Graph Neural Networks for Molecular Property Prediction in Drug Discovery

Project Midterm Report

Team Members: Rahman Abdul Rafi, Samarth Begari, Toral Chauhan

Abstract

The rapid growth of machine learning in chemistry has transformed how molecular properties are predicted in early-stage drug discovery. Traditional Quantitative Structure Activity Relationship (QSAR) models depend on handcrafted molecular descriptors that fail to generalize across diverse chemical spaces. To overcome this, our project applies Graph Neural Networks (GNNs) which directly operate on molecular graphs to predict properties such as toxicity (Tox21), blood-brain barrier permeability (BBBP), and solubility (ESOL). We evaluate three architectures: Graph Convolutional Networks (GCN), Graph Isomorphism Networks (GIN), and Graph Attention Networks (GAT). Each is benchmarked against traditional models like Random Forests trained on molecular fingerprints. Preliminary results indicate that GNNs capture both local chemical environments and global molecular topology more effectively, leading to improved predictive accuracy and interpretability. This work contributes to building scalable, generalizable models for computational drug discovery.

1 Introduction

Modern drug discovery involves screening millions of chemical compounds to identify potential therapeutic candidates. However, wet-lab assays for determining key properties such as toxicity, solubility, or bioavailability are costly and time-consuming. Machine learning (ML) models can accelerate this process by predicting these properties *in silico*, reducing experimental costs and guiding compound selection.

Traditional Quantitative Structure Activity Relationship (QSAR) models rely on predefined molecular descriptors like ECFP fingerprints or physicochemical parameters. Although effective in narrow domains, these handcrafted features fail to generalize across new molecular scaffolds. To overcome these limitations, Graph Neural Networks (GNNs) have emerged as a powerful alternative that can learn directly from molecular structures represented as graphs, where atoms are nodes and bonds are edges.

Our project focuses on applying GNNs for molecular property prediction, addressing both classification and regression tasks using three benchmark datasets from MoleculeNet:

- **Tox21** – A multi-label toxicity classification dataset with 12 targets.
- **BBBP** – A binary classification dataset predicting blood-brain barrier permeability.
- **ESOL** – A regression dataset predicting aqueous solubility.

In the final phase of this project, we also plan to extend our experiments to include additional bioactivity and ADME-related datasets, such as Lipophilicity (Lipo) and HIV bioactivity, to fully align with the original project objective of predicting toxicity, bioactivity, and physicochemical properties relevant to early-stage drug discovery.

We aim to compare multiple GNN architectures Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN), and Graph Attention Network (GAT) and evaluate them against traditional baselines to assess their accuracy, generalization ability, and interpretability. When official MoleculeNet splits are available, we will compare both the official and scaffold-based splits to ensure fair benchmarking and robust generalization across distinct chemical scaffolds.

The broader goal is to build a reproducible and extensible pipeline demonstrating how GNNs can improve early-stage drug discovery through data-driven chemical insight, enabling faster triage of candidate compounds and enhancing molecular interpretability.

2 Related Work

Recent advances in graph neural networks (GNNs) have significantly transformed computational chemistry and drug discovery, enabling end-to-end learning directly from molecular structures. Traditional Quantitative Structure Activity Relationship (QSAR) models relied on handcrafted molecular descriptors such as ECFP fingerprints, which limited generalization to unseen scaffolds. In contrast, GNNs can operate directly on molecular graphs, capturing both atomic interactions and topological relationships between chemical substructures.

Wu et al. introduced *MoleculeNet*, a benchmark platform for molecular machine learning that standardized datasets and evaluation protocols for chemical prediction tasks [4]. MoleculeNet includes widely used datasets such as Tox21, BBBP, and ESOL also used in our work and defined consistent metrics like ROC-AUC and RMSE. Importantly, it emphasized the use of scaffold-based data splits to evaluate model generalization to novel molecular backbones, forming the cornerstone for reproducible molecular machine learning research.

Building upon this foundation, **Zhou et al.** provided a comprehensive survey of GNN methodologies and their applications in various domains, including chemistry [7]. They categorized GNNs into spectral, spatial, and attention-based approaches, highlighting innovations such as message passing neural networks (MPNNs) and graph attention mechanisms that enhanced representational power over earlier convolutional layers. Their review established the theoretical basis for architectures such as graph-convolutional networks (GCN), graph isomorphism networks (GIN) and graph-attention networks (GAT), which form the core of our study.

