Syrian Private University
Faculty of Informatics Engineering
Department of Artificial Intelligence and Data Science

This thesis was prepared:
To complete my graduation1 project in the field of Artificial Intelligence and Data Science

Titled:
Knowledge Graph-Enhanced Drug-Gene Interaction Prediction in Alzheimer's Disease:
A GNN-LLM Approach

Prepared by:
Maryam Adel Abdul Aal

Supervised by:
Dr. Maysaa Abu Al-Qasim _ Engineer Aya Al-Aswad

First semester

2025/2026

Abstract

Alzheimer's disease represents a major global health challenge with limited effective treatment options. Drug repurposing offers a promising strategy to identify novel therapeutic applications for existing drugs, but traditional experimental approaches remain time-consuming and costly. This research introduces a novel approach for drug-gene interaction prediction in Alzheimer's disease by integrating knowledge graphs with graph neural networks and large language models.

We constructed an Alzheimer-centered subgraph from AlzKB knowledge graph containing 998 nodes and 2,111 edges, encompassing 34 disease variants, 103 Alzheimer-associated genes, and 861 related drugs. Our methodology employs a weighted ensemble of Relational Graph Convolutional Networks (RGCN) and Relational Graph Attention Networks (RGAT) with focal loss to handle class imbalance, enabling multi-class prediction of drug-gene interaction types (Binds, Increases Expression, Decreases Expression, or No Relation).

The ensemble model achieved 89.7% overall accuracy and 94.2% macro-averaged AUC, demonstrating effective performance across all interaction classes. The system incorporates Large Language Models for generating scientific explanations of predicted interactions, enhancing interpretability for researchers and clinicians. An interactive web interface was developed to facilitate real-time predictions and comprehensive analysis.

The approach demonstrates the potential of combining structured biomedical knowledge with advanced deep learning techniques to accelerate drug discovery for Alzheimer's disease, providing a scalable framework for precision medicine applications.

Keywords: Alzheimer's disease, drug repurposing, knowledge graphs, graph neural networks, relational graph convolutional networks, large language models, drug-gene interactions, precision medicine, computational drug discovery.

Index

table of contents

Chapter One

Introduction

1.1 Background and Research Motivation

Alzheimer's disease (AD) is one of the most prevalent neurodegenerative disorders worldwide and represents a major public health challenge due to its progressive nature, irreversible cognitive decline, and growing socioeconomic impact. The disease is characterized by complex pathological mechanisms involving genetic, molecular, and biochemical interactions, which makes its study particularly challenging from both biomedical and computational perspectives. The traditional de novo drug discovery process for AD has historically proven lengthy and costly, underscoring the necessity for accelerated therapeutic strategies, such as drug repurposing.

In parallel with advances in biomedical research, there has been a rapid growth in large-scale biomedical databases and knowledge bases that store structured information about diseases, genes, drugs, and their interactions. To effectively unify and analyze this highly heterogeneous data, Knowledge Graphs (KGs) have emerged as a powerful paradigm for representing such complex systems by modeling biomedical entities as nodes and their relationships as edges. When combined with modern Graph-based Learning techniques, KGs enable integrated analysis and support data-driven discovery, particularly in Drug Repurposing. This thesis is motivated by the potential of Generative Graph-based Artificial Intelligence methods to enhance the computational identification of novel AD therapeutics through structured representation and learned relationship prediction. This approach aims to move beyond simple link classification towards the probabilistic generation of novel, high-confidence therapeutic hypotheses.

1.2 Problem Statement

Despite the availability of extensive biomedical data related to AD, extracting meaningful and novel therapeutic insights remains a significant challenge. The data are distributed across heterogeneous sources, involve multiple entity types, and exhibit complex relational dependencies that are difficult to model using conventional analytical techniques. As a result, important structural patterns and **indirect relationships between existing drugs and AD-related gene targets** may remain undiscovered.

There is a critical need for robust computational frameworks that can represent, integrate, and analyze Alzheimer-related biomedical data in a unified manner, while preserving both local and global relational structure. Addressing this challenge requires graph-based representations and **Generative Link Prediction** methods capable of capturing the complexity of biomedical knowledge at scale to prioritize drug candidates efficiently.

