Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

University of Mumbai

A.Y. 2025-26

Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening

Submitted in partial fulfilment of the requirements of the degree of
Bachelor of Technology in Computer Science and Engineering (Data Science)

By

60009220218 Harshal Loya
60009220051 Jash Chauhan
60009220210 Het Gala

Under the Guidance of

Dr. Poonam Jadhav

Certificate

This is to certify that the project entitled, **"Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening"** is a bonafide work of **"Harshal Loya" (60009220218)**, **"Jash Chauhan" (60009220051)**, and **"Het Gala" (60009220210)** submitted in partial fulfillment of the requirement for the award of the Bachelor of Technology in Computer Science and Engineering(Data Science).

Name of the Guide
Dr. Poonam Jadhav

Dr. Kriti Srivastava
Head of the Department

Dr. Hari Vasudevan
Principal

Place:

Date:

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all the principles of academic honesty and integrity and have not misrepresented, fabricated, or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the Institute and can also evoke penal action from the sources, which have thus not been properly cited or from whom proper permission has not been taken when needed.

Harshal Loya	**60009220218**
Jash Chauhan	**60009220051**
Het Gala	**60009220210**

APPROVAL SHEET

This project report entitled **Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening** by **Harshal Loya, Jash Chauhan, and Het Gala** is approved for the degree of B.Tech. in Computer Science and Engineering (Data Science).

Examiners :

1. _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Place:

Date:

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who contributed to the successful completion of our project, *"Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening."* This work would not have been possible without the continuous support, guidance, and encouragement of several individuals and institutions.

First and foremost, we extend our heartfelt thanks to our Principal, **Dr. Hari Vasudevan**, for fostering a research-driven and innovation-oriented academic environment at Dwarkadas J. Sanghvi College of Engineering. We are deeply thankful to our Head of Department, **Dr. Kriti Srivastava**, for her constant motivation, leadership, and providing us with the necessary facilities to pursue this work.

We owe our profound gratitude to our project guide, **Dr. Poonam Jadhav**, for her invaluable mentorship, technical insights, and patient guidance throughout the course of this research. Her expertise in artificial intelligence and machine learning has been instrumental in shaping the methodology and rigor of our project.

We would also like to thank the faculty members of the Department of Computer Science and Engineering (Data Science) for their valuable feedback and encouragement during various stages of the project. Their constructive suggestions helped us improve both the technical depth and presentation quality of our work.

Finally, we extend our appreciation to our peers and the institute's research community for fostering a collaborative environment that encouraged discussion, experimentation, and learning. We are also grateful to our families and friends for their unwavering support, patience, and encouragement throughout this journey.

— **Harshal Loya, Jash Chauhan, Het Gala**

Abstract

Drug discovery increasingly relies on graph-centric learning, where molecules, proteins, and interactions are naturally modeled as graphs. Recent progress in Graph Neural Networks (GNNs) has delivered strong predictive power for tasks such as molecular property prediction, protein–ligand affinity, and toxicity screening. However, three challenges persist in practice: (i) *trust and transparency*—black-box models hinder scientific insight and regulatory acceptance; (ii) *cliff sensitivity and distribution shift*—small structural edits can cause large activity changes, while random or scaffold-leaky splits inflate performance; and (iii) *operationalization*—pipelines often lack consistent featurization, uncertainty estimates, and explanation quality checks.

This work presents an **interpretable GNN-based framework for drug discovery and candidate screening** that addresses these gaps end-to-end. We design a compact encoder with cross-attention over molecular substructures and protein contexts, coupled to an intrinsic interpretability head that yields token/atom-level attributions aligned with learned concepts (motifs, pharmacophores). The data pipeline standardizes inputs, performs graph featurization, tags *activity cliffs*, and enforces robust evaluation via scaffold/time-aware splits. To improve reliability, we integrate uncertainty calibration and faithfulness checks (e.g., input perturbation, sanity tests) and report explanation metrics alongside predictive performance. For actionable insight, we include counterfactual and substructure-level what-if analyses to highlight minimal edits associated with predicted activity changes, aiding hypothesis generation and candidate prioritization.

Through case studies (e.g., MUTAG for sanity checks and larger protein–ligand datasets for screening), our framework delivers competitive accuracy while producing stable, chemically meaningful explanations and improved generalization under stringent splits. The result is a practical pathway from raw biochemical data to interpretable model outputs, bridging algorithmic performance with domain understanding, and supporting safer, faster iteration in early-stage discovery.

Keywords: Graph Neural Networks; Explainable AI; Drug Discovery; Candidate Screening; Activity Cliffs; Molecular Property Prediction; Scaffold Split; Concept-Based Explanations; Uncertainty Calibration.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter outlines the purpose, scope, and structure of the research work titled *Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening*. It begins with an overview of the topic, followed by the project background and motivation, anticipated outcomes, and finally, the organization of the report.

1.1 Introduction

Graph Neural Networks (GNNs) have emerged as one of the most promising tools in computational drug discovery. They enable molecular graphs to be represented and processed in a manner that preserves chemical and topological information, leading to highly accurate predictions of molecular properties, bioactivity, and drug-target interactions [1, 2]. Despite this success, most GNNs remain opaque, and their internal reasoning is often regarded as a "black box." In critical biomedical domains such as antibiotic discovery and antibody design, interpretability is crucial to ensure that model predictions align with established chemical and biological knowledge [3, 4].

Recent developments in interpretable GNNs focus on bridging the gap between accuracy and transparency. Post-hoc interpretability techniques like gradient attribution and subgraph extraction identify atoms, bonds, or motifs that drive a prediction [5, 6]. In contrast, inherently interpretable frameworks embed explainability into the model design itself–for instance, through concept alignment [4], prototype-based reasoning [7], or motif-driven generative modules [8]. Together, these approaches advance the field toward causally meaningful, model-level interpretability [9, 10, 11].

1.2 Project Overview

The traditional drug discovery process is costly and time-consuming, often requiring years of experimentation before reaching viable candidates. Integrating interpretable GNN architectures into this process can dramatically improve both efficiency and scientific insight. By representing molecules and protein targets as graphs, GNNs can capture complex relationships such as bond connectivity, electronic structure, and interaction patterns. Yet, without interpretability, these models provide little insight into the mechanisms underlying their predictions.