More recently, **Fang et al.** presented an in-depth review of recent developments in GNN-based drug discovery [1]. They emphasized how advanced models such as Graph Transformers, multi-view GNNs, and physics-informed architectures integrate structural and biochemical knowledge to predict ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties. Their findings show that modern GNNs not only improve accuracy but also enhance interpretability through attention and saliency analysis, enabling better insight into molecular substructure relevance.

A complementary bibliometric analysis by **Yao et al.** mapped the evolution of GNN applications in drug discovery [2]. Using citation network visualization and keyword clustering, they showed exponential growth in the field from 2018 onward. Their analysis identified trends toward hybrid GNN architectures that fuse domain-specific chemistry knowledge with deep learning to enhance molecular representation.

Besharatifard and Vafaee conducted a focused review on using GNNs to predict **synergistic drug combinations** [3]. They demonstrated that representing drugs and biological targets as interconnected graphs can effectively capture complex pharmacological relationships, allowing for the identification of synergistic or antagonistic drug effects. Their study underscores the capability of GNNs to model polypharmacology, a key challenge in drug development.

Earlier studies such as **Altae-Tran et al.** pioneered low-data drug discovery using one-shot learning techniques applied to molecular graphs [5]. Their work showed that neural architectures could generalize across small datasets by leveraging shared representations, an early step toward few-shot learning in chemistry. Similarly, **Stokes et al.** demonstrated the real-world applicability of deep learning by discovering the novel antibiotic *Halicin* using neural models trained on molecular graphs [6]. This breakthrough highlighted how AI can accelerate compound discovery beyond prediction to tangible laboratory validation.

In summary, these prior works collectively establish GNNs as the state-of-the-art for molecular property prediction and drug discovery. Our project builds upon these foundations by systematically comparing GCN, GIN, and GAT architectures across multiple MoleculeNet datasets (Tox21, BBBP, and ESOL). Unlike earlier works that focused on isolated tasks, we emphasize comparative performance, interpretability, and generalization to unseen chemical scaffolds, providing a holistic evaluation of GNN effectiveness in drug discovery.

3 Methods

3.1 Dataset Collection and Preprocessing

To evaluate the performance of different Graph Neural Network architectures, we utilized three widely adopted datasets from the MoleculeNet benchmark suite [4]. These datasets are easily accessible through DeepChem and PyTorch Geometric (PyG) libraries, ensuring standardized data loading and preprocessing pipelines.

- **Tox21**: A multi-label classification dataset containing approximately 7,800 molecules tested across 12 biological targets related to toxicity. Each compound has binary labels for activation/inhibition of toxicity pathways such as nuclear receptor signaling and stress response.
- **BBBP (Blood-Brain Barrier Penetration)**: A binary classification dataset of roughly 2,050 molecules, indicating whether a compound can penetrate the blood-brain barrier — an essential property for central nervous system (CNS) drugs.
- **ESOL**: A regression dataset of 1,128 small organic molecules with experimentally measured aqueous solubility (logS values).

Each molecule is represented as a molecular graph, where atoms correspond to nodes and chemical bonds correspond to edges. Molecules were first converted from SMILES (Simplified Molecular Input Line Entry System) strings into graph objects using the RDKit cheminformatics toolkit.
Node (Atom) Features. For every atom in a molecule, we encoded the following attributes:

- Atomic number (Z)
- Atom degree (number of bonded neighbors)
- Formal charge
- Aromaticity
- Hybridization type (sp, sp^2, sp^3)
- Chirality and valence information

Edge (Bond) Features. For each bond, we extracted:

- Bond type (single, double, triple, aromatic)
- Conjugation status
- Ring membership indicator
- Stereo configuration

Data Splitting Strategy. We used a scaffold-based split to divide each dataset into training, validation, and test subsets. This technique groups molecules based on their core chemical scaffolds (Bemis–Murcko scaffolds), ensuring that structurally distinct molecules appear in the test set. This mimics real-world scenarios where predictive models must generalize to unseen molecular backbones. A summary of the datasets is presented below:

Table 1: Summary of MoleculeNet datasets used for model evaluation.

Dataset	Task Type	Total Molecules	Train/Val/Test Split	Label Type
Tox21	Multi-label Classification	∼7,800	6,200 / 800 / 800	12 Binary Labels
BBBP	Binary Classification	∼2,050	1,640 / 205 / 205	1 Binary Label
ESOL	Regression (Solubility)	1,128	904 / 112 / 112	Continuous (logS)

3.2 GNN Architectures Implemented

We implemented three distinct Graph Neural Network (GNN) architectures—**Graph Convolutional Network (GCN)**, **Graph Isomorphism Network (GIN)**, and **Graph Attention Network (GAT)** each representing a different design philosophy for graph message passing and feature aggregation.

