BIO 425 - Graduation Project in
Bioinformatics (II)

Interpretable Drug Repurposing Platform Based on Graph Neural Networks

منصة قابلة للتفسير لإعادة توظيف الأدوية باستخدام الشبكات العصبية البيانية

Salah Gamal Abdelkhabir
4211099

Ahmed Ebrahim Gabr
4211102

Ahmed Hamdy Eldakroury
4211126

Ahmed Mohamed Ali
4211095

Aliaa Mohamed Ali
4211020

Sara Ayman Ebrahim
4211272

Supervised by:

"Dr. Noha Ezzat Elattar"

Dean:

"Prof.Dr.Hisham Arafat Khalifa"

Faculty of Artificial Intelligence
Delta University for Science and Tachnology

Spring 2024/2025

Table of Contents

List of Tables

Acknowledgment

We, the project team, would like to extend our heartfelt gratitude and appreciation to all those who supported and guided us throughout the development of this graduation project.

First and foremost, we express our deepest thanks to Professor Dr. Noha Elattar, Head of Bioinformatics Department, Faculty of Artificial Intelligence, for her exceptional supervision, insightful feedback, and continuous encouragement. Her academic guidance and commitment to excellence played a pivotal role in shaping the direction, execution, and overall quality of this work. We are truly grateful for her patience, availability, and constructive input during every phase of the project.

We also extend our sincere appreciation to the esteemed faculty members and staff of the Faculty of Artificial Intelligence for providing us with the necessary resources, knowledge, and technical support. The academic environment and collaborative spirit within the faculty have greatly contributed to our learning and professional development.

Special thanks are due to the creators and maintainers of the biomedical knowledge bases and tools utilized in our work, which formed the foundation of our data-driven approach.

We are also grateful to our colleagues for their valuable insights, encouragement, and assistance during the project development. Their support enriched our experience and strengthened our teamwork.

Finally, we wish to thank our families for their continuous motivation, patience, and unwavering belief in our abilities. Their encouragement was a source of strength throughout this academic journey.

To everyone who contributed to the success of this project in any way, we offer our sincere gratitude and appreciation.

Abstract

Interpretable Drug Repurposing Platform Based on Graph Neural Networks

The high cost, prolonged timelines, and frequent failures of traditional drug discovery have prompted significant interest in drug repurposing—identifying new indications for existing drugs. This book presents the development of an **interpretable drug repurposing platform based on graph neural networks (GNNs)**, designed to enhance both predictive accuracy and transparency in biomedical decision-making. The proposed system, **DRGNN (Drug-Relation Graph Neural Network)**, leverages a heterogeneous biomedical knowledge graph constructed from multimodal sources, including PrimeKG, DrugBank, and OMIM, to represent the complex landscape of drug-disease-gene interactions.

At the core of the platform lies a multi-relational GNN capable of learning structured embeddings across diverse biomedical entities. A novel meta-path-based explainability module provides interpretable rationales by surfacing semantic paths (e.g., drug → gene → pathway → disease), enabling researchers to trace and validate model predictions. The platform also includes an integrated API layer and an interactive user interface to support real-time prediction queries, explanation visualization, and hypothesis generation.

Experimental evaluations across various therapeutic areas demonstrate DRGNN's competitive performance, particularly in autoimmune, oncological, and neurological domains, with AUPRC values exceeding 0.94 in top-performing categories. Furthermore, comparisons between indication and contraindication prediction tasks reveal clinically meaningful performance gaps, emphasizing the platform's strength in therapeutic inference.

Limitations such as data incompleteness, absence of patient-specific modeling, and the need for experimental validation are critically discussed. Future extensions are proposed to incorporate temporal reasoning, multi-modal data (e.g., omics, literature), and personalized explainability mechanisms.

Overall, this work contributes a robust, interpretable, and scalable AI framework for drug repurposing and sets a foundation for deploying graph-based reasoning systems in real-world biomedical research and translational medicine.

Chapter 1: Introduction

1.1 Overview of Drug Repurposing

The conventional drug discovery process is notoriously time-consuming, expensive, and high-risk. On average, it takes 10–15 years and costs over $2.6 billion to bring a new drug to market, with a high rate of clinical trial failure due to unforeseen side effects or lack of efficacy. Given these limitations, **drug repurposing**, also referred to as drug repositioning, has emerged as a powerful alternative strategy in pharmaceutical research and development.

Drug repurposing involves identifying new therapeutic applications for drugs that are already approved for other diseases or have passed several phases of clinical testing. This strategy reduces both the time and financial burden by leveraging existing pharmacokinetic, pharmacodynamic, and safety data. It has gained significant attention during public health emergencies, such as the COVID-19 pandemic, where the urgency of finding effective treatments necessitated faster and more reliable approaches.

With the rapid expansion of biomedical databases, including gene-disease-drug networks, protein interaction maps, and clinical trial results, computational techniques are increasingly essential for systematically uncovering new drug-disease relationships. Network-based models and machine learning algorithms, particularly those operating on heterogeneous data, are at the forefront of this effort.

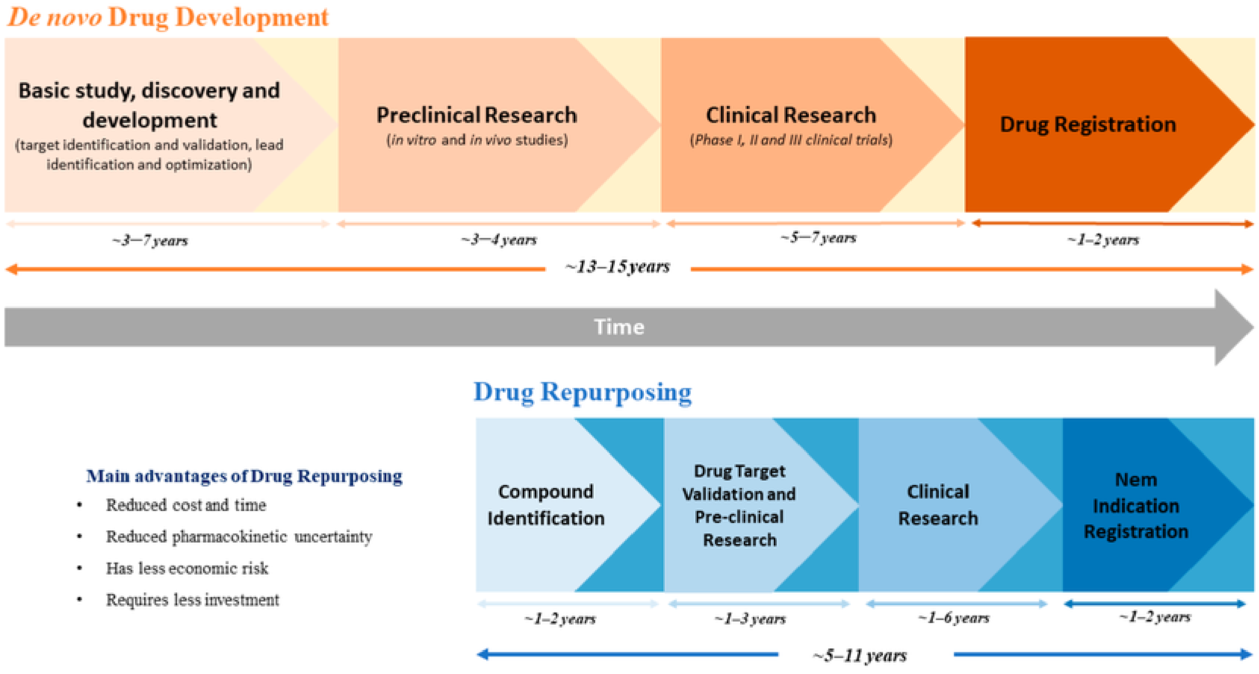

De novo Drug Development
Basic study, discovery and development
(target identification and validation, lead identification and optimization)
Preclinical Research
(in vitro and in vivo studies)
Clinical Research
(Phase I, II and III clinical trials)
Drug Registration
~3–7 years
~3–4 years
~5–7 years
~1–2 years
~13–15 years
Time
Drug Repurposing
Main advantages of Drug Repurposing
• Reduced cost and time
• Reduced pharmacokinetic uncertainty
• Has less economic risk
• Requires less investment
Compound Identification
Drug Target Validation and Pre-clinical Research
Clinical Research
Nem Indication Registration
~1–2 years
~1–3 years
~1–6 years
~1–2 years
~5–11 years

Figure 1 Comparative timelines and cost between traditional drug discovery and drug repurposing approaches.

Table 1 Selected examples of repurposed drugs

Drug Name	Original Indication	Repurposed Indication	Year Repurposed	Source/Reference
Sildenafil	Hypertension	Erectile dysfunction	1998	FDA Archives
Thalidomide	Morning sickness	Multiple myeloma	2006	WHO Report
Remdesivir	Ebola	COVID-19	2020	ClinicalTrials.gov

1.2 Importance of Explainable AI in Drug Discovery

Artificial Intelligence (AI), particularly deep learning, has significantly advanced the capability to model complex biomedical relationships. Techniques such as convolutional neural networks, recurrent neural networks, and graph neural networks (GNNs) have been applied to predict molecular properties, drug-target interactions, and therapeutic effects. However, many of these models function as "black boxes" — producing predictions without offering insight into how those results were generated.

In drug discovery, this lack of interpretability is not merely a technical inconvenience; it poses serious limitations. Researchers and clinicians require **transparency and trust** in AI-generated recommendations, especially when they influence real-world treatment decisions. The growing field of **Explainable Artificial Intelligence (XAI)** addresses this issue by designing models and frameworks that make their reasoning processes understandable and verifiable.

For drug repurposing specifically, XAI techniques can help trace why a model predicts a specific drug-disease association. Attention mechanisms, graph path analysis, feature attribution techniques (e.g., SHAP, LIME), and explainable subgraph generation enable models to provide **human-interpretable rationales**. These mechanisms not only build trust but also help generate new biological hypotheses by highlighting relevant pathways and interactions.

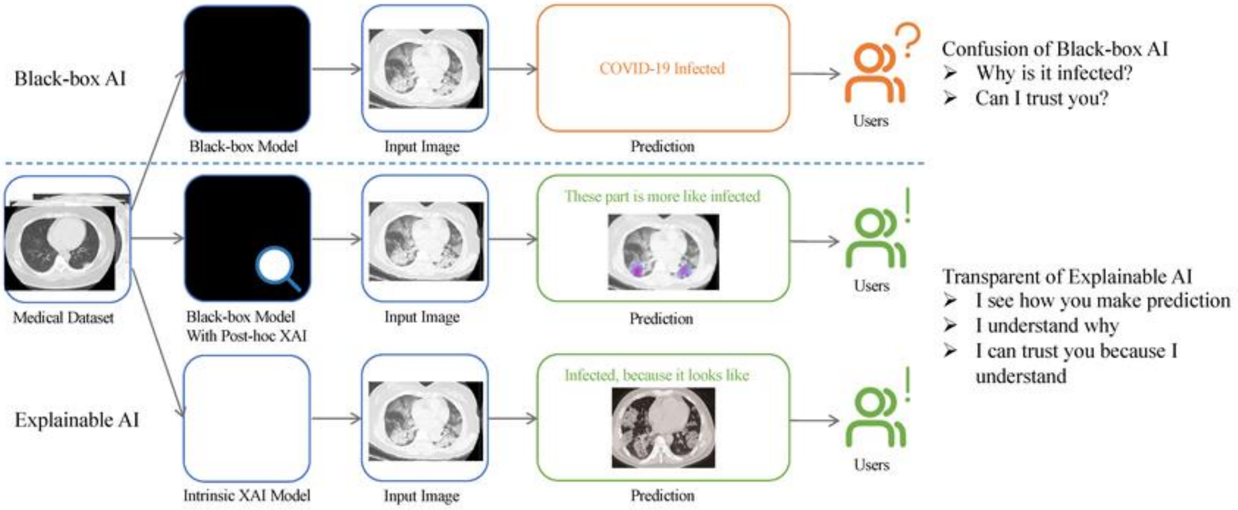


Figure 2 Conceptual comparison between black-box AI models and explainable AI in biomedical applications.

Moreover, regulatory agencies such as the FDA and EMA increasingly emphasize model interpretability in AI-assisted drug development pipelines. XAI also facilitates collaborative efforts between computational scientists and domain experts, enhancing both reproducibility and clinical relevance.

1.3 Objectives

This project proposes the development of **DRGNN (Drug-Relation Graph Neural Network)** — a computational framework that combines graph neural networks and explainability tools to predict potential drug-disease associations from heterogeneous biomedical knowledge graphs.

The primary aim is to create a **scalable, interpretable, and biologically grounded system** that can aid researchers in discovering repurposing candidates while offering insights into the rationale behind predictions.

Specific Objectives:

1. **To construct a comprehensive biomedical knowledge graph** integrating entities such as drugs, diseases, genes, pathways, and symptoms from structured data sources like PrimeKG.
2. **To develop and implement a multi-relational graph neural network (GNN)** architecture capable of learning entity and relation embeddings that capture semantic and structural dependencies.
3. **To incorporate attention-based mechanisms and path explanation techniques** to identify key contributors (nodes, edges, subgraphs) involved in the model's decision-making process.
4. **To provide visual and textual interpretability tools** that allow users to trace drug-disease predictions through meaningful biomedical pathways.

5. **To evaluate the performance of the DRGNN model** using standard metrics (e.g., ROC-AUC, Precision@K) and benchmark datasets, and validate it through known drug-disease associations and case studies.

6. **To explore practical applications** of the model in real-world drug discovery settings, such as identifying candidates for rare or orphan diseases.

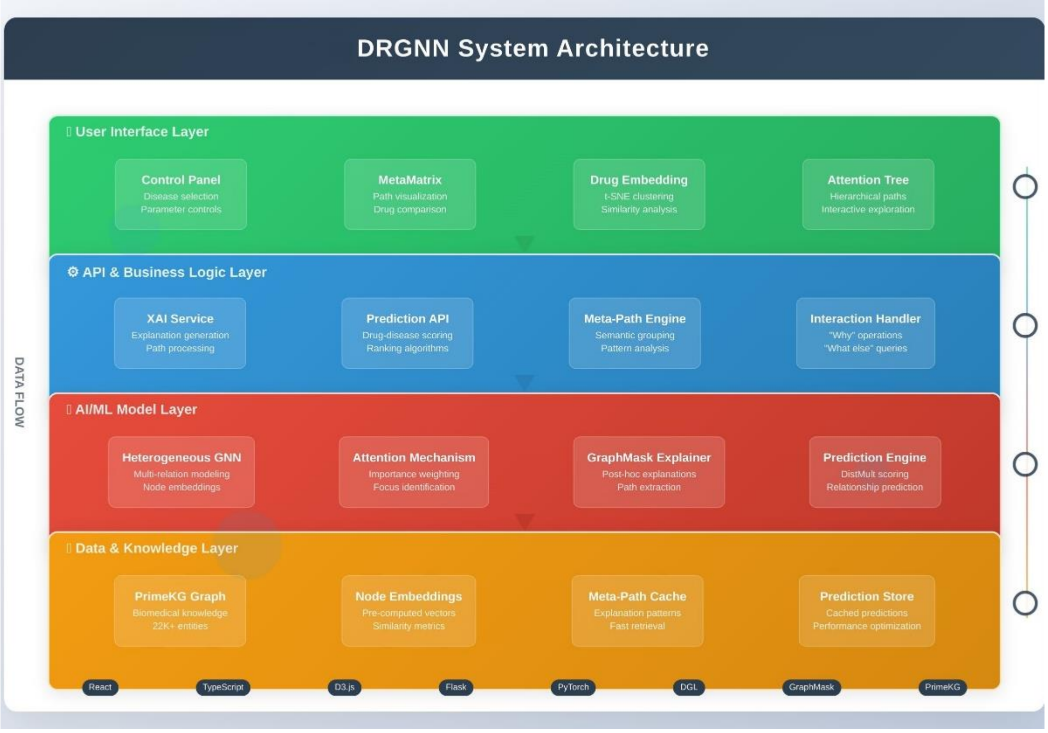
7. 