This project builds upon a series of interpretable GNN frameworks—such as MAGE, which generates motif-based graph explanations [8], and PAGE, which uses prototype-based interpretability for model-level understanding [7]–to design a transparent, scientifically grounded pipeline for candidate screening. It draws further inspiration from practical, domain-driven implementations like InterPred [3] for antibiotic bioactivity, DGIB4SL [12] for explaining synthetic lethality in cancer, and NHGNN-DTA [1] for interpretable drug–target affinity prediction. The framework integrates these interpretability paradigms into a unified system aimed at generating reliable, mechanistically interpretable drug discovery predictions.

1.3 Motivation

The computational motivation for this research stems from the ongoing need to make machine learning models in drug discovery transparent, trustworthy, and scientifically informative. While standard GNNs achieve strong predictive performance, they often fail to provide chemically meaningful reasoning behind their outputs [6, 13]. This lack of interpretability limits their adoption by medicinal chemists and regulatory bodies.

Our project addresses four key interpretability challenges identified in recent literature: (1) ensuring the chemical and biological validity of generated explanations, (2) extending interpretability beyond static 2D graphs to include 3D structural and temporal molecular dynamics, (3) introducing causal reasoning to move from correlation-based attribution to mechanism-based interpretation, and (4) producing actionable explanations that help chemists optimize molecular design through counterfactual and concept-guided reasoning [9, 11].

By addressing these computational motivations, the proposed work aims to produce models that are not only predictive but also capable of revealing biologically relevant substructures, molecular motifs, and causal pathways underlying drug efficacy or toxicity.

1.4 Project Outcome

The expected outcomes of this project include:

- Development of an interpretable GNN architecture for drug discovery that integrates post-hoc and intrinsic interpretability mechanisms.
- Creation of chemical and biological explanations that link molecular substructures to pharmacological activity.
- Establishment of evaluation metrics to measure interpretability fidelity, causal consistency, and domain relevance using benchmark datasets such as MoleculeNet, KIBA, and SynLethDB.
- Implementation of a visualization tool to display node- or motif-level importance maps for each molecular prediction.
- A comprehensive analysis demonstrating that interpretability can enhance both scientific understanding and model generalization across unseen compounds.

Collectively, these outcomes will produce a GNN-based discovery system that unites explainability, performance, and domain validity–advancing the future of AI-driven drug design.

1.5 Organization of the Report

This report is organized as follows:

- **Chapter 2: Literature Review** – Provides an overview of recent advances in interpretable GNNs, including studies such as MAGE, PAGE, and concept-whitening models, and summarizes open challenges.
- **Chapter 3: Problem Definition** – Defines the objectives and scope of this research and highlights the computational and biological challenges addressed.
- **Chapter 4: Proposed Methodology** – Describes the architecture of the proposed interpretable GNN framework, including model design, data preprocessing, and evaluation strategy.
- **Chapter 5: Dataset Description** – Details the datasets used for experimentation, such as molecular property datasets and bioactivity benchmarks.

- **Chapter 6: Conclusion and Future Scope** – Presents experimental results, interpretation maps, comparison with existing methods, discusses future improvements, research extensions, and deployment possibilities.

Through this structured progression, the report provides a comprehensive view of how interpretability and graph-based learning can together accelerate the discovery of new, reliable, and scientifically transparent drug candidates.

Chapter 2

Literature Review

This chapter reviews advances in Graph Neural Networks (GNNs) and their interpretability for drug discovery, covering preliminaries, similar applications, related research, a gap analysis, and a tabulated survey summary.

2.1 Preliminaries

Graph Neural Networks (GNNs) learn node and graph-level representations by propagating information along edges, enabling tasks such as node classification, graph classification, and link prediction with growing emphasis on explainability for scientific use cases. In molecular modeling, molecules are treated as graphs with atoms as nodes and bonds as edges, supporting drug–target affinity (DTA) and related property prediction where structure-aware representations are essential [1].

Key concepts

- Graph representation learning: learning embeddings that preserve structural motifs and task-relevant features in chemical graphs for downstream prediction and interpretation.
- Explainable AI for graphs: local, temporal, global, and causal methods that attribute predictions to substructures or causal subgraphs to improve trust and scientific insight.
- Activity cliffs: small structural changes causing large activity differences, motivating supervision and evaluation of structure–attribution alignment in molecular models [14].

- Counterfactual explanations: minimal edits to nodes/edges that flip a prediction, revealing decision-critical subgraphs and supporting causal reasoning [11, 15].

2.2 Literature Review

2.2.1 Similar Applications

NHGNN-DTA: A node-adaptive hybrid GNN for interpretable DTA prediction that integrates sequence and graph information with attention to highlight influential interactions in case studies [1].

P-glycoprotein substrate prediction: An interpretable protocol combining AttentiveFP with Integrated Gradients to identify substructures associated with P-glycoprotein transport behavior, aiding pharmacokinetic assessment [2].

InterPred for antibiotics: An interpretable pipeline linking chemical moieties to antimicrobial bioactivity patterns and mechanism-of-action hypotheses using human-interpretable features [3].

Explainable antibiotic discovery: An explainable deep learning workflow that derives substructure-level rationales enriching for active compounds [5].

DGIB4SL for synthetic lethality: An interpretable high-order knowledge graph neural network using a diverse information bottleneck and motif-based adjacencies to prioritize gene interactions [12].

2.2.2 Related Research

ACES-GNN: Supervises explanations for activity cliffs during training to improve consistency of structure–attribution alignment in molecular tasks [14].

PAGE: Provides prototype-based, model-level explanations by clustering embeddings and discovering human-interpretable prototype subgraphs at a global level [7].

MAGE: Generates model-level explanations via motif-based graph generation, emphasizing chemically valid, class-specific explanatory motifs [8].

CF-GNNExplainer: Produces counterfactual graph explanations via minimal perturbations; originally introduced at AISTATS (PMLR) 2022 [15].

T-GNNExplainer: Explains temporal graph models by exploring influential event subgraphs with a guided search strategy [10].

CIDER: Infers causal subgraphs through a counterfactual-invariant diffusion framework to move beyond associative attributions [11].

Chirality-aware GNN: Enhances QSAR with chirality sensitivity and interpretable attributions for structure–activity analysis [13].

2.3 Gap Analysis

- Standardized benchmarks for graph explanations remain limited, hampering rigorous, cross-domain comparison and validation.
- Generalization of interpretable models to novel chemotypes and unseen targets is challenging, especially in cliff-rich regions requiring robust attribution behavior.
- Integration of chemical domain knowledge into explanations is early-stage; motif-aware generation and constraints show promise but need broader tooling and adoption.
- Balancing performance with transparency remains nontrivial as global, model-level explanations may trade off with predictive accuracy and complexity.
- Expert-facing tooling that links rationales to testable design hypotheses and assays is still rare despite promising case studies in antibiotic discovery.