1.3 Research Objectives

1. To construct a high-confidence, AD-centric Knowledge Subgraph from the comprehensive biomedical knowledge base (ALZKB).
2. To apply **Graph Representation Learning (GRL)** techniques, including DeepWalk, to generate robust feature representations (embeddings) that capture the structural and semantic properties of the subgraph nodes.
3. To develop and evaluate advanced Graph Neural Network models, specifically Relational Graph Convolutional Networks (RGCN) and Relational Graph Attention Networks (RGAT), for multi-class drug-gene interaction prediction with focal loss to handle class imbalance.
4. To implement a weighted ensemble approach combining RGCN and RGAT predictions to improve overall classification performance across interaction types (Binds, Increases Expression, Decreases Expression, No Relation).
5. To integrate Large Language Models (LLMs) for generating scientific explanations of predicted drug-gene interactions, enhancing interpretability and providing biological context for computational predictions.
6. To develop an interactive web interface for real-time prediction and analysis, enabling researchers to explore drug-gene interactions and receive AI-generated explanations with confidence scores.

1.4 Research Gap

Existing computational studies on Alzheimer's disease (AD) drug repurposing often rely on predictive models or simple Knowledge Graph Embedding (KGE) techniques that do not fully exploit rich structural and semantic information embedded within large-scale biomedical knowledge graphs. Many existing approaches further depend on limited or uni-modal network features, which restrict their ability to capture complex drug–gene–disease relationships. Additionally, most current frameworks lack interpretability mechanisms to explain predicted interactions in biologically meaningful terms.

This research gap highlights the need for a unified graph-based framework that:

- Integrates multi-level graph features, including network topology (degree-based features, PageRank), semantic information (node labels), and structural context (DeepWalk embeddings).
- Utilizes advanced graph neural architectures (RGCN/RGAT ensemble) with focal loss for handling class imbalance in multi-class interaction prediction.
- Provides interpretable predictions through LLM-generated explanations, bridging the gap between computational results and biological understanding.
- Offers interactive tools for researchers to explore and validate predictions in real-time scenarios.

1.5 Scope and Limitations of the Work

The scope of this research is strictly defined by the following boundaries:

• **Disease Focus:** The study is exclusively focused on Alzheimer’s Disease (AD) and targets drug–gene interactions associated with its known molecular pathology.

• **Data Source:** The project relies primarily on information derived from the ALZKB Knowledge Graph. Consequently, the predictions are constrained by the quality, coverage, and completeness of the underlying biomedical data.

• **Computational Focus:** The methodological core of this work centers on graph representation learning using DeepWalk embeddings and graph-based link prediction models. These predictive results serve as the foundation for downstream generative analysis using large language models to produce biological explanations and research hypotheses.

• **Limitations:** This study is computational in nature. The identified drug repurposing candidates represent high-confidence predictions that require subsequent experimental validation through wet-lab studies or clinical trials.

1.6 General Methodology of the Project

This research adopts a structured, multi-phase computational methodology, where each phase builds upon the outputs of the previous one. The implemented methodology includes the following stages:

• **Data Preparation and Subgraph Construction:** Extracting and refining an Alzheimer's disease (AD)-centric knowledge subgraph from the ALZKB database in order to reduce noise and focus the learning process on disease-relevant entities.

• **Feature Engineering and Representation Learning:** Applying graph representation learning techniques, such as DeepWalk, to generate dense vector embeddings that capture the structural and semantic properties of biomedical entities within the constructed subgraph.

•**Graph Neural Network Modeling:** Implementing and training advanced GNN architectures including RGCN and RGAT with focal loss for multi-class drug-gene interaction prediction. The models utilize both structural embeddings and topological features to learn complex relational patterns.

•**Ensemble Learning and Evaluation:** Combining RGCN and RGAT predictions using weighted ensemble (0.7 RGCN + 0.3 RGAT) to optimize classification performance. Models are evaluated using accuracy, macro-AUC, and F1-score with early stopping (patience=50) to prevent overfitting.

•**LLM Integration for Explainability:** Incorporating Large Language Models to generate scientific explanations for predicted interactions, providing biological context and interpretation of computational results in natural language format.

•**Interactive System Development:** Implementing a web-based interface using Streamlit for real-time prediction, visualization, and report generation. The system supports bilingual functionality (Arabic/English) and PDF report export for comprehensive analysis.