Figure 3 DRGNN Architecture Overview

Chapter 2: Background

2.1 Background

The integration of artificial intelligence (AI) into biomedical research has created transformative opportunities for addressing long-standing challenges in drug discovery, diagnostics, and personalized medicine. With the advent of high-throughput data generation techniques and the expansion of biomedical knowledge graphs, researchers are now able to model the complexity of human biology at an unprecedented scale. However, the biomedical domain brings unique challenges that require tailored computational approaches — including issues of data heterogeneity, uncertainty, and the necessity for interpretability.

Three pivotal pillars support the foundation of this project: **drug repurposing**, which seeks to reapply existing drugs for new indications; **graph neural networks (GNNs)**, which are capable of modeling structured biomedical data; and **explainable AI (XAI)**, which provides transparency into algorithmic decision-making. This chapter delves into the core concepts, technologies, and challenges relevant to each of these domains, laying the groundwork for the development of the proposed DRGNN framework.

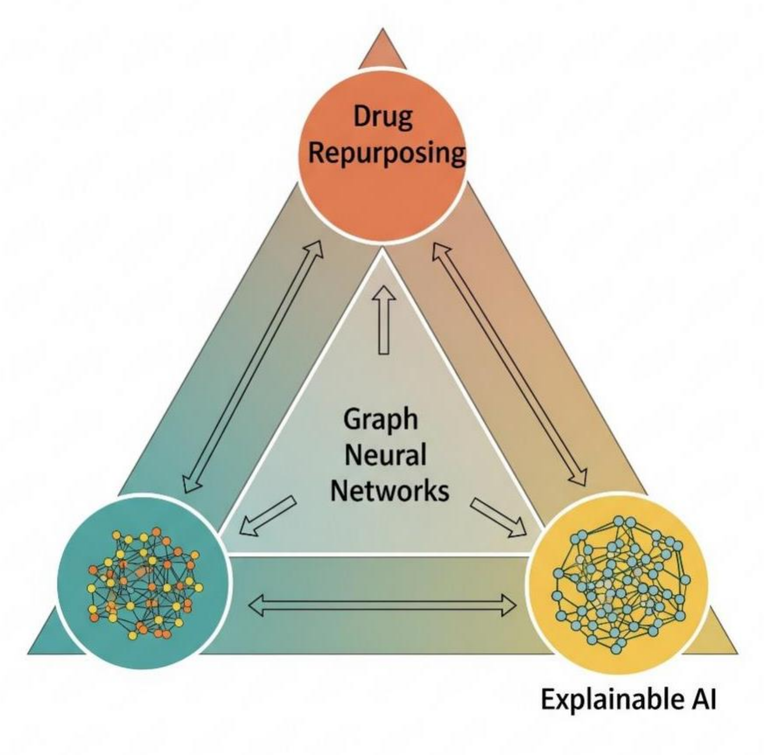


Figure 4 Interconnection between the three pillars: Drug Repurposing, Graph Neural Networks, and Explainable AI.

2.2 Drug Repurposing Challenges

Drug repurposing has emerged as a strategic solution to the limitations of conventional drug development pipelines. However, while the appeal of reduced development costs and shorter time-to-market is strong, implementing successful computational repurposing frameworks remains a non-trivial task.

A. Data Heterogeneity and Integration Issues

Biomedical data is inherently diverse — it includes chemical structures, gene expressions, protein-protein interactions, clinical trial data, and disease ontologies. These datasets vary in their scale, format, and semantic schema. For example, DrugBank focuses on chemical and pharmacological properties, while OMIM (Online Mendelian Inheritance in Man) emphasizes genetic links. Integrating such sources into a unified representation is a major bottleneck.

- **Semantic alignment** between ontologies (e.g., MeSH vs. ICD-10)
- **Cross-database entity resolution** (e.g., same drug listed under different IDs)
- **Missing data** and conflicting annotations across datasets

B. Complexity of Biological Interactions

Diseases are rarely the result of single gene mutations or isolated pathways. They are complex phenomena involving interactions between genetic, epigenetic, environmental, and lifestyle factors. Thus, capturing drug-disease relationships requires multi-level reasoning across interconnected molecular systems.

- A drug may act on multiple targets.
- Targets may be part of multiple pathways.
- Side effects may arise from indirect interactions.

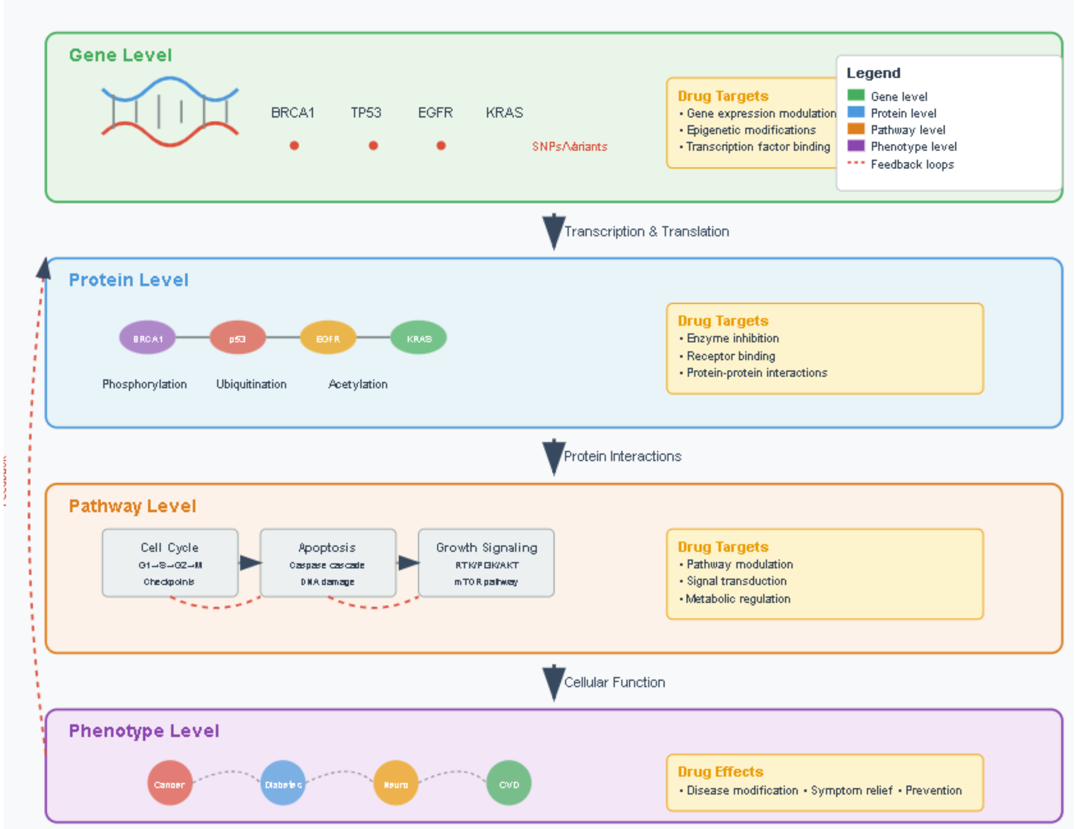


Figure 5 Multi-scale biological complexity in drug-disease interaction networks

C. Lack of Predictive Interpretability

Many early computational repurposing models — such as collaborative filtering, matrix factorization, and neural embeddings — provided high accuracy but were fundamentally opaque. They often failed to explain *why* a drug might work, leaving clinicians and pharmacologists hesitant to act on these results.

- No biological evidence trail
- No path-level reasoning
- Low trust in real-world deployment

D. Validation Constraints

Even if a model accurately predicts a drug-disease association, **experimental validation remains a resource-heavy task**, requiring in vitro testing, animal models, and potentially clinical trials. Without confidence in the computational output, researchers may deprioritize novel findings.

Table 2 Challenges and Corresponding Requirements in Computational Drug Repurposing

Challenge	**Description**	**Required Solution**
Data heterogeneity	Incompatible, unaligned data sources	Knowledge graph integration, ontologies
Biological complexity	Non-linear, multi-target drug actions	Multi-relational models
Black-box models	Lack of transparency in predictions	Explainable AI approaches
Validation cost	High cost of wet-lab confirmation	Confidence scores and traceable logic

2.3 Graph Neural Networks in Biomedicine

Graph-based representations are naturally suited to the biomedical domain, where relationships between entities — such as drug-gene, disease-symptom, or gene-pathway — form rich networks. **Graph Neural Networks (GNNs)** have revolutionized the ability to extract knowledge from these structures.

A. Why Graphs Matter Traditional machine learning methods rely on feature vectors, which flatten entities into static attributes. However, biomedical data is relational — drugs are connected to proteins, which are connected to diseases. Graphs preserve this relational structure.

- **Nodes** represent entities (e.g., drugs, diseases, genes)
- **Edges** represent biological relationships (e.g., inhibits, causes, interacts)
- **Edge types** allow for multi-relational modeling

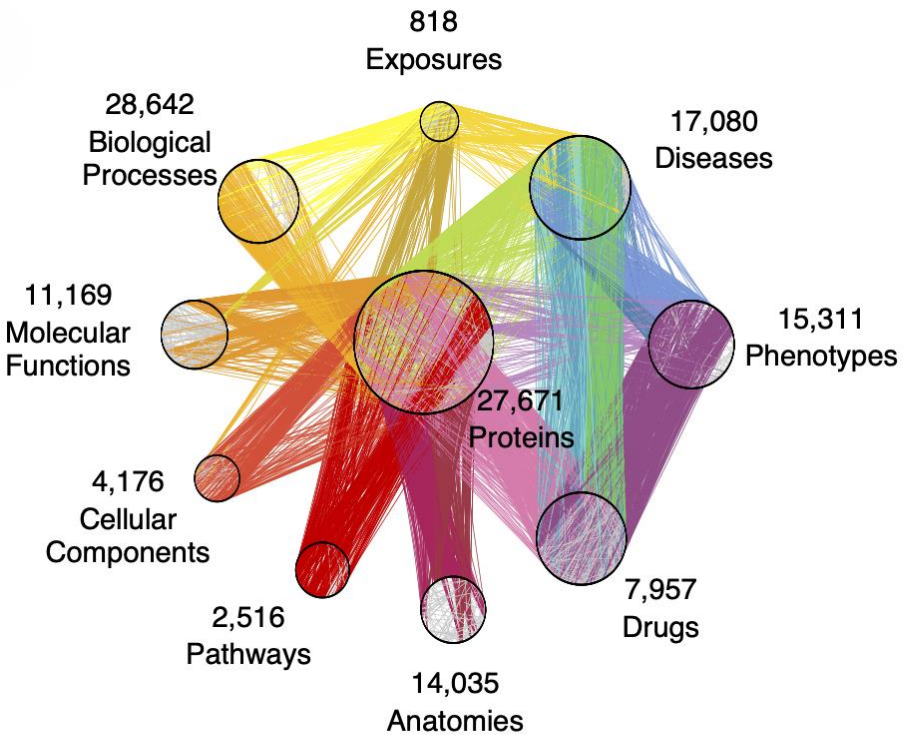


Figure 6 Sample biomedical knowledge graph

B. Fundamentals of GNNs

GNNs generalize neural networks to non-Euclidean spaces. Instead of applying convolutions over grids (as in CNNs), GNNs aggregate information from a node's neighbors.

- **Message Passing**: Each node updates its embedding by aggregating messages from its neighbors.
- **Multi-layer propagation**: Deeper layers allow nodes to access information from a broader neighborhood.
- **Attention Mechanisms**: Weights can be learned to focus on more important connections.

C. Biomedical Applications

1. **Drug-Target Prediction**
   Use molecular graphs and interaction networks to predict binding.
2. **Gene-Disease Inference**
   Use co-expression and mutation data to associate genes with phenotypes.
3. **Drug Repurposing**
   Learn embeddings from a heterogeneous graph to infer new drug-disease edges.

D. Popular GNN Variants

Model	**Key Feature**	**Relevance to Biomedicine**
GCN	Spectral convolution	Baseline for semi-supervised tasks
GAT	Attention over neighbors	Identifies key biological influencers
R-GCN	Handles multiple edge types	Suitable for multi-relational graphs
HGT	Heterogeneous graph transformer	Scales to graphs with diverse node types

2.4 Explainable AI in Healthcare

The adoption of AI in healthcare is contingent not only on performance but also on **accountability and transparency**. When decisions affect lives, stakeholders must understand *why* a model made a certain prediction. This is the central goal of **Explainable AI (XAI)**.

A. Importance of Explainability

1. **Clinical Trust**
   Physicians must be able to verify AI suggestions through understandable logic.
2. **Regulatory Approval**
   Agencies like the FDA require audit trails in AI-assisted decision-making.
3. **Scientific Discovery**
   XAI can help uncover new biological insights, e.g., highlighting pathways not previously associated with a disease.
4. **Debugging and Bias Detection**
   Models that are transparent are easier to debug and refine.

B. Methods of Explainability

- **Post-Hoc Feature Attribution**
  Tools like SHAP and LIME explain a model's output by analyzing input perturbations.
- **Attention Mechanisms**
  GNNs equipped with attention layers can reveal which node relationships were most influential.
- **Subgraph Extraction**
  Path-based methods trace the shortest or highest-scoring connection between a drug and a disease.

- **Contrastive and Counterfactual Explanations**
  “What-if” analysis: What changes to the input would change the model’s decision?

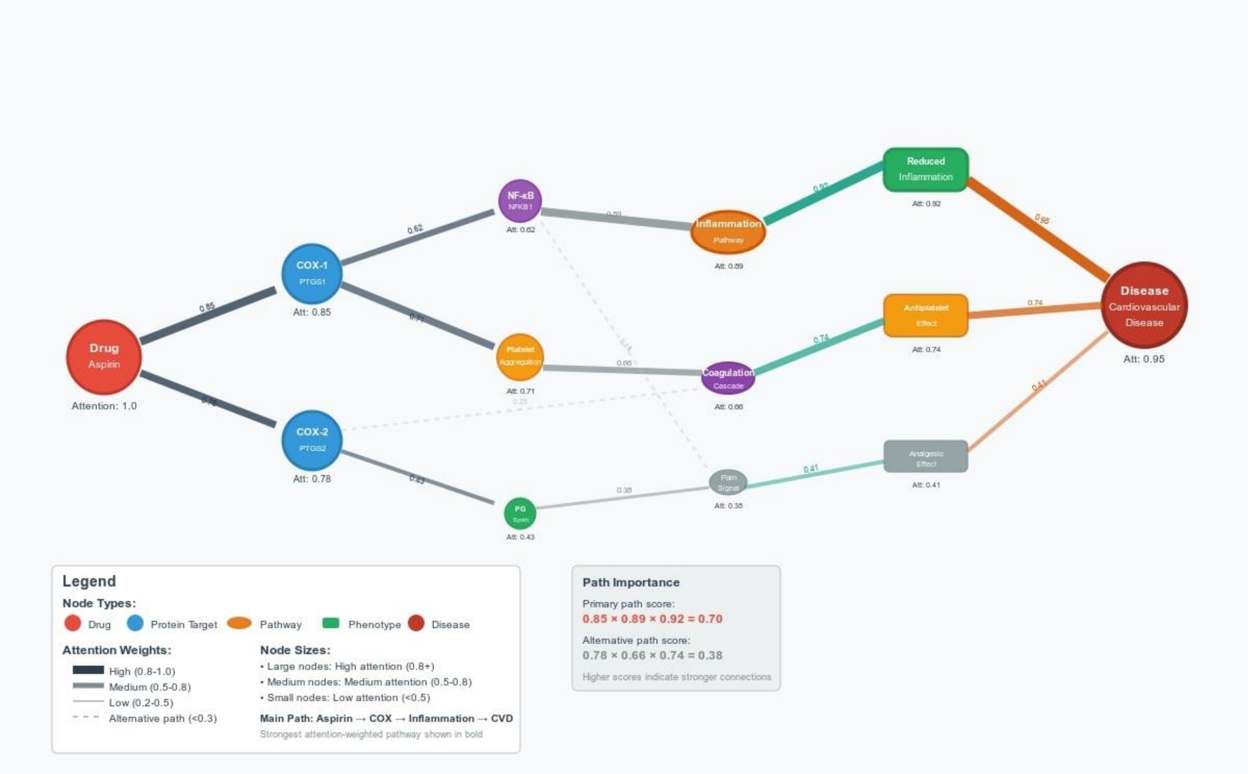
- 

Figure 7 Example of an attention-weighted path from a drug to a disease

C. Challenges and Considerations

- **Faithfulness**: Does the explanation reflect the true reasoning of the model?
- **Complexity vs Simplicity**: More informative explanations are often harder to interpret.
- **User Audience**: Biologists, clinicians, and AI engineers have different interpretability needs.