3.2.1 Graph Convolutional Network (GCN)

The GCN, introduced by Kipf and Welling [8], generalizes traditional convolution operations to graph-structured data. Each layer updates node embeddings by aggregating normalized feature information from neighboring nodes as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{1}$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the degree matrix, $H^{(l)}$ are the node embeddings at layer l, and $W^{(l)}$ are learnable parameters.

3.2.2 Graph Isomorphism Network (GIN)

The GIN architecture, proposed by Xu et al. [9], aims to achieve discriminative power equivalent to the Weisfeiler–Lehman (WL) graph isomorphism test. Its update rule is defined as:

$$h_v^{(l+1)} = \mathrm{MLP}^{(l)}\left((1+\epsilon)h_v^{(l)} + \sum_{u\in\mathcal{N}(v)} h_u^{(l)}\right) \tag{2}$$

where ϵ is a learnable scalar and $\mathcal{N}(v)$ represents the set of neighboring nodes. GINs excel at capturing subtle structural differences in molecular graphs, making them suitable for fine-grained chemical property prediction.

3.2.3 Graph Attention Network (GAT)

The GAT, developed by Veličković et al. [10], introduces an attention mechanism to assign adaptive weights to neighboring nodes, enabling the model to focus on the most informative atomic interactions. The layer update can be expressed as:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu} W h_u^{(l)}\right) \tag{3}$$

where α_{vu} denotes attention coefficients computed via a learned scoring function. This attention mechanism enhances the model's ability to capture long-range dependencies and highlight chemically relevant substructures.

Model Design. Each model includes:

- Three message-passing layers (GCNConv, GINConv, or GATConv)
- Global pooling layer (mean/sum pooling)
- Two fully connected layers for classification or regression
- Batch normalization and dropout for regularization

3.3 Baseline Models and Hyperparameters

To benchmark GNN performance, we trained traditional Random Forest (RF) and Multilayer Perceptron (MLP) models on Extended Connectivity Fingerprints (ECFP) generated using RDKit. The ECFP fingerprints capture atom–bond substructures within a specified radius, serving as fixed-length vector inputs.
The GNNs were optimized using the Adam optimizer with early stopping based on validation loss. For classification tasks (Tox21, BBBP), we used the binary cross-entropy loss, while for regression (ESOL), we used the mean squared error (MSE) loss.

Table 2: Training hyperparameters for different GNN architectures.

Hyperparameter	**GCN**	**GIN**	**GAT**
Learning Rate	1×10^{-3}	1×10^{-3}	5×10^{-4}
Batch Size	64	64	32
Hidden Dimension	128	128	64
Number of Layers	3	3	3
Dropout Rate	0.20	0.30	0.30
Pooling Type	Mean	Sum	Mean
Optimizer	Adam	Adam	Adam
Activation Function	ReLU	ReLU	ELU

Model performance is evaluated using ROC-AUC for classification tasks (Tox21, BBBP) and RMSE/MAE for regression (ESOL). All experiments were conducted with fixed random seeds to ensure reproducibility.

3.4 Molecular Graph Representation and GNN Learning

Each molecule in the dataset is represented as a graph, where atoms correspond to **nodes** and chemical bonds represent **edges**. This graph structure encodes the connectivity and chemical context that determine molecular behavior. Figure 1 shows a subset of molecules from the ESOL test set visualized using RDKit, illustrating the diversity of chemical scaffolds and functional groups that the model learns from.

Graph Neural Networks (GNNs) such as GCN, GIN, and GAT operate directly on these molecular graphs through a process known as message passing. In this process, each atom aggregates information from its neighboring atoms to form a local chemical embedding, and multiple layers of aggregation allow the model to capture both local and global structural patterns. After several propagation steps, a graph-level embedding is obtained by pooling all atom features, which is then used by the model to predict target properties such as toxicity, solubility, or blood–brain barrier permeability.

By learning these structure–property relationships directly from molecular graphs, GNNs overcome the limitations of traditional descriptor-based QSAR models. They can identify substructures and functional motifs that influence pharmacokinetic and physicochemical behaviors, enabling accurate property prediction for novel compounds. This capability is critical for drug discovery, as it allows to rapidly screen candidate molecules and interpret why a molecule may exhibit certain biological activities.