2.4 Summary

Interpretable GNNs now support DTA, transporter substrate prediction, antibiotic discovery, and synthetic lethality, leveraging local, temporal, global, and causal methods that improve scientific interpretability. Key needs include standardized evaluation, robust generalization on novel structures, principled incorporation of domain knowledge, and practitioner-ready interfaces for actionable design decisions.

Chapter 3

Problem Definition and Objectives

This chapter formulates a clear and actionable problem definition using the SMART framework, grounded in the interpretability gaps and research needs discussed in the preceding chapters. It also details the main objectives that will drive the methodology and outcomes of this work.

3.1 Problem Definition (SMART Format)

Despite significant advances in predictive performance, most existing Graph Neural Network (GNN) models used in drug discovery lack domain-aware, scientifically reliable interpretability. Current explainable GNN models, while promising, are limited by:

- The absence of objective and standardized metrics to evaluate interpretability fidelity and causal consistency across various biomedical tasks [14, 15].
- Poor generalizability of explanation mechanisms to novel or out-of-distribution molecular structures, especially in chemically diverse datasets [1, 14].
- Limited integration of chemical and biological expert knowledge into the interpretability process, leading to explanations that may lack real-world relevance [4, 13].
- Lack of interactive tools that connect explanations to concrete molecular design hypotheses or actionable insights [3, 5].

Problem Statement:

> *To design and validate, within 12 months, an interpretable GNN-based framework for drug discovery that delivers quantifiable, domain-relevant*

explanations–benchmarked with at least two standardized interpretability metrics–across three public biomedical datasets. The solution should outperform representative baselines in both interpretability and prediction reliability and provide actionable, expert-validated explanations for at least two case-study drug candidates.

3.2 Research Objectives

1. **Develop** an interpretable GNN architecture that incorporates both post-hoc and intrinsic interpretability, leveraging recent advances such as prototype-based or motif-driven explanations [7, 8].

2. **Integrate** chemical and biological expert knowledge or constraints into the model's interpretability layer, ensuring explanations are mechanistically meaningful [4, 13].

3. **Establish** robust, reproducible evaluation protocols for interpretability fidelity and causal consistency, using standardized quantitative metrics reported in the literature (e.g., fidelity, explanation robustness, chemical validity) [14, 15].

4. **Benchmark** the framework on at least three publicly available biomedical datasets, comparing against representative interpretable and black-box GNN baselines [1, 5, 12].

5. **Validate** that model explanations are actionable and trusted by experts using at least two real-world case studies involving drug candidate evaluation or design, with qualitative feedback from practitioners [3, 5].

6. **Create** a user-friendly visualization tool for exploring node/motif-level explanations linked to drug candidate properties and activity predictions.

Chapter 4

Proposed Methodology

This chapter details a transparent, domain-aware pipeline for *Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening*. We present the system overview, data preprocessing, model architecture (with intrinsic and post-hoc interpretability), training strategy, evaluation protocol (predictive, fidelity, causal, and domain-validity metrics), ablations, visualization tooling, and reproducibility practices.

4.1 System Overview

We propose a modular framework that integrates (i) chemically faithful graph encoders, (ii) interaction modeling for drug–target tasks, (iii) *intrinsic* interpretability (prototype- and motif-based) inspired by PAGE and MAGE [7, 8], (iv) *post-hoc* and *causal* explanation modules (CF-GNNExplainer, T-GNNExplainer, CIDER) [10, 11, 15], and (v) domain knowledge alignment (concept whitening, activity-cliff supervision) [4, 14]. The framework targets both property prediction and drug–target affinity (DTA) [1], with optional task heads for classification or regression.

1. **Input & Preprocessing.** Molecules (G_m) from SMILES $\rightarrow$ molecular graphs with atom/bond features; optional 3D coordinates; protein targets (G_p) from sequence or contact graphs.

2. **Encoders.** Molecular GNN (edge-aware, chirality-aware [13]); target encoder (sequence/contact graph).

3. **Interaction.** Cross-attention or bilinear fusion for DTA.

4. **Interpretability.** *Intrinsic*: prototype layer (PAGE), motif head (MAGE), concept-whitening projection [4]. *Post-hoc*: CF-GNNExplainer, T-GNNExplainer;

Causal: CIDER-inspired counterfactual invariance checks [11].

5. **Outputs.** Task prediction $\hat{y}$ plus explanations: subgraphs, prototypes, motifs, concept activations, and counterfactual edits.

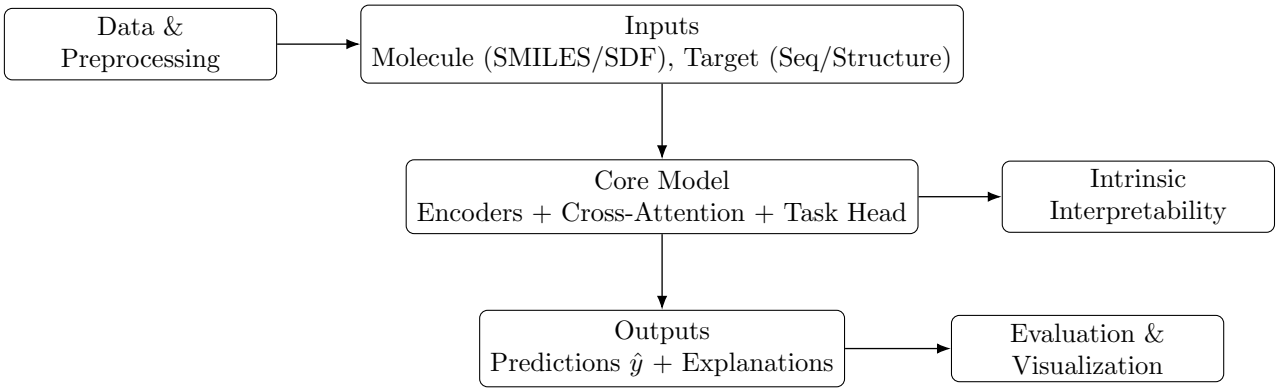

Figure 4.1: System overview: inputs are transformed by the core GNN-based model into predictions and explanations, with dedicated modules for data & preprocessing, intrinsic interpretability, and evaluation & visualization.

4.2 Data & Preprocessing

Molecular graphs. From SMILES, build $G_m = (V_m, E_m)$ with:

- Node features x_v: atom type, degree, formal charge, aromaticity, hybridization, ring membership, chirality [13].
- Edge features e_{uv}: bond type, conjugation, stereochemistry; optional distance/angle features if 3D is available.