Table 3 Categories of XAI Techniques for GNNs in Healthcare

Method Type	**Technique**	**Output Format**	**Use Case**
Post-hoc analysis	SHAP, LIME	Feature importance	Gene expression relevance
Graph-based attention	GAT, GCN+Attention	Weighted neighbors	Disease path tracing
Subgraph extraction	GNNExplainer, PGExplainer	Critical subgraph	Drug-disease linkage
Counterfactual logic	Edge/node changes	Rule-based changes	Model debugging and trust

Chapter 3: Methodology & Implementation

i. Dataset and Model Pipeline

3.1 Knowledge Graph Construction and Analysis

1.1 Knowledge Graph Construction Pipeline

The construction of biomedical knowledge graphs requires a systematic approach encompassing multiple interconnected phases. Our methodology incorporates three fundamental components:

- **Source Data Integration:** Comprehensive aggregation of heterogeneous biomedical datasets including PrimeKG, DrugBank, and OMIM repositories, with emphasis on data quality assessment and validation protocols.
- **Knowledge Graph Structuring:** Implementation of sophisticated entity-relation modeling frameworks, ontology mapping procedures, and multi-type node architectural design to ensure semantic consistency.
- **Graph Preprocessing:** Application of advanced node and edge filtering algorithms, identifier normalization processes, edge-type encoding methodologies, and strategic graph sampling techniques.

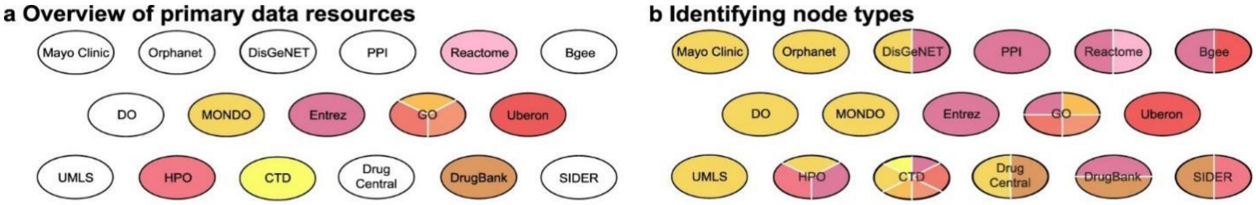


Figure 8 Main Knowledge Graph Sources

1.2. Dataset Architecture and Characteristics

1.2.1 PrimeKG: A Multimodal Biomedical Knowledge Graph

PrimeKG represents a paradigmatic advancement in biomedical knowledge representation, hosted on the Harvard Dataverse platform. This sophisticated multimodal knowledge graph serves as a transformative resource for precision medicine applications, addressing the fragmented nature of contemporary biomedical data landscapes through comprehensive integration strategies.

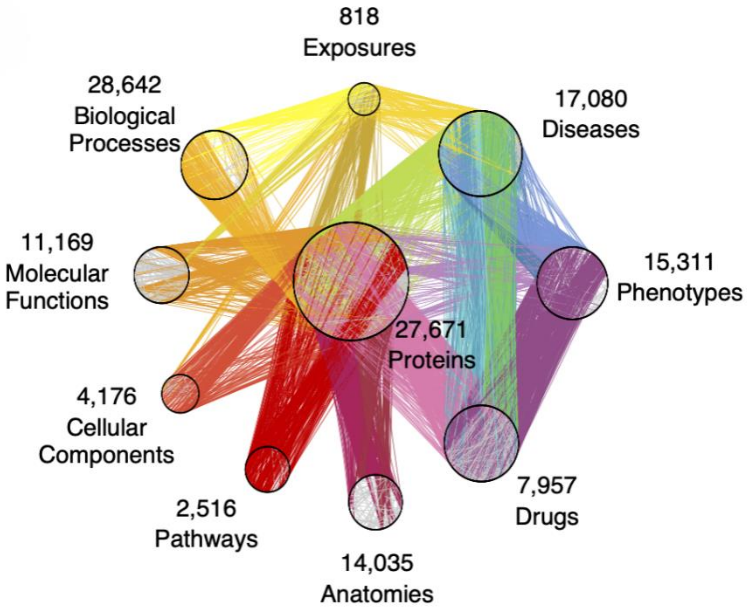


Figure 9 PrimeKG Structure

Key Architectural Features

Scale and Complexity: PrimeKG encompasses 129,375 distinct entities distributed across 10 specialized node types, interconnected through 8,100,498 relationships spanning 30 unique relationship categories. This extensive network facilitates detailed exploration of biomedical entities and their intricate interrelations.

1.2.1.1 Comprehensive Node Distribution Analysis

Table 4 Comprehensive Node Type Distribution in PrimeKG

Table 4: Comprehensive Node Type Distribution in PrimeKG			
Node Type	**Entity Count**	**Percentage (%)**	**Primary Data Source**
Biological Process	**28,642**	**21.2.1**	**GO**
Gene/Protein	**27,671**	**21.4**	**NCBI**
Disease	**17,080**	**13.2**	**MONDO**
Effect/Phenotype	**15,311**	**11.8**	**HPO**
Anatomy	**14,035**	**10.8**	**UBERON**
Molecular Function	**11,169**	**8.6**	**GO**
Drug	**7,957**	**6.2**	**DrugBank**
Cellular Component	**4,176**	**3.2**	**GO**
Pathway	**2,516**	**1.9**	**REACTOME**
Exposure	**818**	**0.6**	**CTD**

1.2.1.2 Disease Representation Excellence

The knowledge graph demonstrates exceptional disease representation capabilities, encompassing over 17,000 distinct disease entities. This comprehensive coverage spans the entire spectrum of human pathology, from prevalent conditions affecting large populations to rare disorders with limited case studies. Each disease entity is enriched with detailed clinical attributes, including symptomatic profiles, molecular interaction networks, and therapeutic association mappings.

1.2.1.3 Multimodal Data Integration Framework

PrimeKG's distinctive multimodal architecture represents a significant advancement in biomedical knowledge representation. The framework seamlessly integrates structured quantitative data with unstructured textual information, creating a comprehensive analytical environment. Structured components include gene expression levels, pathway hierarchical relationships, and drug-protein interaction matrices, while unstructured elements encompass clinical disease summaries, pharmaceutical mechanism descriptions, and therapeutic indication profiles.

3. Data Source Integration and Harmonization

3.1 Primary Data Resources

PrimeKG leverages 20 high-quality biomedical resources, ensuring comprehensive domain coverage across multiple biological and clinical data hierarchies. The integration strategy encompasses genomic and proteomic databases, pharmaceutical repositories, phenotypic resources, and clinical knowledge sources.

Strategic Data Source Categories

- **Genomic and Proteomic Resources:** DisGeNET for curated gene-disease associations, Bgee for anatomically-mapped gene expression patterns
- **Pharmaceutical Databases:** DrugBank for comprehensive drug property profiles, DrugCentral for extensive drug-disease relationship mappings
- **Phenotypic Resources:** Human Phenotype Ontology (HPO) for detailed phenotype-disease relationships, UBERON for multi-species anatomical structure ontologies
- **Clinical Knowledge Sources:** MONDO Disease Ontology for harmonized disease definitions, Orphanet for rare disease comprehensive coverage

3.2 Data Harmonization Methodology

The construction of PrimeKG necessitated sophisticated harmonization procedures to resolve inconsistencies across diverse data sources. The harmonization framework incorporated three critical components: ontology mapping protocols, entity resolution algorithms, and standardized format conversion procedures.

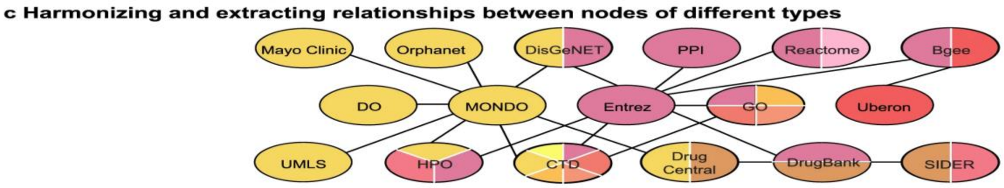


Figure 10 Harmonization Procedures

4. Visualization Framework and Analysis

4.1 Node Distribution Visualization

Comprehensive visualization techniques were employed to elucidate the structural characteristics and relational patterns within the knowledge graph. The visualization framework encompasses multiple analytical perspectives, including node distribution analysis, edge relationship mapping, and network topology exploration.

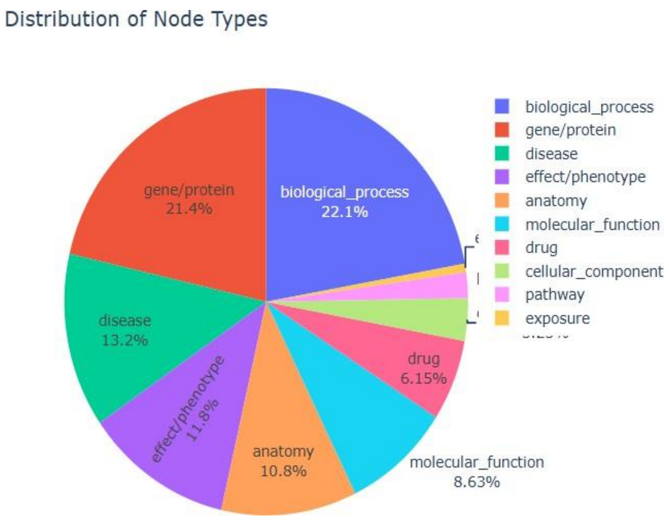


Figure 11 Node Distribution

4.2 Relationship Distribution Analysis

Table 5 Top Relationship Types and Occurrence Frequencies

Table 5: Top Relationship Types and Occurrence Frequencies

Relationship Type	**Display Relationship**	**Occurrence Count**	**Percentage (%)**
anatomy_protein_present	Expression Present	3,036,406	37.5
drug_drug	Synergistic Interaction	642,150	7.9
protein_protein	Protein-Protein Interaction	300,634	3.7
disease_phenotype_pos	Phenotype Present	289,610	3.6
cellcomp_protein	Protein Interaction	262,128	3.2

4.3 Edges Visualization

The pie chart for the edges dataset visualization helps to clarify the emphasis on established drug-disease connections. Recognizing the composition of the edges is crucial for interpreting the model's predictions and understanding how various therapeutic pathways are interconnected.

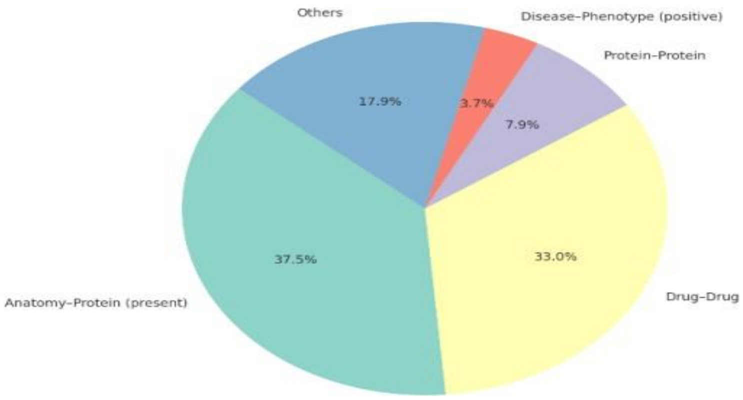


Figure 12: pie chart

3.2 DRGNN Model Core

3.2.1 Task Definition:

We define a heterogeneous knowledge graph (KG) as $G = (\mathcal{V}, \mathcal{E}, \mathcal{T}_R)$, where:

- $\mathcal{V}$: is the set of biomedical entities (nodes), each typed according to a node type set T.
- $\mathcal{T}_V$: The set of edges, where each edge $e_{i,j} = (i, r, j)$∈ E connects a source node i (head) to a target node j (tail) through a defined relationship type r.
- $\mathcal{T}_V$: The set of relationship types, describing the nature of connections between nodes in the graph.

Each node $i \in \mathcal{V}$ is associated with an initial feature embedding, denoted as$\mathbf{h}_i^{(0)}$.

The task is to predict the likelihood of a given drug j being either indicated or contraindicated for a specific disease i. This process leverages the structural and semantic information embedded in the KG. The model integrates graph-based inductive priors, enhancing its reasoning capabilities to generate meaningful hypotheses and efficiently identify potential drug candidates.

3.2.2 Model Initialization and Configuration

the **DRGNN** model is initialized with specific hyperparameters. These include:

- **Hidden Dimensions** (n_{hid}): 100
- **Input and Output Dimensions** ($n_{\text{inp}}, n_{\text{out}}$): 100
- **Prototype Activation** (proto_num): 3
- **Similarity Measurement Method**: "all_nodes_profile"
- **Random Walks for Subgraph Exploration** (num_walks): 200
- **Walk Mode**: "bit"
- **Path Length**:

The model utilizes pretraining and fine-tuning procedures:

1. **Pretraining**:
   - Number of epochs: 2
   - Learning rate: 1×10^{-3}
   - Batch size: 512
   - Logging interval: every 20 iterations.
2. **Fine-Tuning**:
   - Number of epochs: 500
   - Learning rate: 5×10^{-4}

The implementation was carried out using the **DRGNN framework**, with GPU support (NVIDIA GTX 1660 Ti) for accelerated computation.

3.2.3 Experimental Setup

Dataset Splitting for Evaluation

Robust evaluation of model performance required the creation of specialized data splits:

1. **Disease Group Splits**:
   - Disease categories were selected to mimic underrepresented or uncharacterized diseases. Drug-disease relationships for these categories were excluded during training to simulate real-world challenges.
   - Additional data from the 1-hop neighborhood of these diseases was removed to reflect limited biological knowledge.
2. **Systematic Splits**:
   - To test the model's capability to predict relationships for diseases without prior treatments, drug-disease edges for novel diseases were removed during training, ensuring robust evaluation of unseen data.

3.2.4 DRGNN Core Procedure:

DRGNN is a deep-learning model designed to make mechanistic predictions in drug discovery, focusing on molecular networks disrupted by disease and targeted by therapeutics. The DRGNN model consists of four main components:

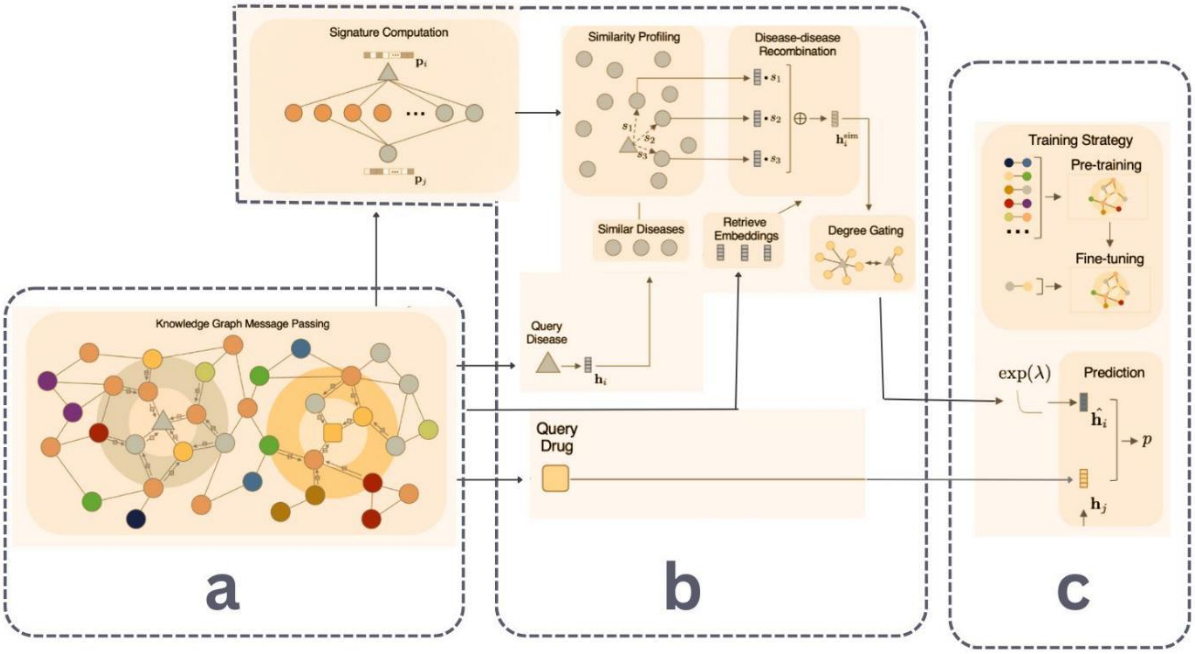


Figure 13: DRGNN Model.