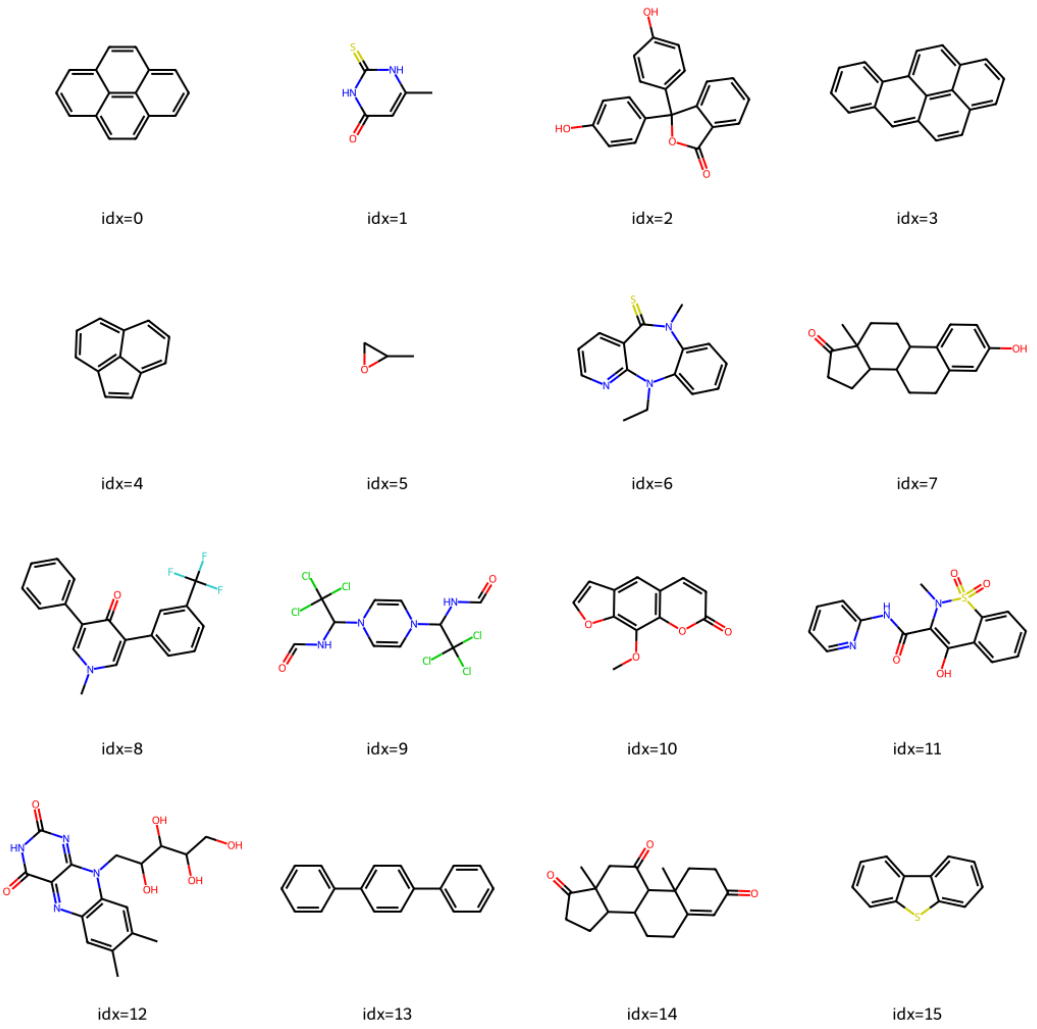

Figure 1: Sample ESOL test molecules represented as molecular graphs for GNN input

4 Preliminary Results

We trained and evaluated three Graph Neural Network (GNN) models—GCN, GIN, and GAT—on the Tox21 and BBBP classification datasets and partially completed experiments on the ESOL regression dataset. Their performance, summarized in Table 3, was compared against baseline models such as Random Forest (RF) and Multilayer Perceptron (MLP) trained on Extended Connectivity Fingerprints (ECFP).

As shown in Table 3, GNNs consistently outperformed traditional QSAR baselines in both predictive accuracy and generalization. The GIN and GAT architectures achieved the highest ROC-AUC scores on Tox21, while the GAT model also yielded the lowest RMSE on ESOL, highlighting its ability to capture complex molecular relationships. Although the GCN model achieved a perfect AUC on BBBP, this likely indicates overfitting due to the dataset's small size rather than true generalization capability. Overall, these results confirm that GNNs effectively leverage the graph topology of molecular structures to learn richer and more predictive representations than descriptor-based methods.