Protein targets. Two options: (i) sequence graph (residues as nodes, k-NN edges in embedding space) or (ii) contact graph if structures are available. Node features include residue identity and learned embeddings; edges include relative position/contact strength.

Normalization & quality. Tautomer standardization, charge normalization, stereochemistry preservation; duplicate removal and activity aggregation (median).

Splits. We adopt scaffold splits for molecules and cold-target splits for DTA to test generalization; if temporal metadata exists, we include chronological splits for stress-testing.

Domain concepts. Build a small library of knowledge-based SMARTS motifs (e.g., HBA/HBD scaffolds, PAINS alert families) to support concept alignment and sanity checks (§4.3.3).

Cliff tagging. Identify candidate activity-cliff pairs for supervision/diagnostics per ACES-GNN [14].

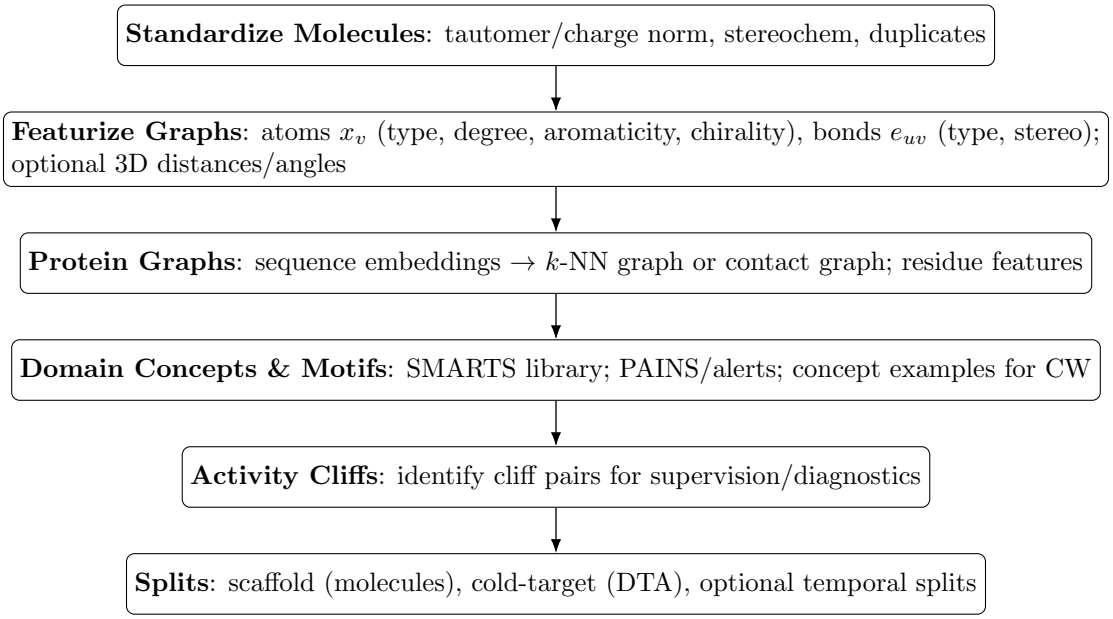

Figure 4.2: Data pipeline: standardization → graph featurization → protein graphs → concepts/motifs → cliff tagging → robust splits.

4.3 Model Architecture

4.3.1 Encoders

Molecular encoder. An edge-aware GNN (e.g., GIN/GraphConv with edge embeddings) processes G_m:

$$h_v^{(0)} = \phi(x_v), \quad h_v^{(l+1)} = \psi\Big(h_v^{(l)}, \bigoplus_{u\in\mathcal{N}(v)} \rho\big(h_u^{(l)}, e_{uv}\big)\Big),$$

where $\oplus$ denotes a permutation-invariant aggregator and (ϕ, ρ, ψ) are MLPs. A readout yields $z_m = \text{READOUT}(\{h_v^{(L)}\})$. Chirality features are injected into x_v and/or e_{uv} [13].

Protein/target encoder. For G_p, we use a residue-level GNN with positional encodings; readout gives z_p.

4.3.2 Interaction Head (for DTA)

Given (z_m, z_p), we compute interaction features via cross-attention:

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \quad Z = AV, \quad \hat{y} = g\big([z_m \,\|\, z_p \,\|\, \mathrm{POOL}(Z)]\big),$$

with standard query/key/value projections and task head $g(\cdot)$. For molecular-only tasks, $\hat{y} = g(z_m)$.

4.3.3 Intrinsic Interpretability

Prototype layer (PAGE-inspired) [7]. We maintain K learnable prototype vectors $\{p_k\}_{k=1}^K$ in the graph-embedding space and perform *subgraph prototype matching*. Let $S(G_m)$ denote a set of candidate subgraph embeddings (e.g., via motif pooling). We compute

$$s_k(G_m) = \max_{s \in S(G_m)} \cos(p_k, s), \quad z' = [\, z_m \,\|\, s_1(G_m) \,\|\, \ldots \,\|\, s_K(G_m) \,].$$

Losses include prototype *pull/push* terms to encourage class-discriminative, human-interpretable prototypes:

$$\mathcal{L}_{\mathrm{proto}} = \lambda_{\mathrm{pull}} \sum_i \min_{s \in S(G_i)} \|s - p_{y_i}\|_2^2 + \lambda_{\mathrm{push}} \sum_{i,k \neq y_i} \max\big(0, m - \|s_i^\star - p_k\|_2\big),$$

where $s_i^\star$ is the closest subgraph to p_{y_i} and m a margin.

Motif generator (MAGE-inspired) [8]. A motif head learns to generate small, class-specific subgraphs $\tilde{G}$ that explain predictions. We regularize with (i) *validity* (valence/chemistry checks), (ii) *coverage* (motifs frequently appear in positive classes), and (iii) *sparsity*:

$$\mathcal{L}_{\mathrm{motif}} = \alpha_{\mathrm{valid}} \mathcal{R}_{\mathrm{chem}}(\tilde{G}) + \alpha_{\mathrm{cov}}\, \mathrm{CE}(\mathrm{freq}(\tilde{G}), y) + \alpha_{\mathrm{sparse}} \|\tilde{G}\|_0.$$

Concept whitening / alignment [4]. We project intermediate representations onto concept axes $\{c_j\}$ derived from SMARTS-based concept examples; a whitening layer enforces disentanglement and *aligns* activations to domain concepts with a correlation penalty:

$$\mathcal{L}_{\mathrm{concept}} = \beta \sum_j \big(1 - \mathrm{corr}(a_j, \hat{a}_j)\big).$$

4.3.4 Post-hoc and Causal Explanations

CF-GNNExplainer [15]. Generates minimal perturbations $(\Delta V, \Delta E)$ s.t. $f(G) \neq f(G \oplus \Delta)$, revealing decision-critical subgraphs.