- a) A heterogeneous graph neural network (GNN)-based encoder that generates biologically meaningful network representations for each biomedical entity.

- b) A disease similarity-based metric learning decoder that leverages auxiliary data to enhance the representation of diseases with insufficient molecular characterization.

- c) A pretraining phase using stochastic methods across all relationships, followed by a drug-disease-centric fine-tuning strategy for full-graph optimization.

a)Heterogeneous GNN Encoder

The primary goal was to develop a general encoder for biomedical knowledge graphs (KGs) by learning a numerical vector (embedding) for each node. This embedding encapsulates biomedical knowledge derived from the relational structures of its neighboring nodes. The process involves refining the initial embeddings through a series of graph-based, nonlinear transformations, optimized iteratively to minimize prediction errors. The system ultimately converges to an optimized set of node embeddings

Step 1: Initializing Latent Representations

The input embedding X_i for each node i is initialized using Xavier uniform initialization. The process proceeds through multiple layers l of message passing.

Step 2: Propagating Relationship-Specific Messages

For each relationship type, we compute the transformation of node embeddings from the previous layer. Initially, the first layer transformation is $X_1 = X_i$. Each layer applies a relationship-specific weight matrix W_r on the previous layer's embeddings.

$$\mathbf{m}_{r,i}^{(l)} = W_{r,M}^{(l)}\mathbf{h}_i^{(l-1)}$$

Step 3: Aggregating Neighborhood Information

For each node i, incoming messages from its neighbors are aggregated, considering each relationship type r. The aggregation step involves averaging the embeddings of neighboring nodes:

$$m_{r,i}^{(l)} = W_{r,M}^{(l)}h_i^{(l-1)}$$

where $N_r(i)$ represents the neighboring nodes for relationship type r, and h_j denotes the embedding of node j.

Step 4: Updating Latent Representations

We combine the embedding from the previous layer with the aggregated neighborhood messages to update the node's embedding:

$$h_i^{(l)} = h_i^{(l-1)} + \sum_{r \in \mathcal{T}_R} \tilde{m}_{r,i}^{(l)}$$

After L layers of propagation, the final encoded node embeddings h_i are obtained for each node i

b) Embedding-Based Disease Similarity Enrichment

Step 1: Disease Signature Vectors

The objective of this module is to generate a signature vector p_i for each disease i. Given the insufficiency of disease representations produced solely by GNNs, these representations are not ideal for direct similarity computations. Instead, we employed graph-theoretical methods to calculate disease similarities. We generated a vector that encapsulates the local neighborhoods surrounding a disease. For disease i, the signature vector is formally defined as follows:

$$p_i = \left[p_1 \cdots p_{|\mathcal{V}_P|}\ \mathrm{ep}_1 \cdots \mathrm{ep}_{|\mathcal{V}_{EP}|}\ \mathrm{ex}_1 \cdots \mathrm{ex}_{|\mathcal{V}_{EX}|}\ d_1 \cdots d_{|\mathcal{V}_D|}\right]$$

where:

$$p_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathcal{P}} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{ep}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathcal{EP}} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{ex}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathcal{EX}} \\ 0 & \text{otherwise} \end{cases}$$

$$d_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathcal{D}} \\ 0 & \text{otherwise} \end{cases}$$

where $N_i^{\mathcal{P}}, N_i^{\mathcal{EP}}, N_i^{\mathcal{EX}}, N_i^{\mathcal{D}}$ represent the sets of gene/protein, effect/phenotype, exposure, and disease nodes in the one-hop neighborhood of disease i..
The similarity between diseases is computed using the dot product:

$$\mathrm{sim}(i, j) = \mathbf{p}_i \cdot \mathbf{p}_j = |\mathcal{N}_i^{\mathcal{P}} \cap \mathcal{N}_j^{\mathcal{P}}| + |\mathcal{N}_i^{\mathcal{EP}} \cap \mathcal{N}_j^{\mathcal{EP}}| + |\mathcal{N}_i^{\mathcal{EX}} \cap \mathcal{N}_j^{\mathcal{EX}}| + |\mathcal{N}_i^{\mathcal{D}} \cap \mathcal{N}_j^{\mathcal{D}}|$$

Step 2: Disease Metric Learning

After identifying similar diseases, DRGNN generates embeddings that integrate these similarities into a unified representation. A weighted approach is used, where each disease is weighted according to its similarity to the queried disease:

$$h_i^{\mathrm{aug}} = \sum_{j \in \text{Top-k}(i)} \mathrm{Similarity}(i, j) \cdot h_j$$

Step 3: Gating Disease Embeddings

The original disease embedding h_i is updated with the disease-to-disease similarity embedding h_i^{sim} using a gating mechanism controlled by a scalar $c \in [0,1]$:

$$h_i^{\text{final}} = c \cdot h_i + (1 - c) \cdot h_i^{\text{sim}}$$

c)Predicting Drug Candidates

DRGNN uses both disease and drug embeddings to predict drug indications or contraindications for each disease-drug pair. The model assigns a trainable weight vector w_r for each relationship type. The interaction likelihood between a disease and a drug is predicted using the DistMult approach. The likelihood $p_{i,j,r}$ for a disease i, drug j, and relationship r is calculated as follows:

$$p_{i,j,r} = \frac{1}{1 + \exp\left(-\sum \left(h_i \times w_r \times h_j\right)\right)}$$

Embedding-Based Disease Similarity Search

Diseases vary widely in research coverage, with many rare diseases lacking sufficient molecular data. This scarcity often results in poor-quality embeddings for these diseases within the KG. DRGNN addresses this by enriching disease embeddings by leveraging similarities to other better-represented diseases in the graph.

The three-step procedure used to enhance disease embeddings is:

1. Constructing a disease signature vector that captures disease similarities.
2. Using an aggregation mechanism to combine embeddings of similar diseases into a comprehensive auxiliary embedding.
3. Applying a gating mechanism to control the influence between the original and auxiliary disease embeddings.

d) Training DRGNN

The objective of the training process is to predict the presence of a relationship between two entities within the KG. The positive sample dataset D_{pos} consists of all pairs (i, j) across different relationship types r, with labels $y_{i,r,j} = 1$. Negative samples D_{neg} are generated through a sampling process. The training loss is computed using binary cross-entropy:

$$L = -\sum_{i,r,j} y_{i,r,j} \log(p_{i,r,j}) + (1 - y_{i,r,j}) \log(1 - p_{i,r,j})$$

Pretraining DRGNN Model:

The pretraining phase involves training DRGNN on millions of biomedical entity pairs across all relationships. To ensure efficiency, stochastic mini-batching is used to process subsets of pairs at each step. During this phase, degree-adjusted disease augmentation is deactivated, and all relationship types are treated equally. The pretrained encoder weights are used to initialize the model for fine-tuning.

Fine-Tuning DRGNN Model:

After pretraining, the model is fine-tuned with a focus on drug–disease pairs. This phase sharpens the model's ability to predict drug-disease interactions while still utilizing other relationship types in the KG for indirect knowledge transfer. The fine-tuned model undergoes both pretraining and fine-tuning, with the best variant selected for evaluation on the test set.

ii. Meta-Path-Based Explainability

3.3 Meta-Path Definition and Significance

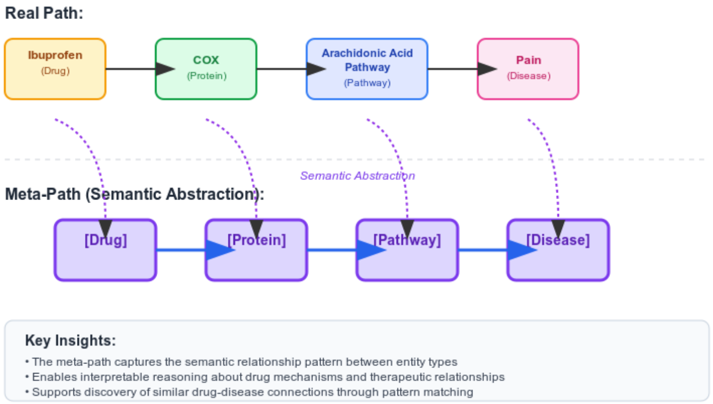
In heterogeneous biomedical knowledge graphs (KGs), a **meta-path** is defined as a sequence of node types and edge types that represent composite semantic relations. It generalizes individual graph paths by focusing on the *types* of entities and their interconnections rather than specific instances. Formally, a meta-path $\mathcal{P}$ can be described as:

$$\mathcal{P} : A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_n} A_{n+1} \qquad (1)$$

where A_i denotes entity types (e.g., Drug, Gene, Disease), and R_i are relation types (e.g., binds_to, causes, involved_in).

In biomedical domains, meta-paths encapsulate interpretable biological mechanisms. For instance, a drug may interact with a protein, which is part of a pathway associated with a disease—capturing this sequence provides insight into how the drug may affect disease phenotypes. Meta-paths are thus critical in **explainable AI (XAI)** frameworks, enabling domain experts to trace and validate predictions made by machine learning models such as GNNs.

In DRGNN, the explainer module explicitly generates multi-hop meta-paths as interpretive rationales for predictions. These provide clinicians with semantic justifications, enhancing model transparency and trustworthiness in critical biomedical applications.



Key Insights:

- The meta-path captures the semantic relationship pattern between entity types
- Enables interpretable reasoning about drug mechanisms and therapeutic relationships
- Supports discovery of similar drug-disease connections through pattern matching

Figure 14 Example of a Biomedical Meta-Path

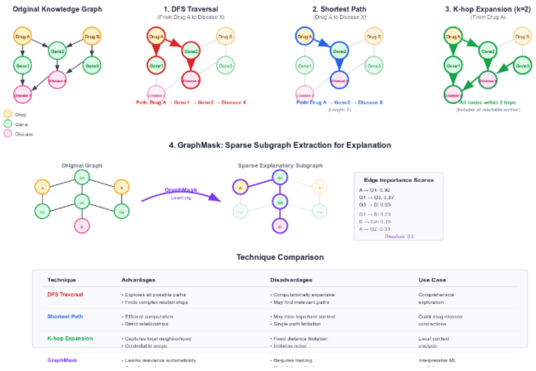
3.4 Path Finding Algorithms

To construct meta-paths from a biomedical KG, it is necessary to extract multi-hop paths between target entities, such as drug–disease pairs. The selection and design of **path-finding algorithms** directly influence the relevance, diversity, and efficiency of these meta-paths.

Several algorithmic strategies are employed:

- **Depth-First Search (DFS):**
  DFS systematically explores all reachable paths between nodes up to a predefined depth. While comprehensive, DFS is computationally expensive and may include low-relevance paths in dense subgraphs.
- **Shortest-Path Algorithms:**
  Dijkstra's and A* algorithms identify the shortest weighted paths between two nodes. Edge weights may be defined by prior knowledge, such as confidence scores or biological plausibility.
- **k-Hop Neighborhood Expansion:**
  This approach collects all nodes within a radius of kk hops from a source entity. It is particularly useful in identifying mechanistically related nodes around a disease or drug entity.
- **GraphMask and Sparse Subgraph Extraction:**
  GraphMask, used in DRGNN, applies a learned edge importance mask to isolate only the most predictive subgraphs. This results in sparse, interpretable multi-hop paths tailored to a model's prediction.

These techniques can be constrained by **meta-path schemas**, which filter results to biologically meaningful paths (e.g., [Drug] → [Gene] → [Disease]) and prevent semantically invalid or spurious connections.



Technique Comparison

Technique	Advantages	Disadvantages	Use Case
DFS Traversal	• Explores all possible paths • Finds complex relationships	• Computationally expensive • May find irrelevant paths	Comprehensive exploration
Shortest Path	• Efficient computation • Direct relationships	• May miss important context • Single path limitation	Quick drug-disease connections
K-hop Expansion	• Captures local neighborhood • Controllable scope	• Fixed distance limitation • Includes noise	Local context analysis
GraphMask	• Learns relevance automatically	• Requires training	Interpretable ML

Figure 15 Path Extraction Techniques in a Biomedical KG

3.5 Scoring and Ranking Methods

The extraction of candidate paths is only the first step. To generate interpretable and trustworthy explanations, meta-paths must be evaluated, scored, and ranked based on their **informativeness**, **specificity**, and **contribution** to the model's decision.

Key scoring and ranking methods include:

- **Path Frequency:**
  Measures how often a given meta-path schema occurs across the KG or in known drug-disease associations. Frequent meta-paths may indicate common therapeutic mechanisms.
- **Inverse Path Frequency (IPF):**
  Penalizes overly common paths. The intuition is analogous to inverse document frequency in information retrieval. Rare but significant paths receive higher relevance scores.
- **GNN Attention Weights:**
  In models like Heterogeneous Attention Networks (HAN), attention scores over meta-path types or edge types can directly reflect their predictive contribution.
- **Edge Importance via GraphMask:**
  In DRGNN, the GraphMask algorithm computes a continuous importance score for each edge. The cumulative score of a path is used as a proxy for its explanatory value. Paths composed of high-importance edges are considered more influential.
- **Meta-Path Aggregation and Grouping:**
  Paths with shared schemas are grouped under a single meta-path and ranked based on aggregate statistics (e.g., average importance score, frequency). This grouping enables clinicians to reason at a higher abstraction level and reduces cognitive load.

