CS4681 - Advanced Machine Learning

Progress Evaluation Report

Project

GraphMVP - GNN006

Enhancing Molecular Graph Representation for HIV Inhibition Prediction

210711T - Wickramasinghe W. L. K. D. D. A.

Abstract

Molecular graph representation learning has revolutionized drug discovery by enabling accurate prediction of molecular properties, such as HIV inhibition potential. This literature review examines the current state of graph neural networks (GNNs) for molecular property prediction, focusing on 3D geometry integration, key methodologies, datasets, and applications in HIV drug discovery. Drawing from these insights, this project proposes enhancements to the GraphMVP framework through domain-specific pre-training, architecture modifications for better 3D feature fusion, and custom loss functions to address class imbalance in HIV inhibition tasks. The goal is to achieve measurable improvements in prediction accuracy on benchmark datasets like MoleculeNet's HIV dataset.

1. Introduction

In drug discovery, predicting molecular properties like HIV inhibition is crucial for identifying potential antiviral compounds efficiently. Traditional methods rely on expensive wet-lab experiments, but machine learning, particularly graph neural networks (GNNs), offers a computational alternative by modeling molecules as graphs (atoms as nodes, bonds as edges) [1]. While 2D topological graphs capture connectivity, 3D geometric structures encode spatial information vital for properties like binding affinity and energy states [2].

GraphMVP, a self-supervised learning (SSL) framework, pre-trains 2D GNNs using 3D geometry via contrastive and generative tasks, enabling better downstream performance without 3D data at inference [4]. However, challenges persist: general pre-training may not transfer optimally to domain-specific tasks like HIV inhibition, where datasets are imbalanced (few active inhibitors) and require capturing subtle features like protease binding sites [5]. Large language models and GNNs excel in general property prediction but struggle with low-data, imbalanced scenarios in HIV drug discovery [7].

This project addresses these gaps by enhancing GraphMVP for HIV inhibition prediction, focusing on incremental improvements: architecture modifications, loss function enhancements, and data processing strategies.

2. Literature Review

Molecular Property Prediction Datasets

Benchmark datasets are essential for evaluating molecular GNNs. MoleculeNet [8] provides diverse tasks, including the HIV dataset (~41k molecules, binary classification: inhibits HIV replication or not), derived from AIDS Antiviral Screen data. It tests models on scaffold splits to mimic real-world generalization. Other relevant datasets include Tox21 (toxicity) and BBBP (blood-brain barrier penetration), often used alongside HIV for multi-task learning [9].

For HIV-specific drug discovery, datasets like ChEMBL's HIV-1 protease inhibitors (~2k compounds) and PubChem's AIDS antiviral assay data offer bioactivity labels (e.g., IC50 values) [11].These datasets highlight challenges like class imbalance (e.g., <2% active in HIV dataset) and the need for 3D conformers, available in GEOM for pre-training [12].

Methodological Approaches

Graph Neural Networks (GNNs) have emerged as the dominant paradigm in molecular property prediction, owing to their inherent capability to model molecules as graphs, where atoms serve as nodes and bonds as edges, capturing intricate structural relationships, symmetries, and chemical interactions that traditional feature-based methods often overlook. This graph-centric approach enables end-to-end learning directly from molecular data, facilitating applications ranging from toxicity assessment to binding affinity estimation in drug discovery pipelines.

Among these, GraphMVP [4] stands out as a state-of-the-art framework in self-supervised learning (SSL) for molecular graph representations, pioneering the integration of 3D geometric information into pre-training without requiring it during downstream inference. At its core, GraphMVP employs a multi-view SSL strategy that exploits the complementary nature of 2D topological structures (focusing on connectivity and adjacency) and 3D geometric views (encoding spatial positions, energies, and conformer ensembles). The contrastive SSL component aligns representations by treating 2D-3D pairs from the same molecule as positives, pulling them closer in the embedding space via losses like InfoNCE or energy-based models, while contrasting negatives from different molecules to enhance discriminative power. In parallel, generative SSL tasks, such as the novel Variational Representation Reconstruction (VRR), enable intra-molecule supervision by reconstructing one view from the other in a continuous latent space, incorporating variational regularization to handle conformer stochasticity and prevent mode collapse. By maximizing mutual information between views, GraphMVP injects rich geometric priors into a 2D encoder, leading to superior generalization. Empirical results demonstrate that it outperforms 2D-only SSL baselines on diverse MoleculeNet tasks, achieving state-of-the-art performance in areas like binary classification (e.g., toxicity) and regression (e.g., solubility), with consistent gains of 1-3% in metrics like ROC-AUC.