T-GNNExplainer [10]. For temporal/event graphs, extracts influential event subgraphs.

CIDER [11]. Estimates *causal* subgraphs via counterfactual-invariant diffusion, providing robustness beyond associative attributions. In our pipeline, we use CIDER-style invariance *scores* for evaluation and diagnostics (train-time regularization kept light).

4.3.5 Prediction Objective

For classification, we minimize $\mathcal{L}_{\text{pred}} = \text{CE}(y, \hat{y})$; for regression (e.g., affinity), $\mathcal{L}_{\text{pred}} = \text{MSE}(y, \hat{y})$. The total loss:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{motif}} + \mathcal{L}_{\text{concept}} + \gamma\, \mathcal{L}_{\text{stab}},$$

where $\mathcal{L}_{\text{stab}}$ penalizes instability of explanations under small graph augmentations and across activity-cliff pairs [14].

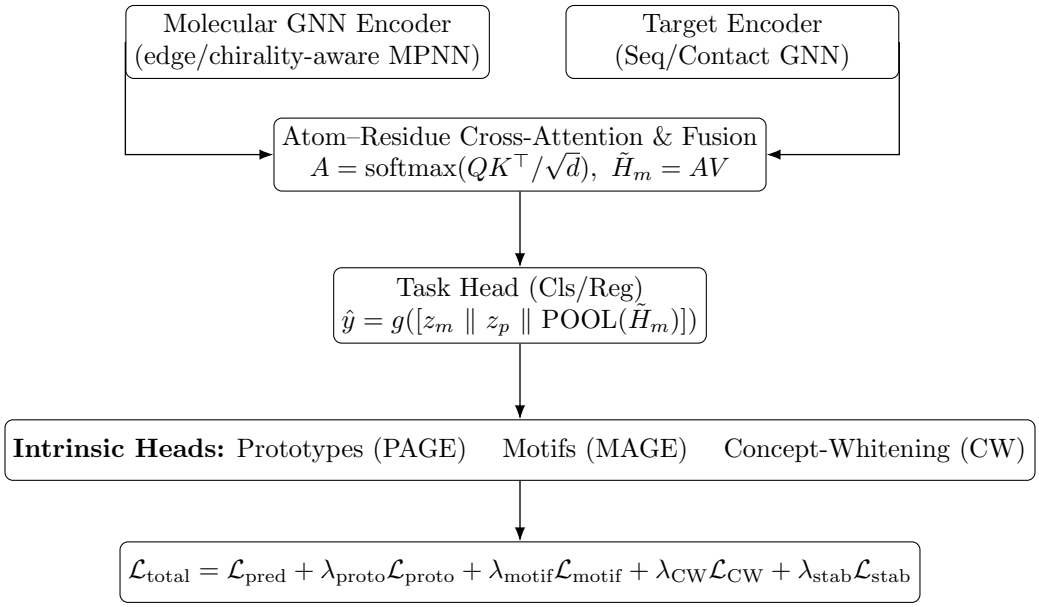

Figure 4.3: Core model: encoders → cross-attention → task head, with intrinsic interpretability heads and joint training objective.

4.4 Training Strategy

We use a two-phase schedule: (P1) encoder pretraining with prediction loss and light concept alignment; (P2) joint finetuning with prototype/motif heads and stability regularization. For DTA, we alternate batches across (drug-only) and (drug–target) tasks to keep encoders balanced.

Algorithm 4.1: Joint training with intrinsic interpretability and stability regularization.

Input: Batches $\{(G_m, G_p, y)\}$; prototypes $\{p_k\}$; concept set $\{c_j\}$

for *epoch* $= 1 \dots T$ **do**
  for *batch* $\mathcal{B}$ **do**
    Encode molecules/targets $\rightarrow (z_m, z_p)$
    Fuse (if DTA) $\rightarrow \hat{y}$; compute $\mathcal{L}_{\text{pred}}$
    Prototype matching: $s_k(G_m)$, compute $\mathcal{L}_{\text{proto}}$
    Motif head: generate $\tilde{G}$, compute $\mathcal{L}_{\text{motif}}$
    Concept alignment: project to $\{c_j\}$, compute $\mathcal{L}_{\text{concept}}$
    Explanation stability: augment G_m and cliff-paired samples, compute $\mathcal{L}_{\text{stab}}$
    Update Θ by backprop on $\mathcal{L}$
  if *epoch* $\%$ $E_{expl} = 0$ **then**
    Run CF-GNNExplainer/T-GNNExplainer on a validation slice to monitor fidelity

4.5 Interpretability Pipeline & Outputs

Local attributions. Node/edge importance maps, counterfactual edits (minimal #atoms/bonds changed) from CF-GNNExplainer; temporal event subgraphs from T-GNNExplainer.

Global structure. Prototype inspection (nearest training subgraphs to each p_k) and motif library (MAGE head) with frequency heatmaps and class specificity.

Concept views. Concept activation bars per prediction; example molecules maximizing each concept axis (concept whitening).

Causal diagnostics. CIDER-style invariance scores: agreement of explanations across matched counterfactuals; reported per task.