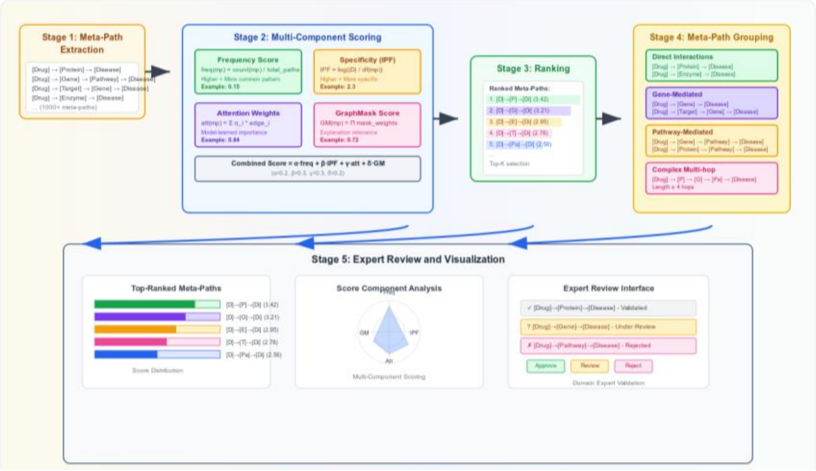
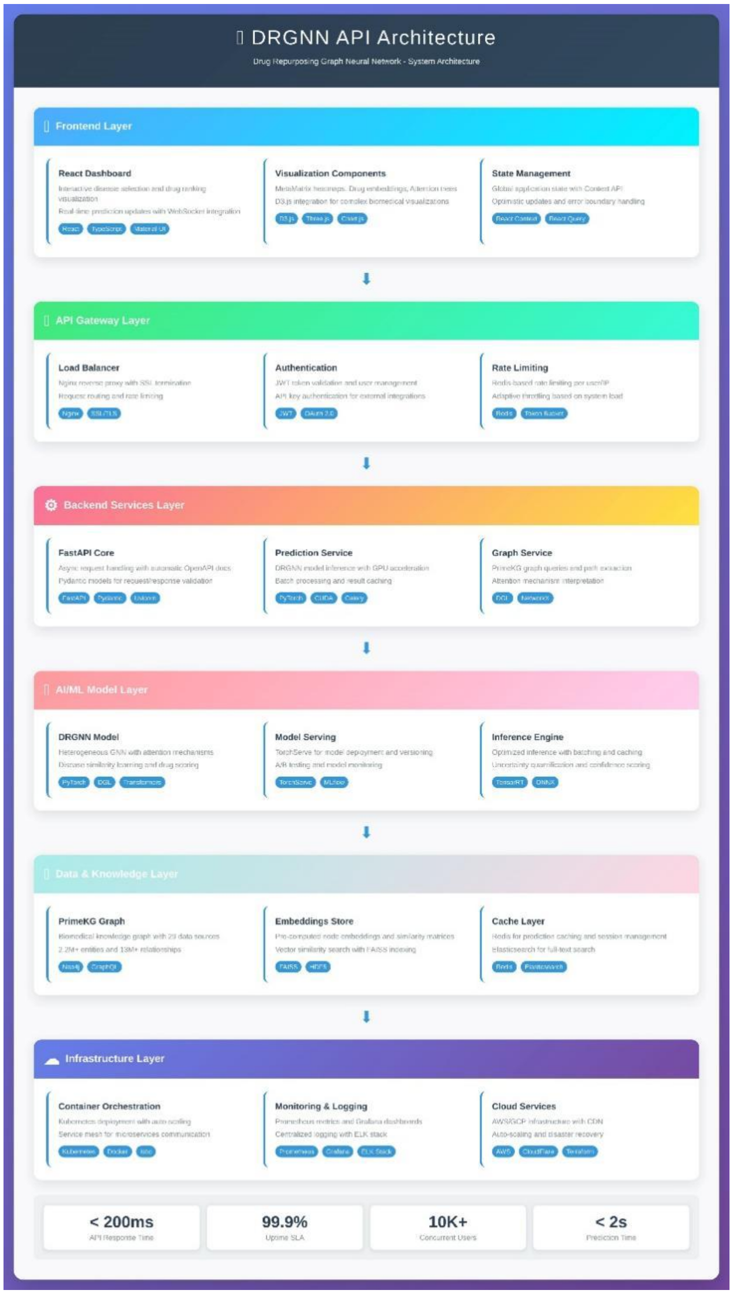


Figure 16 Meta-Path Scoring and Ranking Pipeline

iii. API and User Interface Design

3.6 API Implementation

The architecture of DRGNN is designed as a modular, layered system to support real-time, explainable drug repurposing analysis. Central to its operation is a full-stack API infrastructure that enables seamless communication between the model inference layer, the backend logic, and the user-facing interface.



DRGNN API Architecture

Drug Repurposing Graph Neural Network - System Architecture

Frontend Layer

React Dashboard
Interactive disease selection and drug ranking visualization
Real-time prediction updates with WebSocket integration

Visualization Components
MetaMatrix heatmaps, Drug embeddings, Attention trees
D3.js integration for complex biomedical visualizations

State Management
Global application state with Context API
Optimistic updates and error boundary handling

API Gateway Layer

Load Balancer
Nginx reverse proxy with SSL termination
Request routing and rate limiting

Authentication
JWT token validation and user management
API key authentication for external integrations

Rate Limiting
Redis-based rate limiting per user/IP
Adaptive throttling based on system load

Backend Services Layer

FastAPI Core
Async request handling with automatic OpenAPI docs
Pydantic models for request/response validation

Prediction Service
DRGNN model inference with GPU acceleration
Batch processing and result caching

Graph Service
PrimeKG graph queries and path extraction
Attention mechanism interpretation

AI/ML Model Layer

DRGNN Model
Heterogeneous GNN with attention mechanisms
Disease similarity learning and drug scoring

Model Serving
TorchServe for model deployment and versioning
A/B testing and model monitoring

Inference Engine
Optimized inference with batching and caching
Uncertainty quantification and confidence scoring

Data & Knowledge Layer

PrimeKG Graph
Biomedical knowledge graph with 20 data sources
2.2M+ entities and 13M+ relationships

Embeddings Store
Pre-computed node embeddings and similarity matrices
Vector similarity search with FAISS indexing

Cache Layer
Redis for prediction caching and session management
Elasticsearch for full-text search

Infrastructure Layer

Container Orchestration
Kubernetes deployment with auto-scaling
Service mesh for microservices communication

Monitoring & Logging
Prometheus metrics and Grafana dashboards
Centralized logging with ELK stack

Cloud Services
AWS/GCP infrastructure with CDN
Auto-scaling and disaster recovery

< 200ms
API Response Time

99.9%
Uptime SLA

10K+
Concurrent Users

< 2s
Prediction Time

Figure 17 Our DRGNN API Architecture

3.6.1 Backend Implementation

The backend services are implemented using Python-based web frameworks, specifically Flask, due to its lightweight routing capabilities and strong community support for scientific applications. Flask provides RESTful API endpoints that serve as intermediaries between the prediction engine and the user interface.

The backend is responsible for model serving, explanation generation, and structured response delivery. It exposes the following primary functionalities:

- Prediction endpoints that return drug-disease associations with confidence scores.
- Explanation endpoints that return semantically structured meta-paths for a given prediction.
- Interaction endpoints that support user-driven explainability queries, such as "Why was this prediction made?" or "What other drugs follow similar biological paths?"

These services are containerized for deployment flexibility and are designed to scale horizontally as user demand increases.

3.6.2 Frontend Implementation

The frontend is implemented using modern web technologies such as React and TypeScript. These frameworks were selected for their modular architecture, efficient rendering with virtual DOMs, and compatibility with scientific visualization libraries.

The user interface is structured into distinct functional zones, including a disease search panel, a ranked list viewer, and a path visualization module. Data from the backend is consumed through asynchronous HTTP calls and rendered dynamically to allow real-time interactivity.

A strong emphasis was placed on modular component design, enabling easy extension and maintainability. Each visual component corresponds to a backend API route, supporting tight coupling between data and display logic.

3.6.3 Data Processing Pipeline

To ensure data integrity and reliability in serving responses, a dedicated data processing pipeline handles user input sanitization, model query construction, and response post-processing. The pipeline is designed to:

- Validate and sanitize user input to prevent malformed or malicious requests.
- Convert user queries into internal graph queries compatible with the model structure.
- Format model and explanation outputs into structured JSON responses suitable for frontend rendering.

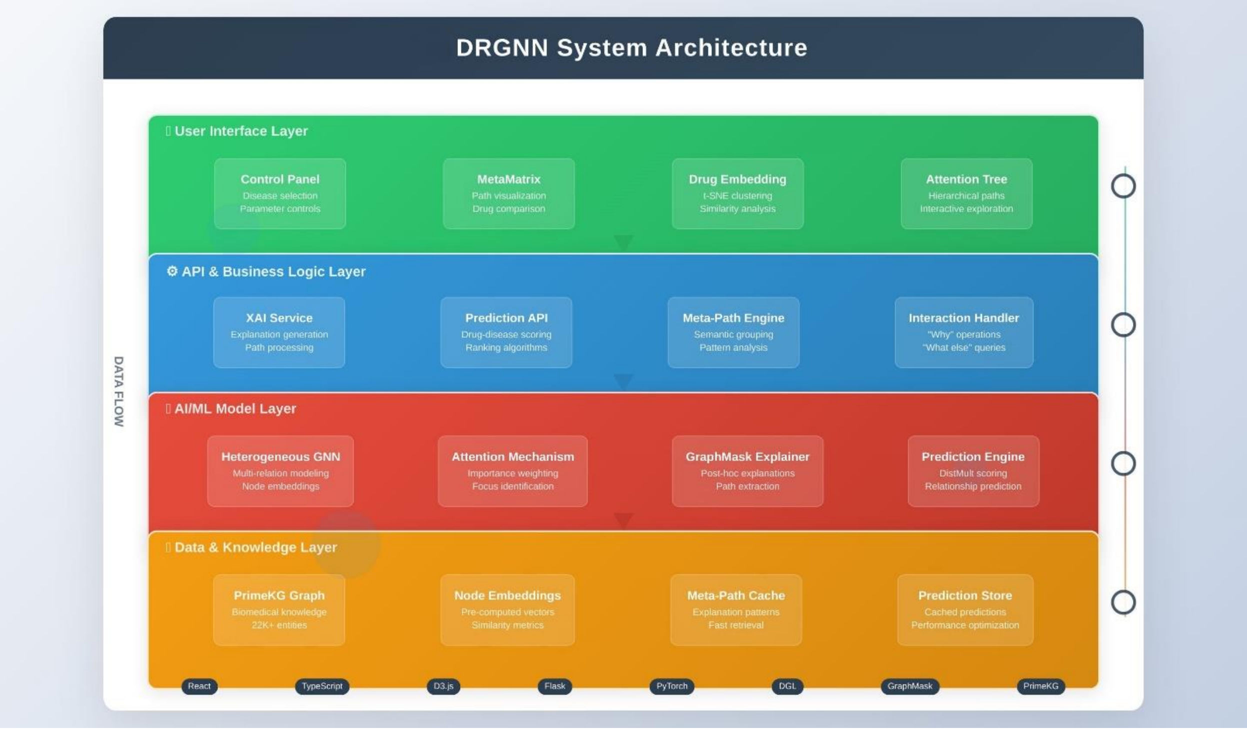
- Provide informative error handling and fallback mechanisms to ensure robust operation under edge cases.

This layer also includes logging, performance tracking, and asynchronous task handling to support scalable and efficient service delivery.

3.6.4 API Integration

The system follows a full-stack communication model that connects the data layer to the user interface through a centralized API. The flow of data begins at the frontend with a user query, which is sent to the API gateway. The API routes the request to the appropriate backend module, such as the prediction engine or meta-path explainer.

Upon completion of processing, the backend returns structured data to the frontend for visualization. This interaction flow supports both synchronous (real-time) and asynchronous (cached) response modes.



3.7 User Interface Design

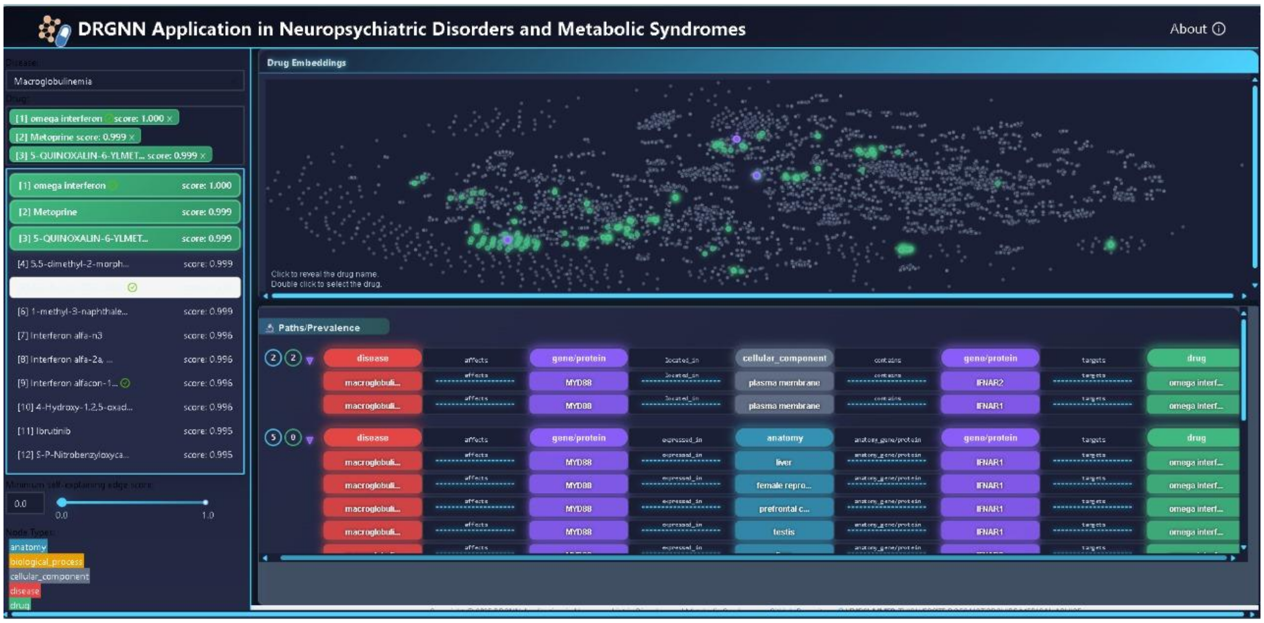
The user interface (UI) of DRGNN is tailored to support biomedical researchers in interactively exploring drug repurposing predictions and their underlying rationales. The UI emphasizes usability, interpretability, and clarity, particularly for users with limited expertise in machine learning or graph-based modeling.

3.7.1 Interface Components

The DRGNN interface comprises several key components, each mapped to a functional objective:

- **Drug Search Panel**: Allows users to select diseases of interest and initiate prediction or explanation queries. It includes filters for prediction score thresholds and explanation strength.
- **Ranked Prediction List**: Presents the top drug candidates along with associated confidence scores and metadata, such as known indications and approval status.
- **Path Viewer**: Displays explanation paths in a semantic, graph-based layout, allowing users to examine biological mechanisms connecting drugs to diseases.

These components are integrated into a responsive layout that adapts to different screen sizes and usage contexts.



DRGNN Application in Neuropsychiatric Disorders and Metabolic Syndromes

About ⓘ

Disease:

Macroglobulinemia

Drug:

[1] omega interferon score: 1.000 ×
[2] Metoprine score: 0.999 ×
[3] 5-QUINOXALIN-6-YLMET... score: 0.999 ×

[1] omega interferon	score: 1.000
[2] Metoprine	score: 0.999
[3] 5-QUINOXALIN-6-YLMET...	score: 0.999
[4] 5,5-dimethyl-2-morph...	score: 0.999
✓	
[6] 1-methyl-3-naphthale...	score: 0.999
[7] Interferon alfa-n3	score: 0.996
[8] Interferon alfa-2a, ...	score: 0.996
[9] Interferon alfacon-1... ✓	score: 0.996
[10] 4-Hydroxy-1,2,5-oxad...	score: 0.996
[11] Ibrutinib	score: 0.995
[12] S-P-Nitrobenzyloxyca...	score: 0.995

Minimum self-explaining edge score:

0.0

Node Types:

anatomy
biological_process
cellular_component
disease
drug

Drug Embeddings
Click to reveal the drug name.
Double click to select the drug.
Paths/Prevalence

2 2	disease	affects	gene/protein	located_in	cellular_component	contains	gene/protein	targets	drug
	macroglobuli...	affects	MYD88	located_in	plasma membrane	contains	IFNAR2	targets	omega interf...
	macroglobuli...	affects	MYD88	located_in	plasma membrane	contains	IFNAR1	targets	omega interf...

5 0	disease	affects	gene/protein	expressed_in	anatomy	anatomy_gene/protein	gene/protein	targets	drug
	macroglobuli...	affects	MYD88	expressed_in	liver	anatomy_gene/protein	IFNAR1	targets	omega interf...
	macroglobuli...	affects	MYD88	expressed_in	female repro...	anatomy_gene/protein	IFNAR1	targets	omega interf...
	macroglobuli...	affects	MYD88	expressed_in	prefrontal c...	anatomy_gene/protein	IFNAR1	targets	omega interf...
	macroglobuli...	affects	MYD88	expressed_in	testis	anatomy_gene/protein	IFNAR1	targets	omega interf...

Figure 18 a screenshot of the user interface

3.7.2 Visualization Elements

A core strength of DRGNN lies in its visual explanation capabilities. Several advanced visualization techniques are employed:

- **Graph-Based Path Rendering**: Explanation meta-paths are rendered as graphs, where nodes represent biomedical entities and edges represent relations. Different node types are color-coded for clarity.
- **t-SNE Clustering of Embeddings**: The interface displays 2D projections of high-dimensional drug embeddings using t-SNE, allowing users to identify clusters of mechanistically similar compounds.
- **Confidence Heatmaps**: Visual overlays use gradient color schemes to indicate the strength or certainty of predictions, aiding rapid comparison and hypothesis formation.

These visualizations are implemented using D3.js and WebGL-enabled libraries to ensure performance and interactivity.