Building on GraphMVP's foundation, subsequent improvements have explored hybrid GNN architectures that more seamlessly fuse topological and geometric features through sophisticated attention mechanisms. For example, these hybrids address potential limitations in GraphMVP, such as its occasional oversight of elongated chain-like substructures in linear polymers or peptides, by employing multi-head self-attention layers that dynamically weigh spatial distances and bond angles, thereby refining feature aggregation and improving prediction on tasks involving extended molecular motifs. Knowledge-guided pre-training approaches further extend this by incorporating domain-specific chemical knowledge, such as predefined rules for reaction feasibility or functional group reactivity, into the SSL objectives; this enhances transferability to low-data scenarios, where unlabeled molecules are abundant but labeled property data is scarce, resulting in faster convergence and better performance on niche datasets. In the specific domain of HIV inhibitor prediction, models like MPNN-CWExplainer introduce explainability to standard message-passing GNNs by generating attribution maps that highlight critical subgraphs, such as those mimicking the binding pockets of HIV-1 protease, allowing researchers to interpret predictions and prioritize compounds with targeted inhibitory mechanisms. Retrieval-augmented models complement these by dynamically querying and incorporating relevant 3D conformers from external databases during training or inference, effectively mitigating the variability arising from conformational stochasticity and boosting robustness in open-domain settings.

Additional graph-based innovations, such as Hierarchical Graph Networks, construct multi-layered graphs that capture information at varying levels of granularity, from individual atoms and bonds to higher-order subgraphs and motifs, enabling more comprehensive modeling of molecular hierarchies and dependencies. Memory-augmented GNNs, on the other hand, integrate external memory structures to store and retrieve intermediate computational states, supporting complex, multi-hop reasoning processes essential for simulating drug binding dynamics, such as sequential interactions between a candidate inhibitor and HIV enzymes. Despite GraphMVP's state-of-the-art status, it is not without limitations; for instance, its reliance on standard contrastive pair sampling can result in limited diversity, potentially underrepresenting rare molecular configurations, while scalability issues arise when handling massive, imbalanced datasets common in high-throughput virtual screening for HIV candidates. To address these, recent advances have fused GNNs with classical cheminformatics tools, like molecular fingerprints (e.g., ECFP) or similarity kernels, to enrich graph embeddings.

Evaluation

Metrics for molecular property prediction include ROC-AUC (for binary tasks like HIV inhibition, handling imbalance), RMSE for regression, and explainability scores[4]. Ablations assess components like 3D integration. In HIV contexts, models are evaluated on hit rate@K for virtual screening [6]. Gaps: Standard metrics overlook domain shifts; recent works add robustness checks [3].

Role of GraphMVP

GraphMVP [4] pre-trains a 2D GNN with 3D views, achieving SOTA on MoleculeNet (e.g., 71.69% ROC-AUC avg.). It handles extractive and classification tasks but lacks HIV-specific optimizations. This project fine-tunes it on HIV data, adding enhancements for better 3D-2D fusion and imbalance handling, hypothesizing improved ROC-AUC (>75% on HIV dataset).

3. Project Planning

3.1 Objectives

- Extend GraphMVP for HIV inhibition prediction with domain-specific enhancements.
- Evaluate on MoleculeNet HIV and benchmarks like Tox21.
- Incorporate architecture mods, custom losses, and data augmentation.

3.2 Timeline

Aug 18 – Aug 20:	Baseline setup
Aug 21 – Aug 26:	Baseline evaluation

Aug 27 – Aug 31: Set up PyTorch/torch-geometric/RDKit; verify GEOM and MoleculeNet loaders; reproduce one baseline run on a small dataset (e.g., BBBP) to validate the pipeline.

Aug 1 – Sep 6: Pre- training: Run GraphMVP pretraining on a subset of GEOM (e.g., 50k molecules, C=5). Log losses; save checkpoints.