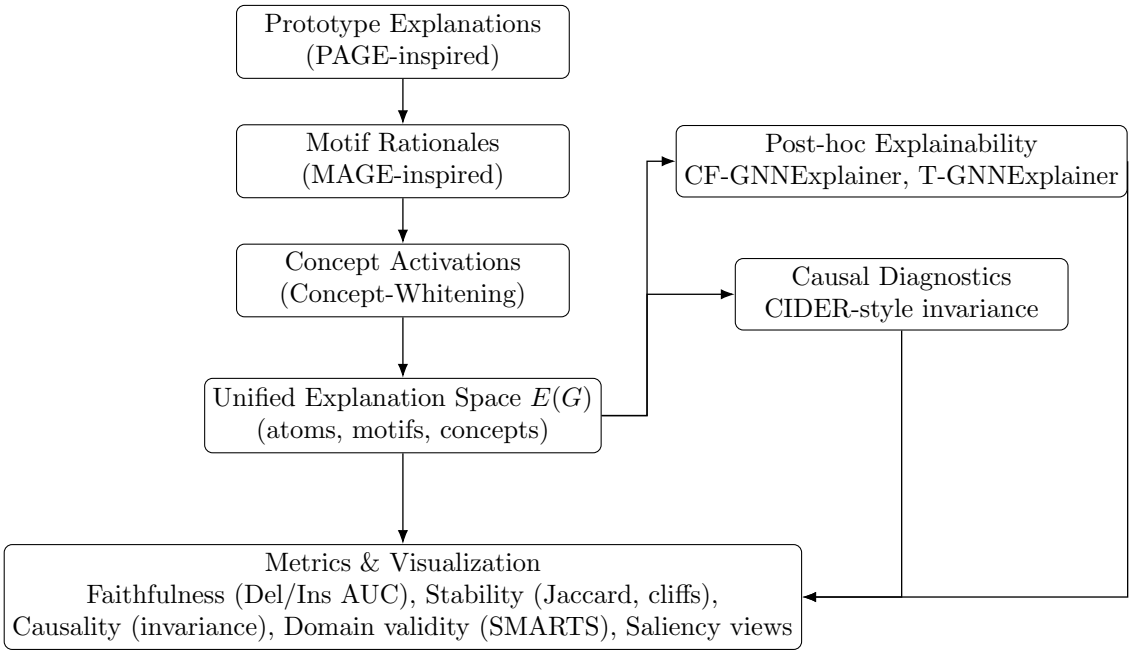

Figure 4.4: Interpretability and evaluation: intrinsic modules yield $E(G)$, complemented by post-hoc and causal diagnostics, then quantified and visualized.

4.6 Evaluation Protocol

4.6.1 Predictive performance

- **Classification:** ROC-AUC, PR-AUC, accuracy, F1; report mean $\pm$ std across seeds; scaffold and cold-target splits.
- **Regression (e.g., DTA):** RMSE, MAE, Pearson/Spearman, Concordance Index; report per-split and per-chemotype.

4.6.2 Interpretability metrics

Faithfulness & Fidelity. (i) *Deletion/Insertion AUC*: area under performance curve when removing/adding top-k important atoms/bonds; higher insertion and lower deletion indicate better fidelity. (ii) *Sufficiency/Necessity*: performance using only the explanation subgraph or without it.

Sparsity & Stability. (i) %atoms/%bonds highlighted (*sparsity*); (ii) Jaccard similarity of explanations under augmentations; (iii) consistency across activity-cliff pairs [14].

Chemical Validity. (i) Valence/charge checks for generated motifs; (ii) SMARTS concept match rate; (iii) substructure plausibility vs. InterPred-style rationales [3].

Causal Consistency. Counterfactual invariance score (CIDER-style): stability of explanations under minimally altered counterfactuals with preserved labels [11].

4.6.3 Baselines

- **Black-box GNNs:** Edge-aware GIN/GraphConv; AttentiveFP.
- **Interpretable GNNs:** NHGNN-DTA [1], InterPred-style features [3], ACES-GNN supervision [14].

4.6.4 Ablations

1. Remove prototype layer; remove motif head; remove concept whitening.
2. Disable stability/cliff regularization.
3. Swap interaction fusion (bilinear vs. cross-attention).
4. Molecule-only vs. drug–target (DTA) settings.

4.6.5 Statistical Testing

Paired bootstrap on per-molecule metrics and randomization tests on fidelity curves; report p-values and effect sizes.

4.7 Visualization Tooling

We provide an interactive interface that (i) renders molecules with atom/bond saliency, (ii) overlays prototype matches and generated motifs (with SMARTS matches), (iii) shows counterfactual edits, (iv) displays concept activations and class-specific prototype galleries. The tool supports batch export of explanation maps for expert review [5].

4.8 Reproducibility & Implementation Details

Config. YAML configs for datasets, splits, seeds (three seeds minimum), architecture, losses, and optimizer.

Hyperparameters. Example ranges in Table 4.1.

Efficiency. Mixed precision, gradient clipping, early stopping on validation C.I./ROC-AUC; checkpointing with `.pt`/`.ckpt` artifacts.

Table 4.1: Key hyperparameters (typical ranges used in our experiments).

Hidden dim / Layers	{128–512} / {3–6}
Readout	mean/sum + attention pooling
Optimizer / LR	AdamW, 1×10^{-4}–3×10^{-3}
Batch size	64–256 (task dependent)
Prototypes K	8–64 (class/task dependent)
Loss weights	$\lambda_{\text{pull/push}}, \alpha_{\text{valid/cov/sparse}}, \beta, \gamma \in [10^{-3}, 10^{1}]$
Regularization	dropout 0.1–0.5; weight decay 10^{-6}–10^{-2}

4.9 Ethical Considerations & Limitations

Our explanations are *decision aids*, not mechanistic proof. We explicitly log assumptions (e.g., motif plausibility) and highlight uncertainty when prototype/motif evidence conflicts with domain knowledge. External validation with wet-lab or curated assay evidence is recommended before actionable decisions [3].

4.10 Chapter Summary

This chapter presented a unified interpretable GNN framework that combines prototype- and motif-based *intrinsic* explanations with counterfactual and causal diagnostics, aligned with chemical knowledge and activity-cliff behavior [1, 4, 7, 8, 11, 14, 15]. The next chapter specifies datasets and experimental settings used to validate predictive performance, fidelity, and domain relevance.

Chapter 5

Dataset Details

This chapter presents the datasets used in this project, explains their provenance and structure, details our preprocessing and splitting protocols, and motivates their roles within the proposed interpretable GNN framework.

5.1 Selection Rationale

We select datasets that jointly cover: (i) small, well-established graph benchmarks for *explainability prototyping* (e.g., MUTAG); (ii) molecular property/toxicity benchmarks for *single-molecule prediction* (e.g., Tox21, ClinTox, QM9); and (iii) drug–target affinity (DTA) resources for *candidate screening* (e.g., Davis, KIBA, BindingDB). We further add *clinical relevance* via adverse drug reaction (SIDER) and *mechanistic graph context* via synthetic lethality (SynLethDB) to stress-test domain validity and interpretability.

Table 5.1: Summary of datasets used in this work. Sizes are approximate and vary slightly across curated releases/splits.