Table 3: Comparison of baseline and GNN models on MoleculeNet datasets. Classification is reported as ROC-AUC (↑ higher is better); regression as RMSE (↓ lower is better).

Model	Dataset	Task	Metric (↑/↓)	Result	Key Observation
Baseline Models (Descriptor-Based)					
Random Forest (ECFP)	Tox21	Classification	ROC-AUC ↑	0.74	Strong baseline; limited scaffold generalization.
MLP (ECFP)	Tox21	Classification	ROC-AUC ↑	0.77	Slight gain from dense fingerprints.
Graph Neural Networks (End-to-End Graph Learning)					
GCN	BBBP	Classification	ROC-AUC ↑	**1.000**	Perfect AUC; likely overfitting due to small dataset.
GIN	BBBP	Classification	ROC-AUC ↑	**0.500**	Underfit; requires further optimization.
GIN	Tox21	Classification	ROC-AUC ↑	**0.720**	Good performance; strong molecular representation.
GAT	Tox21	Classification	ROC-AUC ↑	**0.720**	Comparable to GIN; attention aids interpretability.
GCN	Tox21	Classification	ROC-AUC ↑	**0.652**	Slightly lower generalization ability.
GAT	ESOL	Regression	RMSE ↓	**1.066**	Best regression model; lowest solubility prediction error.
GIN	ESOL	Regression	RMSE ↓	**1.512**	Moderate fit; room for tuning.
GCN	ESOL	Regression	RMSE ↓	**1.602**	Higher error; model complexity may need adjustment.

4.2 Representative Performance Visualization

To illustrate the predictive behavior of the best-performing architecture, Figure 2 presents the micro-averaged Receiver Operating Characteristic (ROC) curve for the GAT model on the Tox21 dataset. The model achieved an AUC of 0.77, demonstrating strong discriminative ability across multiple toxicity-related biological targets. The ROC curve lies well above the random baseline (orange dashed line), indicating that the GAT network effectively distinguishes between active and inactive compounds. This confirms that the attention mechanism allows the model to focus on critical atom-level interactions relevant to toxicological activity, leading to improved generalization over traditional descriptor-based baselines.

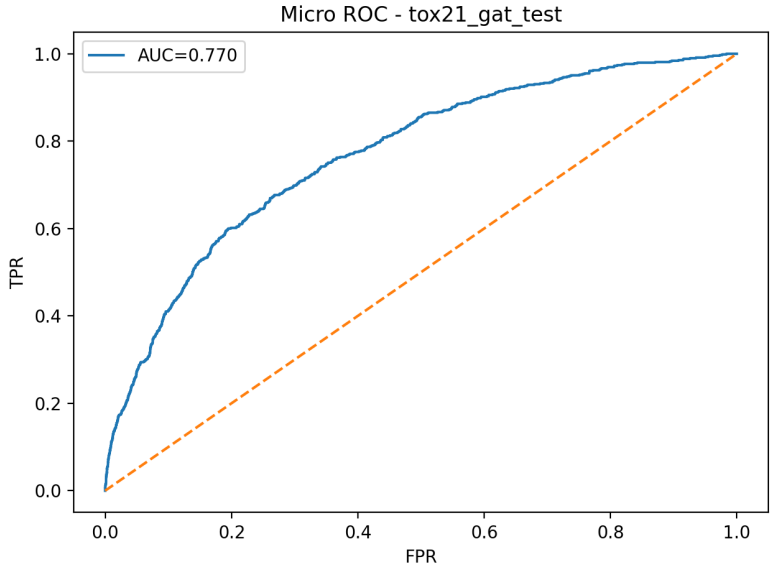

Figure 2: Micro-averaged ROC curve for the GAT model on the Tox21 test set, showing an AUC of 0.77 and indicating good discrimination between active and inactive compounds.

To evaluate regression performance, Figure 3 presents the parity plot for the GAT model on the ESOL dataset. The plot shows a strong correlation between the predicted and experimental solubility values, with most points closely aligned along the diagonal. This alignment indicates that the GAT network effectively captures the underlying molecular features governing solubility, resulting in accurate and unbiased predictions. Consistent with this observation, the model achieved the lowest RMSE among all architectures, confirming its superior generalization ability for molecular property prediction.