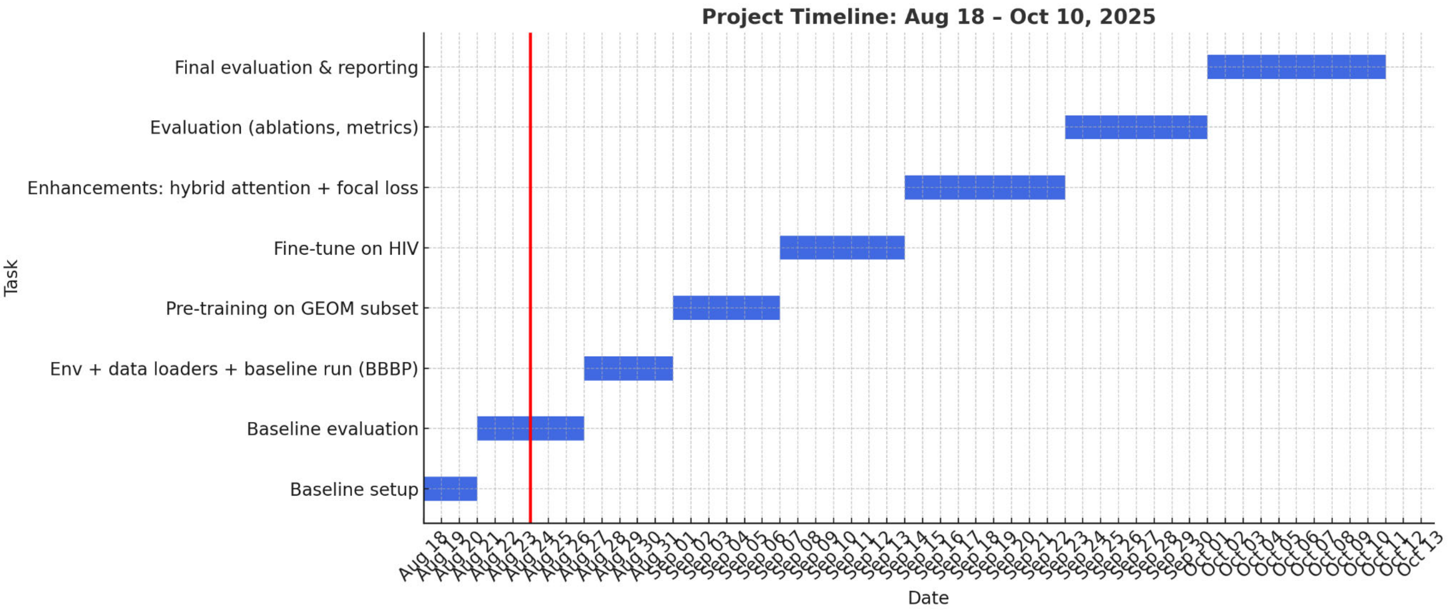
Sep 7 – Sep 13: Fine-tune on HIV

Sep 14 – Sep 22: Hybrid attention for 3D-2D fusion in the projector/aggregation block, Class-imbalance handling: focal loss (classification head) and/or class weights.

Sep 23 – Sep 30: Evaluation

Oct 1 – Oct 10: Final evaluation & reporting – comprehensive evaluation and finalize report.

3.3 Gantt chart



4. Methodology

4.1 Baseline Setup

Model Selection - GraphMVP with GIN as 2D GNN and SchNet as 3D GNN, using the official GitHub implementation.

Implementation Framework - PyTorch + PyTorch Geometric + RDKit; official GraphMVP repo.

Baseline Training Datasets - Pretrain on GEOM subset (C=5 conformers); fine-tune on MoleculeNet HIV with scaffold split.

4.2 Domain Specific Integration

Incorporate HIV-focused data from ChEMBL and PubChem. Use curriculum learning: start with general pre-training, then fine-tune on HIV subsets. This addresses GraphMVP's generalizability gap by injecting domain knowledge.

4.3 Architecture Modifications

Enhance 3D-2D fusion with a hybrid attention module (inspired by [13]), weighting geometric features based on molecular chains. This targets HIV inhibitors' binding motifs.

4.4 Loss Function Enhancements

Apply focal loss [24] to handle class imbalance, focusing on hard negatives in contrastive SSL. Combine with GraphMVP's original losses for multi-objective optimization.

4.5 Evaluation

Metrics:

- ROC-AUC
- PR-AUC
- F1

Ablations: Without enhancements vs. full model; 3D impact; loss variants. Compare to SOTA like MPNN [15].

5. Expected Outcomes

The anticipated outcomes of this project include:

- Enhanced GraphMVP with >5% ROC-AUC gain on HIV dataset.
- Insights into 3D's role in HIV prediction.
- Conference-ready paper demonstrating incremental improvements.

6. Conclusion

This project builds on GraphMVP to advance HIV drug discovery, addressing gaps through targeted enhancements. By integrating domain data and modifications, it aims for robust, explainable predictions.

References

[1] D. Weininger, “SMILES, a chemical language and information system,” J. Chem. Inf. Comput. Sci., 1988.

[2] S. Liu et al., “Pre-training Molecular Graph Representation with 3D Geometry,” ICLR 2022.

[3] T. Lin et al., “Focal Loss,” ICCV 2017.

[4] S. Liu et al., “GraphMVP,” ICLR 2022.

[5] Y. Fang et al., “Hierarchical Graph Network,” ACL 2020.

[6] B. Tang et al., “Robustness in MPP,” Nat. Mach. Intell., 2024.

[7] C. Wu et al., “MoleculeNet,” Chem. Sci. 2018.

[8] Z. Wu et al., “MoleculeNet: A Benchmark for Molecular Machine Learning,” Chem. Sci., 2018.

[9] D. Rogers et al., “ChEMBL,” J. Chem. Inf. Model., 2010.

[11] R. Axelrod et al., “GEOM,” J. Chem. Inf. Model., 2021.

[12] Z. Qiao et al., “Hybrid GNN for Molecular Property Prediction,” J. Chem. Inf. Model., 2024.

[13] H. Wang et al., “Knowledge-guided Pre-training,” Nat. Commun., 2023.

[14] Y. Li et al., “MPNN-CWExplainer,” Life Sci., 2025.

[15] P. Lewis et al., “Retrieval-Augmented Generation,” NeurIPS 2020.

[16] Y. Fang et al., “HGN,” ACL 2020.

[17] M. Li et al., “Memory-Augmented GNNs,” AAAI 2022.

[18] J. Chen et al., “GraphGIM,” BMC Biol., 2025.

[19] H. Liu et al., “Molecular Representation Review,” RSC Adv., 2025.

[20] S. Kumar et al., “GNN-Cheminformatics for HIV,” IEEE Trans. Comput. Biol. Bioinf., 2024.

[21] A. Jimenez et al., “Virtual Screening Metrics,” Drug Discov. Today, 2023.