history length K	5
Decay factor α	0.9
Threshold multiplier λ_σ	1.5
Loss weights $\lambda_1, \lambda_2, \lambda_3$	1.0, 1.0, 1.0
Label smoothing	0.1

Table 8: Pre-trained Encoder Specifications

Modality	Model E_m	Output Dim
SMILES (x^s)	MoLFormer-XL-both-10pct	768
Text (x^t)	MolT5-base	768
HTA (x^h)	MolT5-base (shared)	768
Protein (x^p)	ESM2_t33_650M_UR50D	1280

E.4 Downstream Task Architecture

For drug-target interaction (DTI) prediction evaluation, we employ a lightweight classification architecture that leverages the pre-trained embeddings from our four-modal framework. The downstream architecture is detailed in Algorithm 3 and uses only the drug (SMILES) and protein modalities relevant for binding prediction.

Algorithm 3 Drug-Target Interaction Prediction

Require: Pre-trained embeddings $f^s, f^p \in \mathbb{R}^{512}$
Require: Drug-protein pair (x_i^s, x_j^p), binding label $y_{ij} \in \{0, 1\}$
Ensure: Binding prediction $\hat{y}_{ij}$
$f_i^s \leftarrow \text{FROZEN}(F_\phi^s(E_s(x_i^s)))$ {Use pre-trained SMILES embedding}
$f_j^p \leftarrow \text{FROZEN}(F_\phi^p(E_p(x_j^p)))$ {Use pre-trained protein embedding}
$f^{\text{concat}} \leftarrow [f_i^s; f_j^p] \in \mathbb{R}^{1024}$ {Concatenate embeddings}
$h_1 \leftarrow \text{ReLU}(\text{Linear}_{1024 \to 512}(f^{\text{concat}}))$
$h_1 \leftarrow \text{Dropout}_{0.3}(h_1)$
$h_2 \leftarrow \text{ReLU}(\text{Linear}_{512 \to 256}(h_1))$
$h_2 \leftarrow \text{Dropout}_{0.3}(h_2)$
$\text{logits} \leftarrow \text{Linear}_{256 \to 2}(h_2)$
$\hat{y}_{ij} \leftarrow \arg\max(\text{softmax}(\text{logits}))$
return $\hat{y}_{ij}$

E.5 Evaluation Metrics

We employ five standard binary classification metrics to comprehensively assess DTI prediction performance. Given the confusion matrix with true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the metrics are defined as follows:

Area Under ROC Curve (AUROC) AUROC measures the model's ability to discriminate between positive and negative classes across all classification thresholds:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{3}$$

where $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$.

Area Under Precision-Recall Curve (AUPRC) AUPRC is particularly informative for imbalanced datasets and measures performance across different precision-recall trade-offs:

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall}^{-1}(t))\, dt \tag{4}$$

where $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$ and $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$.

Sensitivity (Recall) Sensitivity measures the proportion of actual positive cases correctly identified:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1-Score F1-score provides the harmonic mean of precision and recall, balancing both measures:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

Accuracy Accuracy measures the overall proportion of correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$