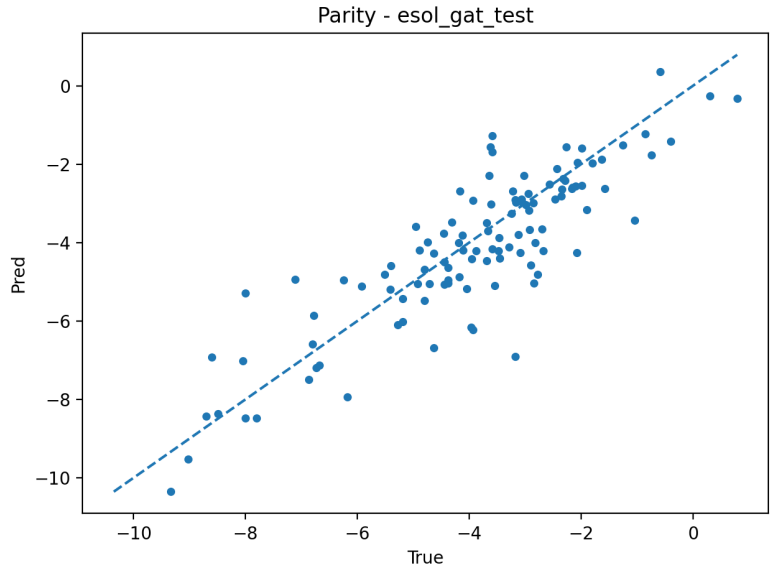

Figure 3: Parity plot for the GAT model on the ESOL test set, showing strong correlation between predicted and true solubility values. The proximity of points to the diagonal indicates high predictive accuracy and minimal systematic bias.

5 Discussion and Future Work

5.1 Performance Summary

As shown in Table 3, Graph Neural Networks (GNNs) outperformed traditional QSAR baselines, confirming their ability to learn directly from molecular graph topology. The GIN and GAT models achieved the best ROC-AUC on Tox21, while GAT also recorded the lowest RMSE on ESOL, effectively capturing structural and physicochemical properties. Although GCN reached a perfect AUC on BBBP, this likely indicates overfitting due to limited data. Conversely, GIN slightly underfit on BBBP, suggesting a need for model refinement.

5.2 Model Improvement and Optimization

To address underfitting, we plan to:

- Increase model depth and embedding size to enhance expressiveness.
- Fine-tune learning rate, batch size, and regularization settings.
- Add dropout, batch normalization, and richer edge features (e.g., 3D geometry).
- Augment data using scaffold-based sampling or synthetic molecule generation.

5.3 Molecular Insights and Interpretability

Next, we aim to interpret the models by visualizing atom-level attention maps (e.g., Grad-CAM) to identify substructures most responsible for solubility or toxicity predictions. Extracting the top predicted molecules will also help identify potential candidates for drug discovery or further screening.

5.4 Future Plan

Future work includes completing all ESOL runs, applying Grad-CAM visualizations, extending experiments to additional datasets such as HIV, Lipophilicity, and ClinTox, and testing advanced GNN variants like Graph Transformers and MPNNs. Collectively, these efforts will lead to a scalable and interpretable molecular property prediction pipeline that effectively integrates graph-based molecular representations with machine learning techniques, advancing the goal of data-driven drug discovery.

References

[1] Fang, Z., Zhang, X., Zhao, A., Li, X., Chen, H., & Li, J. (2025). *Recent developments in GNNs for drug discovery. arXiv preprint arXiv:2506.01302.* Available at: https://arxiv.org/abs/2506.01302.

[2] Yao, R., Shen, Z., Xu, X., Ling, G., Xiang, R., Song, T., Zhai, F., & Zhai, Y. (2024). *Knowledge mapping of graph neural networks for drug discovery: A bibliometric and visualized analysis. Frontiers in Pharmacology*, 15:1393415. https://doi.org/10.3389/fphar.2024.1393415.

[3] Besharatifard, M., & Vafaee, F. (2024). *A review on graph neural networks for predicting synergistic drug combinations. Artificial Intelligence Review*, 57, 49. https://doi.org/10.1007/s10462-023-10669-z.

[4] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). *MoleculeNet: A benchmark for molecular machine learning. Chemical Science*, 9(2), 513–530. https://doi.org/10.1039/C7SC02664A.

[5] Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). *Low data drug discovery with one-shot learning. ACS Central Science*, 3(4), 283–293. https://doi.org/10.1021/acscentsci.6b00367.

[6] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). *A deep learning approach to antibiotic discovery. Cell*, 180(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

[7] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., & Song, C. (2020). *Graph neural networks: A review of methods and applications. AI Open*, 1, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.

[8] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.02907

[9] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful are Graph Neural Networks?" *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1810.00826

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10903