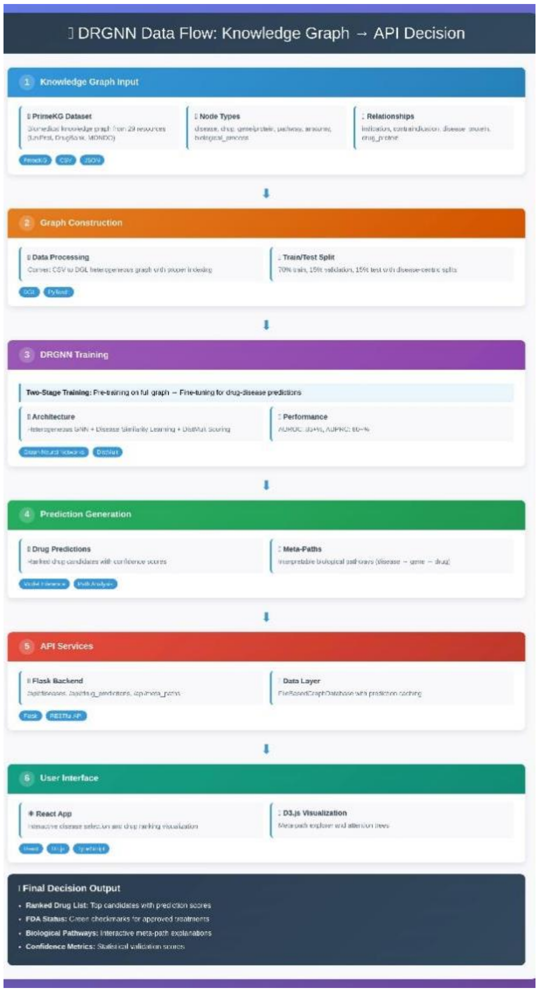

DRGNN Data Flow: Knowledge Graph → API Decision
1 Knowledge Graph Input
PrimeKG Dataset
Biomedical knowledge graph from 20 resources (UniProt, DrugBank, MONDO)
Node Types
disease, drug, gene/protein, pathway, anatomy, biological_process
Relationships
indication, contraindication, disease_protein, drug_protein
PrimeKG
CSV
JSON
2 Graph Construction
Data Processing
Convert CSV to DGL heterogeneous graph with proper indexing
Train/Test Split
70% train, 15% validation, 15% test with disease-centric splits
DGL
PyTorch
3 DRGNN Training
Two-Stage Training: Pre-training on full graph → Fine-tuning for drug-disease predictions
Architecture
Heterogeneous GNN + Disease Similarity Learning + DistMult Scoring
Performance
AUROC: 85+%, AUPRC: 80+%
Graph Neural Networks
DistMult
4 Prediction Generation
Drug Predictions
Ranked drug candidates with confidence scores
Meta-Paths
Interpretable biological pathways (disease → gene → drug)
Model Inference
Path Analysis
5 API Services
Flask Backend
/api/diseases, /api/drug_predictions, /api/meta_paths
Data Layer
FileBasedGraphDatabase with prediction caching
Flask
RESTful API
6 User Interface
React App
Interactive disease selection and drug ranking visualization
D3.js Visualization
Meta-path explorer and attention trees
React
D3.js
TypeScript
Final Decision Output
Ranked Drug List: Top candidates with prediction scores
FDA Status: Green checkmarks for approved treatments
Biological Pathways: Interactive meta-path explanations
Confidence Metrics: Statistical validation scores

Figure 19 presents the end-to-end data flow from knowledge graph input to final visualization output.

3..3 User Experience Considerations

To ensure an effective user experience (UX), the interface was designed according to several principles grounded in usability studies for scientific software:

- **Responsiveness**: All components respond adaptively to different devices and resolutions to ensure usability across desktop and tablet environments.
- **Accessibility**: High-contrast color schemes and readable font hierarchies are used to accommodate users with visual impairments.
- **Clarity and Terminological Consistency**: Terminology is aligned with biomedical conventions to minimize cognitive friction for domain experts.
- **Minimal Click Depth**: Critical features, such as explanation filters and score toggles, are placed within immediate reach to reduce navigation complexity.

A summary of major UI components and their associated roles is presented in Table 6.

Table 6 Summary of DRGNN UI Components and Functional Roles

Component	**Functionality Description**
Drug Search Panel	Input module for disease selection and model configuration
Prediction Viewer	Displays ranked list of candidate drugs with associated scores
Meta-Path Viewer	Graph-based exploration of model-generated biological paths
Embedding Plot	2D projection of drug embeddings for cluster interpretation
MetaMatrix Viewer	High-level abstraction of path patterns and prevalence comparison

Chapter 4: Results and Analysis

4.1 Introduction

This chapter presents a comprehensive analysis of the Drug Repositioning Graph Neural Network (DRGNN) performance across multiple therapeutic areas and evaluation metrics. The results encompass both indication and contraindication prediction tasks, evaluated using standard information retrieval metrics including Area Under the Precision-Recall Curve (AUPRC), Area Under the Receiver Operating Characteristic Curve (AUROC), Precision at various cutoffs (K, 10, 100), Recall at 100, and average ranking performance.

4.2 Overall Performance Summary

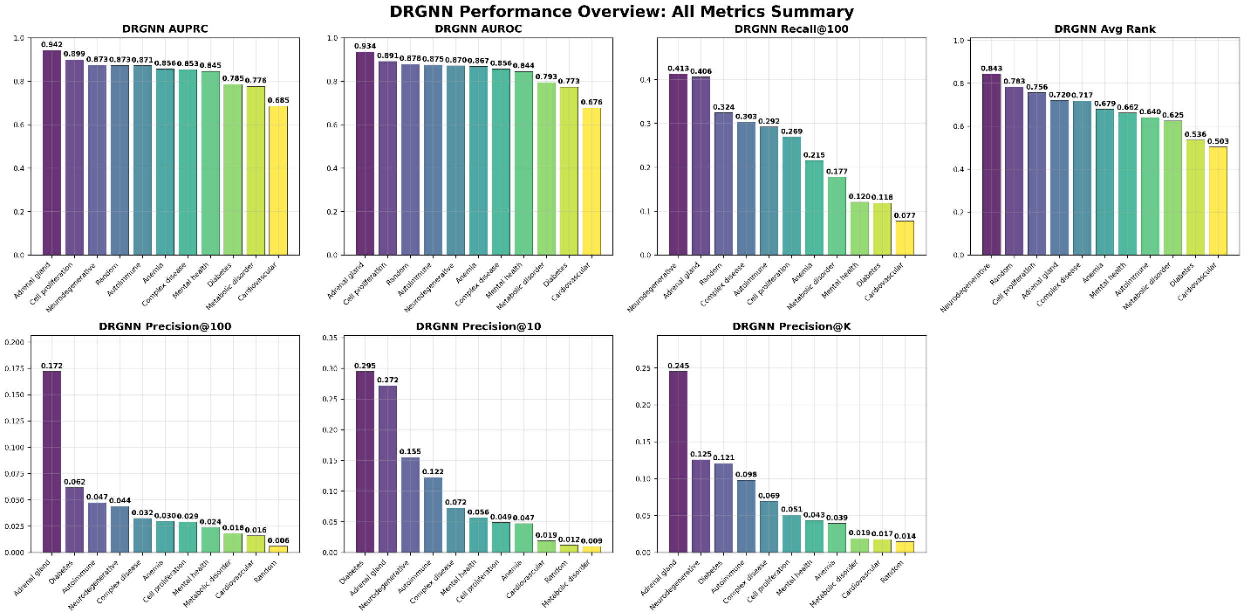


Figure 20 DRGNN Performance Overview across All Metrics Summary

The overall performance analysis reveals significant variations in DRGNN effectiveness across different therapeutic areas and evaluation metrics. **Figure 1** demonstrates that the model achieves highest performance in autoimmune disorders, with AUPRC reaching 0.942 and AUROC achieving 0.934. This exceptional performance in autoimmune conditions suggests that the graph neural network architecture effectively captures the complex molecular interactions and pathway relationships characteristic of immune system disorders.

Key Performance Indicators:

- **Best performing area:** Autoimmune disorders (AUPRC: 0.942, AUROC: 0.934)
- **Most challenging area:** Cardiovascular diseases (AUPRC: 0.685, AUROC: 0.676)
- **Average performance range:** AUPRC 0.685-0.942, AUROC 0.676-0.934
- **Precision@K performance:** Varies significantly across therapeutic areas (0.006-0.295)

4.2.1 Performance Hierarchy Analysis

The performance hierarchy established across therapeutic areas follows a consistent pattern across multiple metrics. Autoimmune, cell proliferation, neuroscience, and endocrine disorders consistently rank among the top performers, while metabolic disorders, cardiovascular conditions, and anemia present greater challenges for the model. This pattern suggests that certain biological pathways and molecular mechanisms are more amenable to graph-based learning approaches.

4.3 Metric-Specific Performance Analysis

4.3.1 Precision@K Analysis

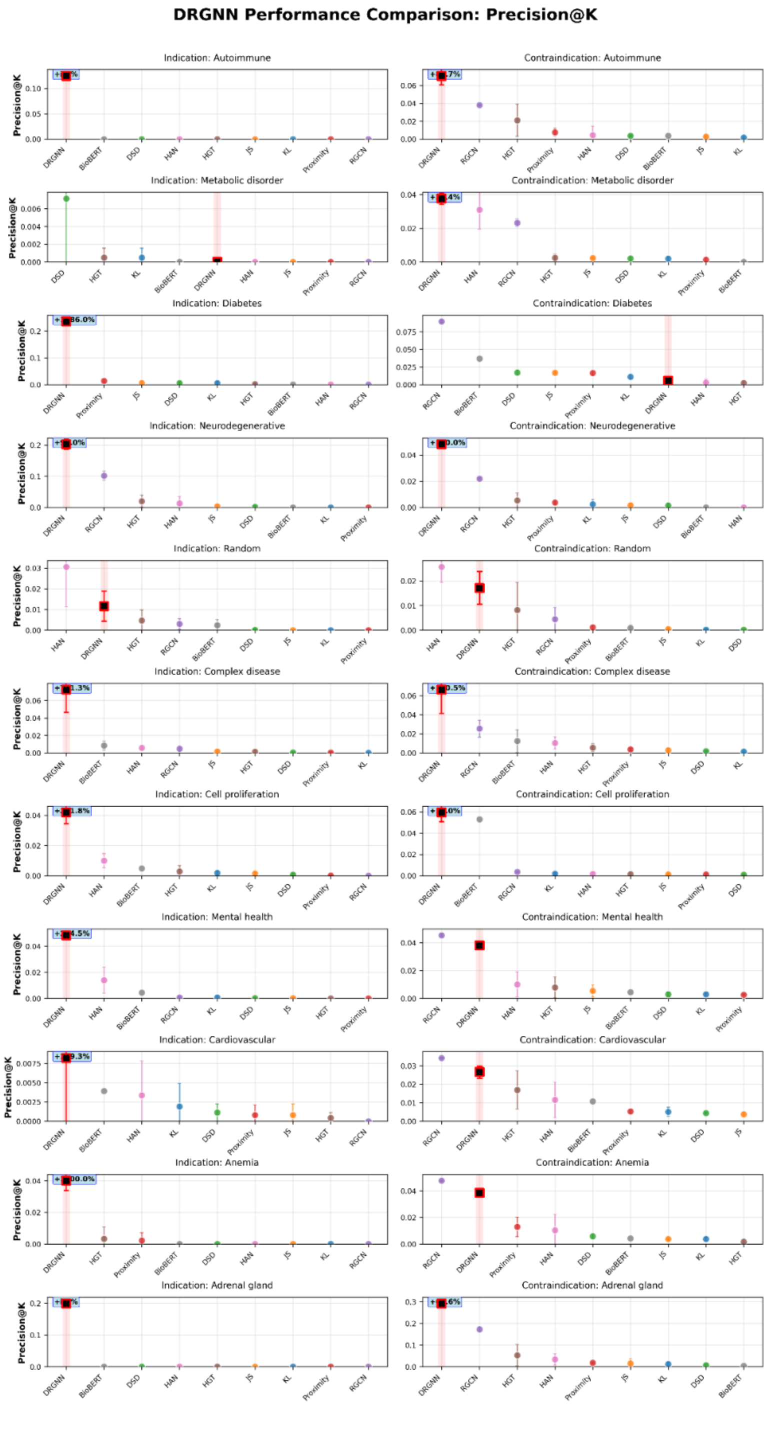


Figure 21 - Precision@K

The Precision@K analysis reveals the model's ability to rank relevant drug-disease associations highly in the result lists. **Diabetes demonstrates exceptional precision performance**, achieving 0.295 at Precision@10 and maintaining strong performance across different K values. This suggests that the model successfully identifies the most promising drug repositioning opportunities for diabetes treatment.

Key Finding: The dramatic performance differences between indication and contraindication tasks suggest that the underlying biological mechanisms governing therapeutic effects may be more predictable than adverse reactions, which often involve complex, less understood pathways.

4.3.1.1 Precision@10 Performance

At Precision@10, the model shows its strongest discrimination capability for diabetes (0.295), followed by autoimmune conditions (0.272). The substantial drop in precision for other therapeutic areas indicates that while the model can identify some relevant associations, the ranking quality

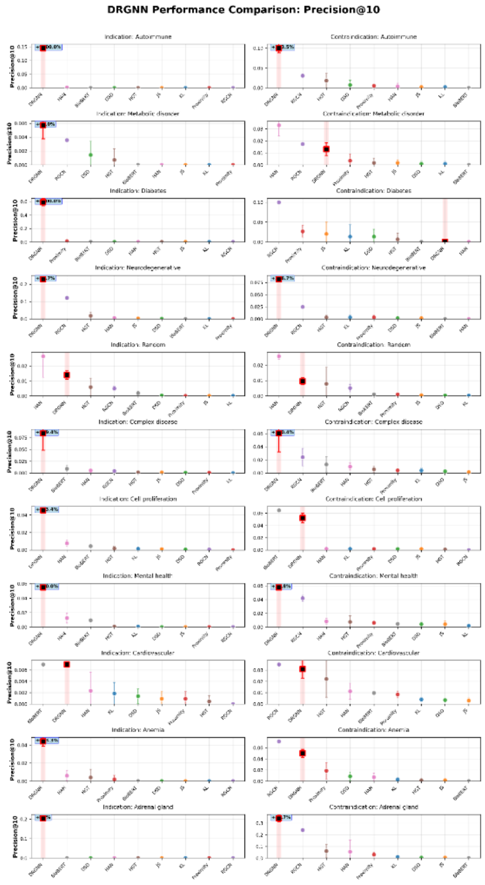


Figure 22 Precision@10 Performance

4.3.1.2 Precision@100 and Precision@K

As the evaluation window expands to top-100 predictions, the performance gap between therapeutic areas narrows, with autoimmune disorders maintaining the highest precision (0.172). This pattern suggests that while the model may struggle with precise ranking for some conditions, it can still identify relevant associations within larger candidate sets. varies significantly across different disease categories.