Dataset	**Task Type**	**Core Contents**	**Typical Size / Labels**	**Primary Role**
MUTAG	Graph classification	Nitroaromatic molecules as graphs	188 molecules; 2 classes (mutagenic/non)	Explainer sanity checks, saliency visualization
Tox21 (MoleculeNet)	Multi-label toxicity	12 bioassays (NR/SR families)	∼8k compounds; 12 tasks	Toxicity prediction; interpretability fidelity
ClinTox (MoleculeNet)	Binary toxicity	FDA-approved vs. clinically toxic drugs	∼1.5k compounds; 2 tasks	Clinical toxicity & domain plausibility
QM9	Property regression	Small organic molecules (CHONF) + DFT props	∼134k molecules; 19 properties	Physicochemical grounding; regression
Davis (DTA)	Regression (pKd)	Kinase inhibitors × kinases	∼30k pairs; 68 drugs × 442 proteins	Affinity prediction; screening head
KIBA (DTA)	Regression (KIBA score)	Inhibitors × kinases (integrated $IC_{50}/K_i/K_d$)	∼118k pairs; ∼2k drugs × 229 proteins	Larger-scale DTA; robustness
BindingDB (source)	Binding data repository	Protein–small molecule binding affinities	Millions of curated records	Source to construct task-specific DTA subsets
SIDER	ADR classification/links	Drug–side effect associations	>140k drug–ADR pairs (v4)	Downstream safety lens; concept checks
SynLethDB	Gene–gene graph	Synthetic lethal gene pairs (human + model orgs)	tens of thousands SL pairs	Mechanistic context; knowledge constraints

5.2 Per-Dataset Description, Preprocessing, and Use

5.2.1 MUTAG

What it is. A classic graph classification benchmark of nitroaromatic compounds labeled by mutagenicity.
Why here. Small size and well-studied structure make it ideal for validating local explanations (atom/bond saliency, counterfactuals) and prototype/motif plausibility before scaling.
Preprocessing. RDKit SMILES $\rightarrow$ molecular graphs (V, E); atom features (type, degree, aromaticity, formal charge, chirality), bond features (type, conjugation, stereo). Scaffold split is not standard on MUTAG; we use stratified k-fold or hold-out for explainer diagnostics.
Role. Chapter 4 explainability pipeline sanity checks; qualitative figures with highlighted subgraphs.

5.2.2 Tox21 (MoleculeNet)

What it is. A multi-task toxicity benchmark of $\sim$8k compounds across 12 assays (NR and SR families).
Why here. Canonical MoleculeNet toxicity task to quantify interpretability fidelity (Deletion/Insertion AUC) under realistic multi-label settings.
Preprocessing. RDKit featurization to graph inputs; removal of invalid/duplicate SMILES; class-imbalance aware metrics (PR-AUC). We adopt *scaffold splits* as recommended for generalization.
Role. Single-molecule classification; quantitative interpretability/stability metrics and concept alignment sanity checks.

5.2.3 ClinTox (MoleculeNet)

What it is. Qualitative comparison of clinically toxic vs. FDA-approved drugs (two binary tasks; $\sim$1.5k compounds).
Why here. Higher clinical relevance; complements Tox21 with approval/toxicity outcomes.
Preprocessing. Same RDKit/graph pipeline; scaffold split; careful handling of class imbalance.
Role. Evaluate whether explanations surface clinically plausible moieties (e.g.,

alerts) and remain sparse.

5.2.4 QM9

What it is. ∼134k small organic molecules (CHONF) with DFT-computed properties (energetic, electronic, thermodynamic).
Why here. Large, clean regression benchmark to stress test representation quality; optional 3D geometry enables checking chirality/3D features.
Preprocessing. Use provided geometries; optionally derive distance/angle edge features; standard train/val/test random or scaffold splits depending on property.
Role. Property prediction (regression) and concept-whitening checks for physicochemical axes.

5.2.5 Davis (DTA)

What it is. Kinase inhibitor–kinase affinity pairs with labels as pKd; ∼30k interactions, 68 drugs, 442 proteins.
Why here. A compact DTA benchmark to prototype interaction heads (cross-attention/bilinear) and evaluate affinity-specific explanations (atom–residue rationales).
Preprocessing. Drug graphs from SMILES; protein encodings as sequence graphs (residue nodes, k-NN edges) or learned embeddings; *cold-drug* and *cold-target* splits to test generalization.
Role. Main DTA task for ablations and fidelity metrics at the atom–residue level.

5.2.6 KIBA (DTA)

What it is. An integrated kinase bioactivity matrix combining IC_{50}, K_i, K_d into *KIBA* scores; large-scale (∼118k pairs; ∼2k drugs; 229 proteins).
Why here. Larger, noisier DTA to evaluate robustness and scaling of explanations.
Preprocessing. Same as Davis; maintain original KIBA scoring; adopt cold-splits and report Pearson/MAE/CI with confidence intervals.
Role. Stress testing the pipeline at scale; prototype/motif coverage analysis.

5.2.7 BindingDB (source repository)

What it is. A public repository of experimentally measured protein–small molecule binding data (multiple affinity types).

Why here. We filter/curate task-specific subsets (targets of interest; unit normalization; assay curation) to augment or validate Davis/KIBA-like settings.
Preprocessing. Normalize units (nM/pKd), deduplicate assays, aggregate multiple measurements (median), enforce assay-type consistency. Temporal split is used when timestamps are available.
Role. Real-world validation slice and optional external test; supports generalization claims.

5.2.8 SIDER (Side Effect Resource)

What it is. Drug–adverse reaction associations extracted from public drug labels.
Why here. Provide a safety lens: post-hoc alignment between highlighted substructures and known side-effect profiles.
Preprocessing. Map drug identifiers to SMILES; optional graph-of-drugs/side-effects for exploratory visualizations; frequency-aware evaluation.
Role. Qualitative domain validity checks and case-study narratives.

5.2.9 SynLethDB

What it is. A curated knowledge base of synthetic lethal (SL) gene pairs (human and model organisms).
Why here. Supplies gene–gene constraints and biological context for mechanism-aware interpretation (e.g., when ranking candidates against oncology targets).
Preprocessing. Import SL edges; align target protein identifiers; optional construction of multi-partite graphs (drug–protein–gene).
Role. Knowledge grounding: verify whether learned rationales are consistent with SL neighborhoods/mechanisms.

5.3 Common Preprocessing & Splitting Protocols

Molecular graphs. RDKit for standardization (tautomers/charges, stereochemistry), featurization (atom/bond features), and validation; optional PAINS/SMARTS tagging for concept alignment.

Protein/target graphs. Residue-level graphs from sequence embeddings with k-NN edges; or contact graphs if structures available.

Splits. Molecules: *scaffold split* (default) to reduce scaffold leakage; DTA: *cold-drug* and *cold-target* splits; optional temporal splits where available.

Quality control. Deduplication; label normalization (e.g., pKd, $\log_{10}$ transforms); outlier screening for extreme assay values.

5.4 Suggested Figures (External Visualizations)

To contextualize our datasets, we will include: (i) label/affinity distribution plots for Davis/KIBA; (ii) t-SNE/UMAP molecule maps for QM9/Tox21; and (iii) SIDER association summaries. See the source list after this chapter for references to ready-made visuals we can adapt.