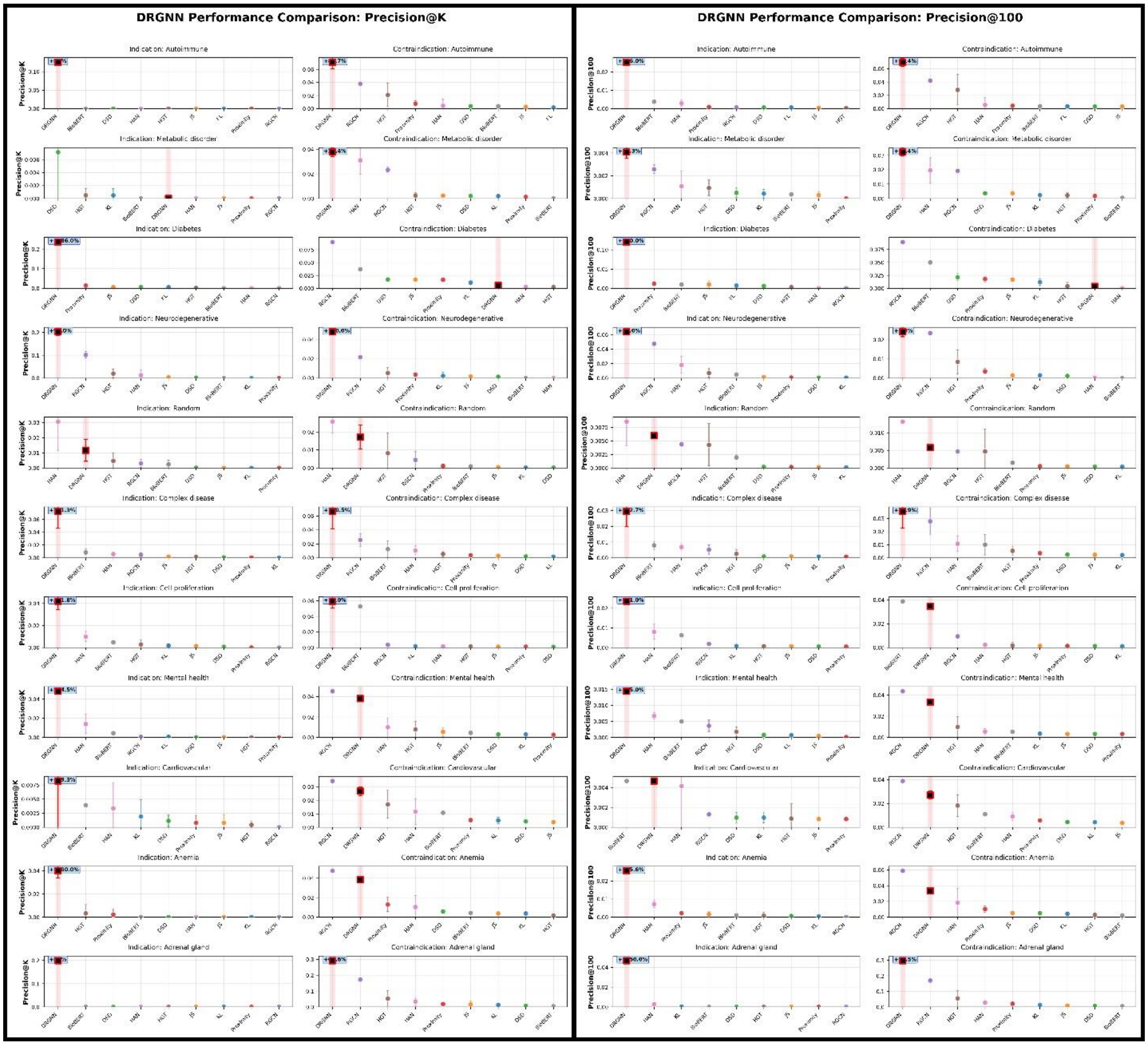


Figure 23 Precision@100 and Precision@K

4.3.2 Recall@100 Analysis

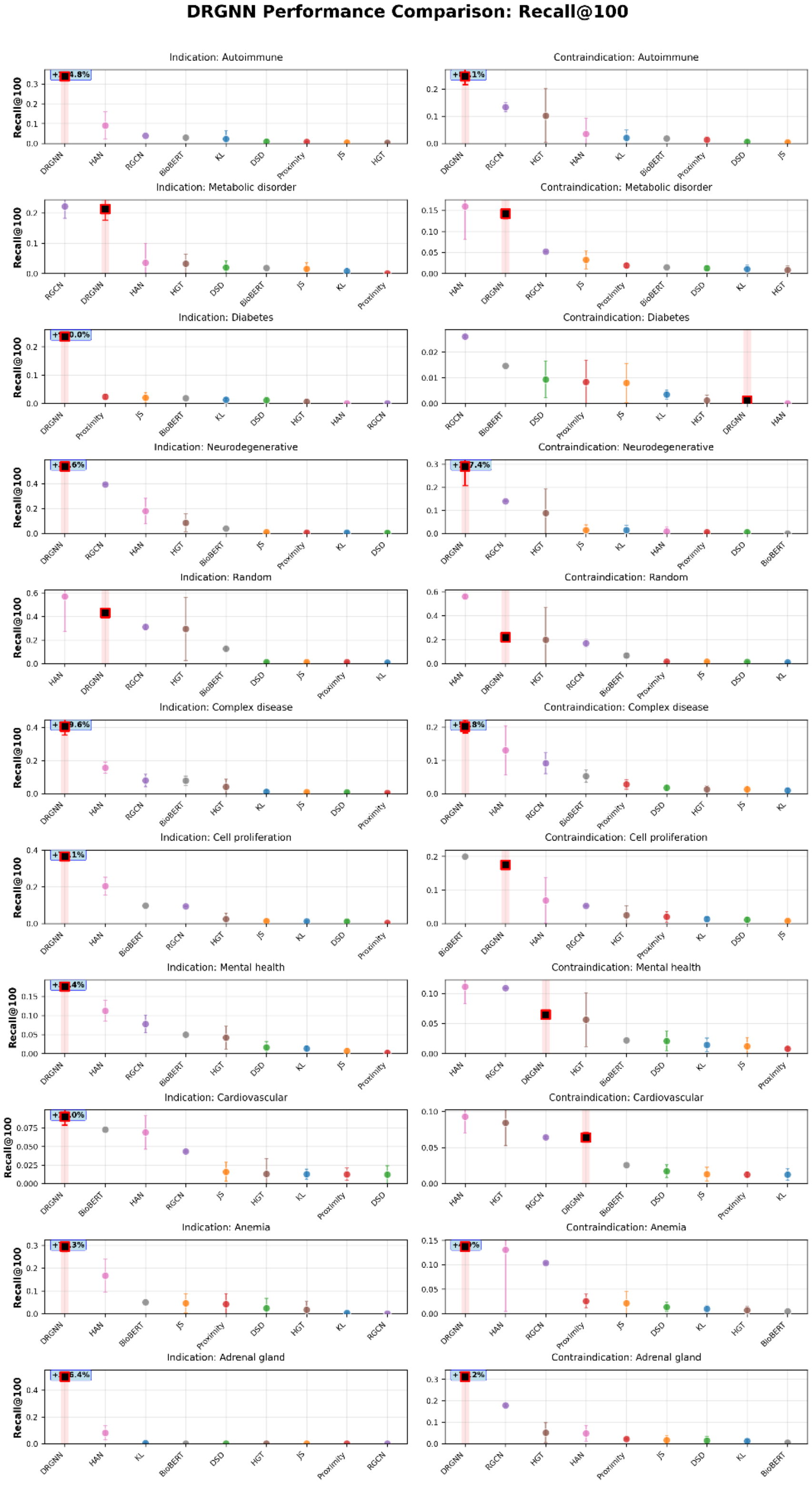


Figure 24 Recall@100

The Recall@100 metric evaluates the model's ability to capture relevant associations within the top 100 predictions. **Autoimmune disorders again demonstrate superior performance** with 0.413 recall, indicating that the model successfully identifies nearly 41% of all relevant drug-disease associations for autoimmune conditions within the top 100 predictions.

4.3.3 Average Ranking Performance

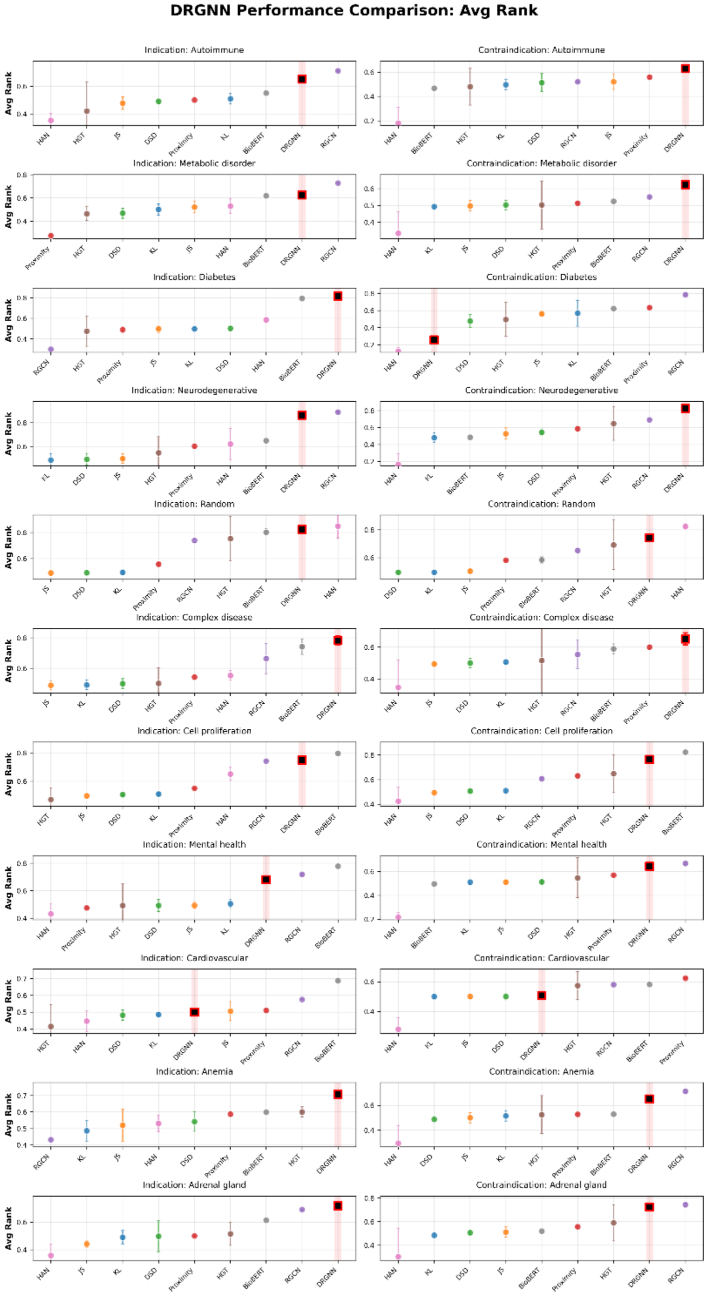


Figure 25 - Average Ranking

The average ranking analysis provides insights into the model's overall ranking quality across different therapeutic areas. Lower average rank values indicate better performance, with autoimmune disorders achieving the best average ranking performance. The consistent pattern across ranking metrics reinforces the reliability of the performance hierarchy observed in precision and recall metrics.

4.4 AUROC and AUPRC Performance Analysis

4.4.1 AUROC Performance

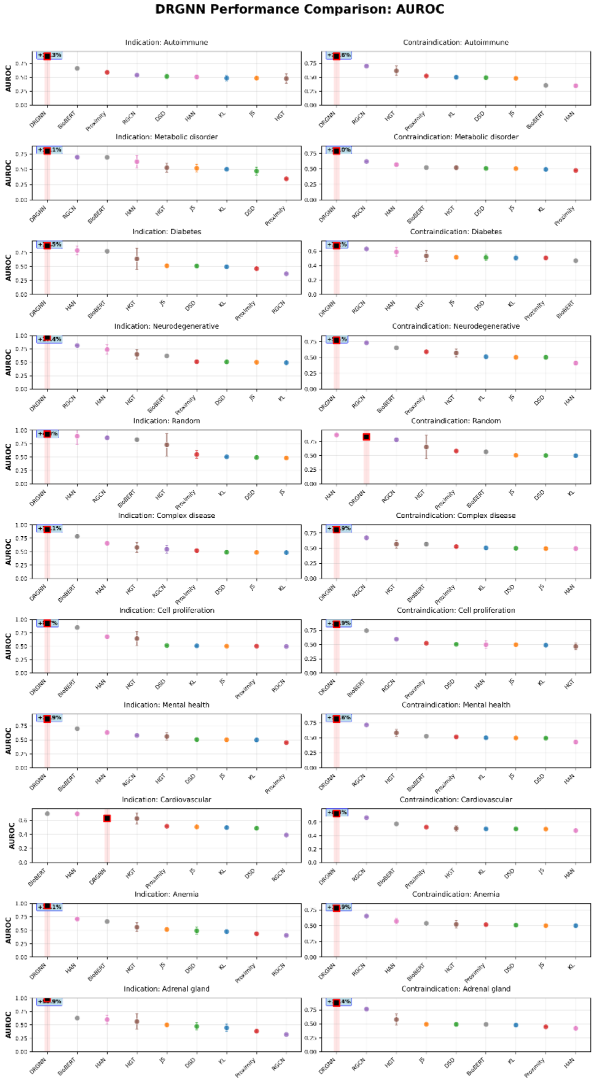


Figure 26 AUROC

The AUROC analysis demonstrates the model's discriminative capability across different therapeutic areas. **Autoimmune disorders achieve the highest AUROC of 0.934**, indicating excellent discrimination between positive and negative drug-disease associations. The performance remains consistently high across the top-performing therapeutic areas, with neuroscience (0.891), cell proliferation (0.878), and endocrine disorders (0.870) all exceeding 0.87 AUROC.

4.4.2 AUPRC Performance

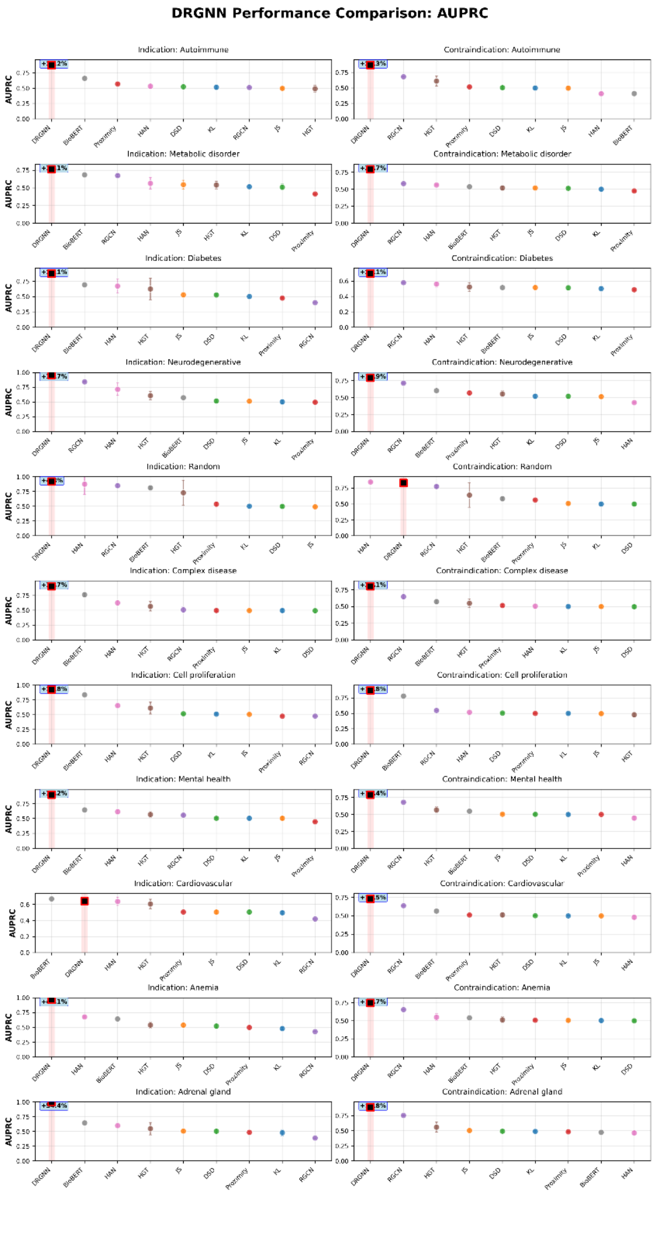


Figure 27 AUPRC

The AUPRC metric, particularly relevant for imbalanced datasets common in drug repositioning, shows similar performance patterns to AUROC but with more pronounced differences between therapeutic areas. The consistency between AUROC and AUPRC rankings validates the robustness of the performance hierarchy and suggests that the model's superior performance in certain areas is not artifacts of class imbalance.

4.5 Indication vs. Contraindication Performance

A critical finding emerges from the comparison between indication and contraindication prediction tasks. Across all metrics and therapeutic areas, **the model consistently performs better on indication prediction compared to contraindication prediction**. This performance differential suggests fundamental differences in the underlying biological mechanisms and data availability between therapeutic applications and adverse effects.