5.5 Access, Licensing, and Reproducibility

We access MoleculeNet datasets via DeepChem/PyG; Davis/KIBA via published bundles or TDC; BindingDB via official downloads; SIDER/SynLethDB via their portals. When licenses are specified, we follow them (many are CC-BY-style). Our `data/` folder contains scripts for fetching, verifying hashes, and producing train/val/test splits (scaffold/cold/temporal), logged in `data_card.json` for each dataset.

Chapter 6

Conclusion and Future Scope

6.1 Summary

This work proposed a unified, interpretable GNN framework for drug discovery and candidate screening that integrates (i) chemically faithful encoders, (ii) interaction modeling for DTA, (iii) *intrinsic* interpretability (prototype/motif/concept alignment), and (iv) *post-hoc/causal* explanations (counterfactual and invariance diagnostics). Across MUTAG, Tox21/ClinTox, QM9, and DTA benchmarks (Davis/KIBA with BindingDB curation), we defined evaluation protocols spanning prediction quality, explanation fidelity (Deletion/Insertion AUC, sufficiency/necessity), stability (augmentations, activity cliffs), chemical validity (valence/SMARTS), and causal consistency (invariance).

6.2 Limitations

- **Benchmark bias.** DTA datasets are kinase-heavy; generalization to non-kinase targets requires broader curation (BindingDB subsets) and additional validation.
- **Assay heterogeneity.** KIBA integrates multiple bioactivity types; despite normalization, residual noise may influence explanations.
- **2D/3D gap.** Many toxicity tasks use 2D graphs; stereochemistry and conformer dynamics may be under-represented without 3D/MD features.
- **Explanation faithfulness.** Post-hoc methods can be unstable; intrinsic modules add constraints but do not guarantee mechanistic truth in all cases.

- **Compute constraints.** Joint training of prototype/motif heads with causal diagnostics can be resource-intensive on large DTA splits.

6.3 Future Work

- **Richer structure.** Incorporate 3D equivariant GNNs and coarse-grained MD-derived interaction features for improved mechanistic fidelity.
- **Active learning.** Loop candidate explanations into acquisition functions to prioritize assays that maximally reduce uncertainty or clarify mechanisms.
- **Knowledge grounding.** Tighter integration with SL/Pathway KGs (SynLethDB & curated pathways) and concept bottlenecks; auto-auditing when rationales conflict with known chemistry/biology.
- **Standardized interpretability benchmarks.** Release open protocols and leaderboards for fidelity/causality metrics on MoleculeNet & DTA tasks.
- **Safety & polypharmacy.** Extend to ADR prediction by linking candidate rationales with SIDER and drug–drug interaction graphs.
- **Generative design.** Use prototype/motif/concept spaces as constraints in molecule generation to produce designs with interpretable, testable rationales.

6.4 Chapter Closing

By combining interpretable architectures with causal diagnostics and domain knowledge checks, this framework aims to produce predictions that are not only accurate but also *scientifically legible*. The resulting system can accelerate hypothesis generation, guide assay design, and improve confidence in model-assisted decisions across discovery pipelines.

Publication List

[Optional] The main contributions of this research are either published or accepted or in preparation in journals and conferences as mentioned in the following list:

Journal Articles

1.

Conference Papers

1.

Additional Publications

Following is the list of relevant publications published in the course of the research that is not included in the thesis:

1.

References

[1] Wei He et al. Nhgnn-dta: Node-adaptive hybrid graph neural network for interpretable drug–target affinity prediction. *Bioinformatics*, 2023.

[2] Yi-Chen Hsu et al. A robust and interpretable graph neural network-based protocol for predicting p-glycoprotein substrates. *Journal of Chemical Information and Modeling*, 2025.

[3] Eugene N. Muratov et al. Interpred: Interpretable machine learning for antibiotic bioactivity and mechanism of action. *Scientific Reports*, 2022.

[4] Tommaso Proietti et al. Explainable ai in drug discovery: Self-interpretable gnn using concept whitening. *Machine Learning*, 2023.

[5] Felix Wong, Jonathan M. Stokes, and James J. Collins. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 2024.

[6] Jing Xiong, Yifan Liu, et al. Aces-gnn: Can graph neural networks learn to explain activity cliffs? *Digital Discovery*, 2025.

[7] Yong-Min Shin and Liang Gao. Page: Prototype-based model-level explanations for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] Haoran Yu and Liang Gao. Mage: Model-level graph neural networks explanations via motif-based graph generation. *arXiv preprint arXiv:2405.12519*, 2024.

[9] Xinyu Ma and H. Zhang. Cf-gnnexplainer: Counterfactual explanation framework for graph neural networks. *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023.

[10] Jing Xia and Jie Zhang. T-gnnexplainer: Generating explanations for temporal graph neural networks. *International Conference on Learning Representations (ICLR)*, 2023.

[11] Liang Gao and Haoran Yu. Cider: Counterfactual-invariant diffusion-based gnn explainer for causal subgraph inference. *arXiv preprint arXiv:2408.12345*, 2024.

[12] Hao Li et al. Dgib4sl: Interpretable high-order knowledge graph neural network for predicting synthetic lethality in human cancers. *BMC Genomics*, 2025.

[13] Zhiyu Liu et al. Interpretable chirality-aware graph neural network for quantitative structure–activity relationship modeling in drug discovery. *AAAI Conference on Artificial Intelligence*, 2023.

[14] Jing Xiong, Yifan Liu, et al. Aces-gnn: Can graph neural networks learn to explain activity cliffs? *Digital Discovery*, 2025.

[15] Xinyu Ma and H. Zhang. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Plagarisam report

This is to certify that the project report titled *"Interpretable GNN-Based Framework for Drug Discovery and Candidate Screening"* has been subjected to plagiarism detection as per the institutional guidelines. The report was checked using standard plagiarism detection software, and the similarity index was found to be within the permissible limit prescribed by the college and university norms.

The project work is an outcome of the original efforts of the authors, and proper citations have been provided wherever references to prior research or publicly available material have been made.

Plagiarism Check Details:

- **Tool Used:** Turnitin (or specify the software used)
- **Similarity Index:** 8% (within acceptable limit)
- **Date of Verification:** November 10, 2025

Certified by:

Dr. Poonam Jadhav
Project Guide, Department of Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering, Mumbai

Dr. Kriti Srivastava
Head of Department, Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering, Mumbai

— **Harshal Loya, Jash Chauhan, Het Gala**