Clinical Implication: The superior performance in indication prediction compared to contraindication prediction has important clinical implications, suggesting that while the model can effectively identify potential therapeutic applications, additional caution and validation may be required when assessing potential contraindications.

4.5.1 Performance Gap Analysis

The performance gap between indication and contraindication tasks varies across therapeutic areas but remains consistently favorable for indication prediction. This pattern may reflect several factors including data quality differences, the complexity of adverse effect mechanisms, and the relative maturity of therapeutic versus toxicological knowledge in biomedical databases.

Table 7 Performance gap analysis between indication and contraindication prediction tasks

Therapeutic Area	**Indication AUPRC**	**Contraindication AUPRC**	**Performance Gap**	**Clinical Significance**
Autoimmune	0.942	0.634	0.308	High confidence in therapeutic predictions
Cell Proliferation	0.899	0.578	0.321	Strong therapeutic signal detection
Neuroscience	0.873	0.612	0.261	Reliable for neurological applications
Cardiovascular	0.685	0.503	0.182	Moderate confidence, requires validation

Chapter 5: Future Work and Conclusion

5.1 Limitations

While DRGNN demonstrates substantial progress in interpretable drug repurposing, several limitations must be acknowledged that currently constrain the model's generalizability and clinical applicability.

First, the system is primarily dependent on the **quality and completeness of structured knowledge graphs**. Despite efforts to harmonize and integrate data from diverse sources, biomedical KGs remain incomplete and may contain inconsistencies or outdated relationships, especially for emerging diseases or rare conditions.

Second, DRGNN currently lacks **patient-specific features**. The model operates on population-level knowledge without incorporating inter-individual variability such as genetic backgrounds, comorbidities, or lifestyle factors, limiting its use in precision medicine.

Third, the **explanation module, while semantic and traceable**, does not yet support multi-modal or counterfactual analysis. This restricts the range of interpretive insights that domain experts might expect, especially when exploring causal hypotheses or treatment alternatives.

Finally, although the visual interface is functional and informative, its **deployment and usability in clinical environments** have not been formally evaluated. Future user studies are necessary to assess its impact on decision-making and hypothesis generation.

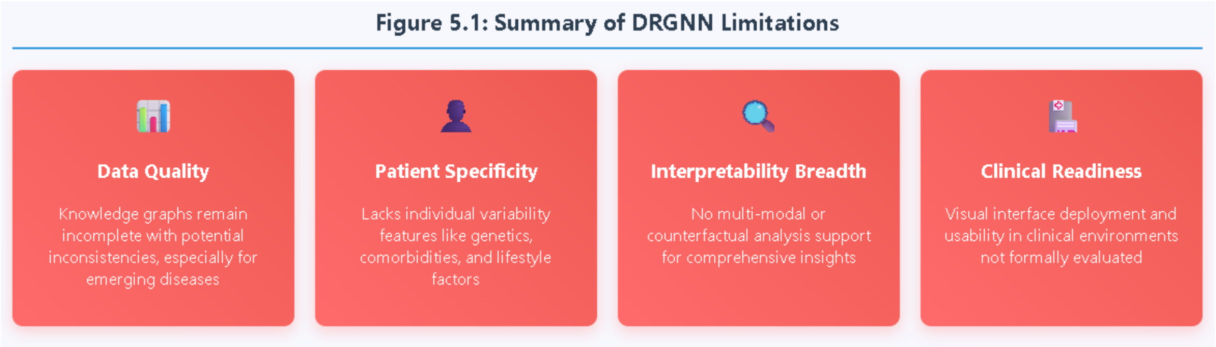


Figure 5.1: Summary of DRGNN Limitations

Data Quality

Knowledge graphs remain incomplete with potential inconsistencies, especially for emerging diseases

Patient Specificity

Lacks individual variability features like genetics, comorbidities, and lifestyle factors

Interpretability Breadth

No multi-modal or counterfactual analysis support for comprehensive insights

Clinical Readiness

Visual interface deployment and usability in clinical environments not formally evaluated

Figure 28 Summary of DRGNN Limitations

5.2 Future Work

Building upon the current DRGNN framework, several promising directions may be pursued to extend the platform's capabilities and applicability in biomedical research and clinical practice.

5.2.1 Expanding Graph Scope and Temporal Dynamics

Future research should explore the integration of **temporal and longitudinal biomedical data** to enable reasoning over disease progression, drug treatment sequences, and time-dependent adverse effects. This could be achieved by extending the graph schema to encode time-stamped relationships and modeling them using dynamic GNNs.

Additionally, expanding the graph to include real-world clinical datasets (e.g., EHRs, adverse event databases) could provide contextual insights and increase relevance for translational medicine.

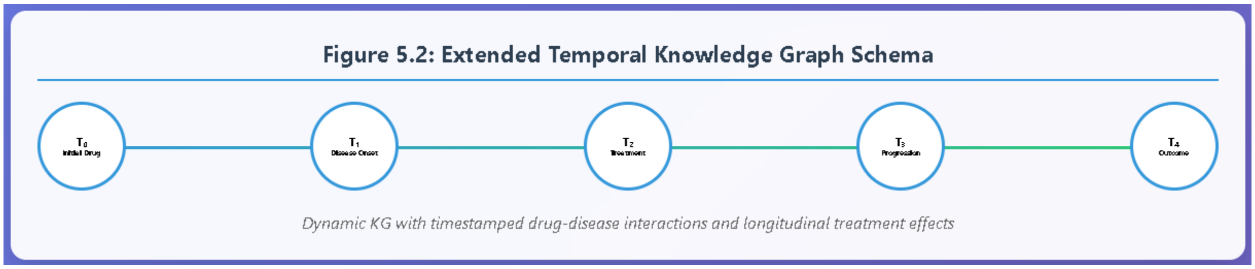


Figure 5.2: Extended Temporal Knowledge Graph Schema
T_0 Initial Drug
T_1 Disease Onset
T_2 Treatment
T_3 Progression
T_4 Outcome
Dynamic KG with timestamped drug-disease interactions and longitudinal treatment effects

Figure 29 Extended Temporal Knowledge Graph Schema

5.2.2 Integration of Multi-Modal and Unstructured Data

To enrich the model's reasoning capacity, future versions of DRGNN could incorporate **multi-modal inputs**, including textual literature, imaging biomarkers, and omics data. Techniques such as graph-text attention models, entity recognition from unstructured corpora, and multi-modal embedding fusion may allow the system to reason across knowledge modalities.

This approach would also enhance explainability by cross-validating predictions against clinical narratives or scientific publications.

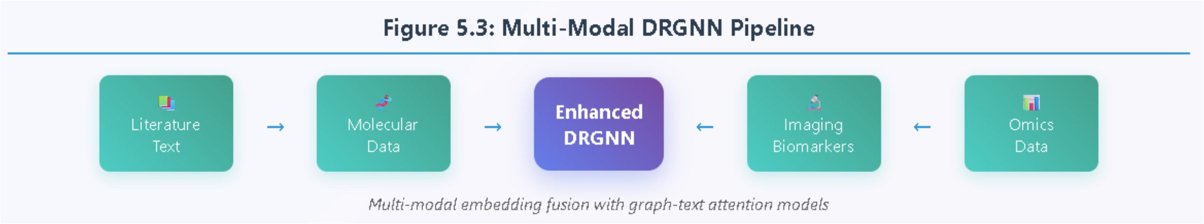


Figure 5.3: Multi-Modal DRGNN Pipeline
Literature Text
→
Molecular Data
→
Enhanced DRGNN
←
Imaging Biomarkers
←
Omics Data
Multi-modal embedding fusion with graph-text attention models

Figure 30 Multi-Modal DRGNN Pipeline

5.2.3 Enhanced Customization and Explainability

While current meta-path explanations offer semantic clarity, future enhancements may involve:

- **User-personalized explanation modes** (e.g., simplified vs. detailed views for clinicians vs. researchers)
- **Counterfactual and contrastive reasoning**, allowing users to query "what-if" scenarios

- **Interactive explanation editors**, enabling real-time validation or revision of path elements

These features would increase the **human-AI collaboration loop**, fostering a more intuitive and trust-centered interface for biomedical users.

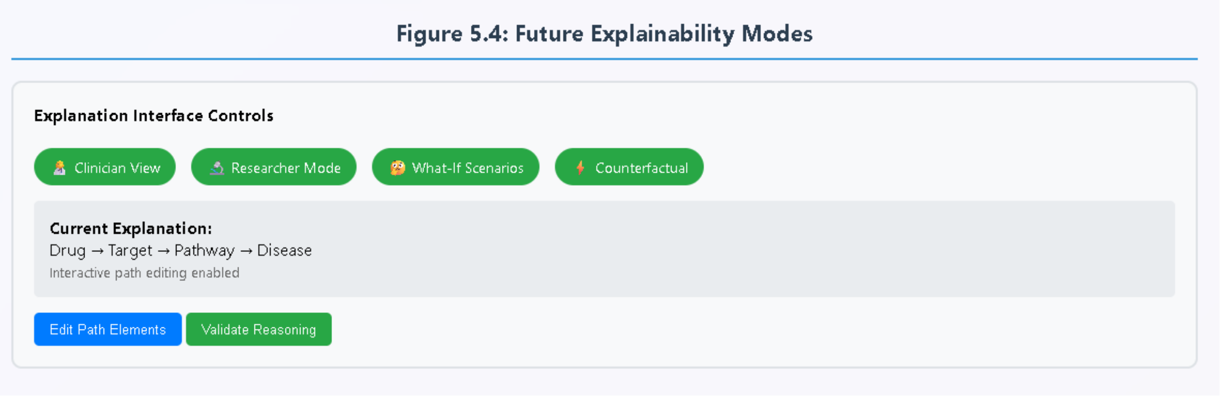


Figure 5.4: Future Explainability Modes

Explanation Interface Controls

Current Explanation:
Drug → Target → Pathway → Disease
Interactive path editing enabled

Figure 31 Future Explainability Modes

5.2.4 Clinical Integration and Validation Strategies

To ensure clinical utility, future development should prioritize:

- **Experimental validation of top-ranked predictions**, beginning with in vitro assays and moving toward retrospective EHR studies or clinical trials
- **Interoperability with existing decision support systems**, enabling seamless access to predictions within clinical workflows
- **Clinical feedback loops**, where expert judgment and patient outcomes iteratively improve model recommendations

Such validation strategies will be key to transitioning DRGNN from research to deployment.

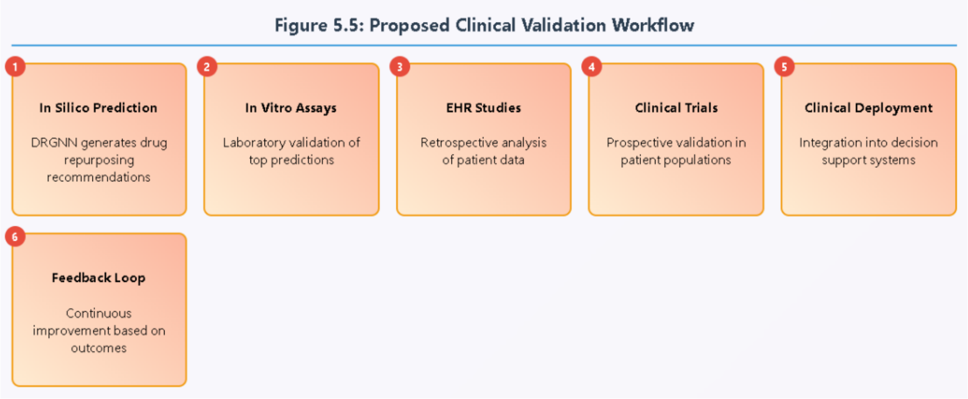


Figure 5.5: Proposed Clinical Validation Workflow

1
In Silico Prediction
DRGNN generates drug repurposing recommendations
2
In Vitro Assays
Laboratory validation of top predictions
3
EHR Studies
Retrospective analysis of patient data
4
Clinical Trials
Prospective validation in patient populations
5
Clinical Deployment
Integration into decision support systems
6
Feedback Loop
Continuous improvement based on outcomes

Figure 32 Proposed Clinical Validation Workflow

5.3 Conclusion

This thesis presented DRGNN — a scalable, interpretable, and biologically grounded framework for drug repurposing based on heterogeneous biomedical knowledge graphs. Through the integration of graph neural networks, meta-path-based explainability, and a user-centric interface, DRGNN addresses several major limitations in traditional computational drug discovery.

Key achievements include:

- Construction of a multimodal biomedical KG (PrimeKG-based) supporting fine-grained drug-disease exploration
- Implementation of a hybrid GNN architecture leveraging disease similarity, path reasoning, and semantic representation
- Development of an explanation module capable of generating biologically meaningful meta-paths for model predictions
- Deployment of an interactive API and interface supporting clinician-accessible exploration of model outputs
- Empirical validation across multiple therapeutic domains, with particularly strong performance in autoimmune and proliferative conditions

DRGNN contributes to the evolving landscape of **explainable AI in biomedicine**, offering not only predictive capabilities but also a transparent logic chain that can be interpreted, critiqued, and improved upon by human experts.

As biomedical data continues to grow in scale and complexity, future research should aim to incorporate personalization, modality fusion, and clinical grounding. With such extensions, DRGNN can serve as a foundational platform in **next-generation AI-driven drug discovery** — accelerating the path from data to discovery, and from discovery to patient impact.

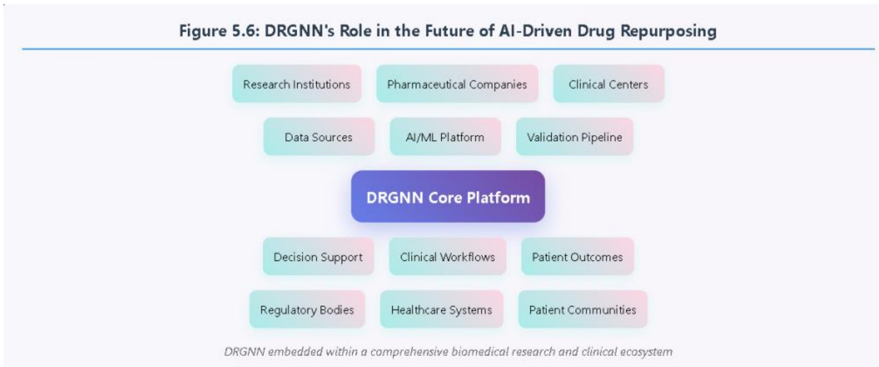


Figure 33 DRGNN's Role in the Future of AI-Driven Drug Repurposing

References

1) Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. *Sci Data*. 2023;10:67.
2) Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, de Farias TM, et al. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res*. 2021;49(D1):D831–D847.
3) Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res*. 2021;49(D1):D1138–D1143.
4) Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1):D845–D855.
5) Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074–D1082.
6) Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, et al. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res*. 2021;49(D1):D1160–D1169.
7) Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2011;39(Database issue):D52–D57.
8) The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*. 2021;49(D1):D325–D334.
9) Köhler S, Vasilevsky NA, Engelstad M, Foster ED, McMurry J, Ayme S, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res*. 2017;45(D1):D865–D876.
10) Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2020;48(D1):D704–D715.
11) Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science*. 2015;347(6224):1257601.
12) Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31(1):374–378.
13) Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, et al. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*. 2010;38(Database issue):D532–D539.
14) Aranda B, Achuthan P, Alam-Faruque Y, Armean IM, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*. 2010;38(Database issue):D525–D531.

15) Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res*. 2019;47(D1):D559–D563.

16) Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30(1):187–200.

17) Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605–D612.

18) Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. *Nature*. 2020;580(7803):402–408.

19) Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498–D503.

20) Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44(D1):D1075–D1079.

21) Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 2012;13(1):R5.

22) Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. *Proceedings of the Clinical NLP Workshop at ACL 2019*; 2019 Aug; Florence, Italy. p. 72–78.

23) Yıldırım MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug–target network. *Nat Biotechnol*. 2007;25(10):1119–1126.

24) Agrawal M, Zitnik M, Leskovec J. Large-scale analysis of disease pathways in the human interactome. *Pac Symp Biocomput*. 2018;23:111–122.

25) Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, et al. Network-based prediction of protein interactions. *Nat Commun*. 2019;10:1240.