1.7 Thesis Structure

This thesis is organized into five main chapters:

Chapter	Title	Primary Focus
Chapter One	**Introduction**	Establishes the research motivation, problem statement, objectives, research gaps, and the scope of the work.
Chapter Two	**Related Work and Theoretical Background**	Reviews fundamental concepts, including Knowledge Graphs, Graph Representation Learning, Graph Neural Networks, and Generative Artificial Intelligence, and critically analyzes existing literature on graph-based drug repurposing for Alzheimer's disease.
Chapter Three	**Methodology**	Details the implementation steps of the proposed framework, including data acquisition, AD-centric subgraph construction, feature engineering, and the design of graph-based link prediction models.
Chapter Four	**Experimental Setup and Results**	Presents experimental design, hyperparameter settings, evaluation metrics (accuracy, macro-AUC, F1-score), and quantitative results obtained from RGCN/RGAT ensemble models, including confusion matrices and class-wise performance analysis.
Chapter Five	**Conclusion and Future Work**	Summarizes the contributions of the thesis, discusses limitations, and outlines future directions, including the integration of generative language models for biological explanation and hypothesis generation.

Chapter Two

Related Work and Theoretical Background

2.1 Fundamental Scientific Concepts

2.1.1 Pathophysiology and Key Drug Targets in Alzheimer's Disease

This section discusses the primary molecular and cellular targets central to AD therapeutic interventions. These include targets related to the **Amyloid Precursor Protein (APP)** processing, **Tau protein hyperphosphorylation**, genetic risk factors like **APOE** and **PSENs**, and inflammatory pathways often associated with neurodegeneration. Understanding these mechanisms justifies the adoption of a network-based approach capable of targeting multiple, interacting biological entities rather than single targets.

2.1.2 The Role of Knowledge Graphs (KGs) in Biomedical Data Integration

Knowledge Graphs (KGs) play a crucial role in modeling and integrating the complex and heterogeneous nature of biomedical data. They represent biomedical entities—such as drugs, genes, proteins, diseases, and biological pathways—as nodes, while their curated relationships—such as drug–gene interactions, gene–disease associations, and gene–pathway memberships—are encoded as typed and directed edges.

Heterogeneous Knowledge Graphs (HKGs) are particularly important in biomedical research, as they enable the integration of diverse data modalities originating from multiple sources, including genomic data, chemical interactions, and disease annotations, into a unified and structured representation. This graph-based paradigm surpasses the limitations of traditional relational databases by explicitly preserving the relational and topological dependencies between entities, thereby providing a suitable foundation for downstream graph-based learning tasks such as representation learning and link prediction.

2.1.3 Graph Representation Learning (GRL) and DeepWalk

Graph Representation Learning (GRL) involves generating low-dimensional, dense vector representations (embeddings, $\mathbf{z} \in \mathbb{R}^d$) for nodes in a graph. These embeddings mathematically encode the node's position, role, and structural context. **DeepWalk**, a foundational GRL method, uses **simulated truncated random walks** on the graph to generate node sequences, which are then treated as sentences and processed using the **Word2Vec (Skip-Gram)** model. This approach effectively translates the non-Euclidean graph topology into a vector space where computational models can efficiently perform tasks like classification and link prediction.

2.1.4 Generative AI Models

Generative AI models, particularly Large Language Models (LLMs) such as Google's Gemini series (Gemini-2.5-Flash, Gemini-2.5-Pro), have emerged as transformative tools in biomedical research and drug discovery. These models leverage vast pre-trained knowledge to generate coherent scientific explanations, interpret complex biological mechanisms, and bridge the gap between computational predictions and human understanding.

In the context of drug repurposing for Alzheimer's disease, LLMs serve three critical functions: (1) **Interpretability Enhancement** - converting numerical predictions from graph neural networks into scientifically rigorous, human-readable explanations that incorporate biological context from knowledge graphs; (2) **Hypothesis Generation** - synthesizing therapeutic candidates from biomedical literature through iterative prompting and evidence synthesis; and (3) **Clinical Translation** - providing mechanistic explanations that support the transition from computational predictions to experimental validation.

The integration of LLMs with graph-based drug-gene interaction predictions represents a paradigm shift toward explainable AI in precision medicine, where model transparency and biological plausibility are essential for clinical adoption and regulatory approval.

2.2 Review of Related Studies

2.2.1 Gogineni (2021): Analysis of Drug Repurposing Knowledge Graphs for Covid-19 [1]

This landmark work provided the community with the **Drug Repurposing Knowledge Graph (DRKG)**, aiming to create a standardized, large-scale resource for drug repurposing research.

- **Data Foundation**

  Scale and Diversity: This study introduced **DRKG**, a massive knowledge graph comprising 97,238 nodes and 5,874,261 edges.
  Entity Interaction: It includes 13 entity types and 107 relationship types, capturing complex biological interactions as illustrated in **Figure 2.2**.
  Integration: Data were harmonized from six major databases, including **DrugBank**, **String**, and **KEGG**.

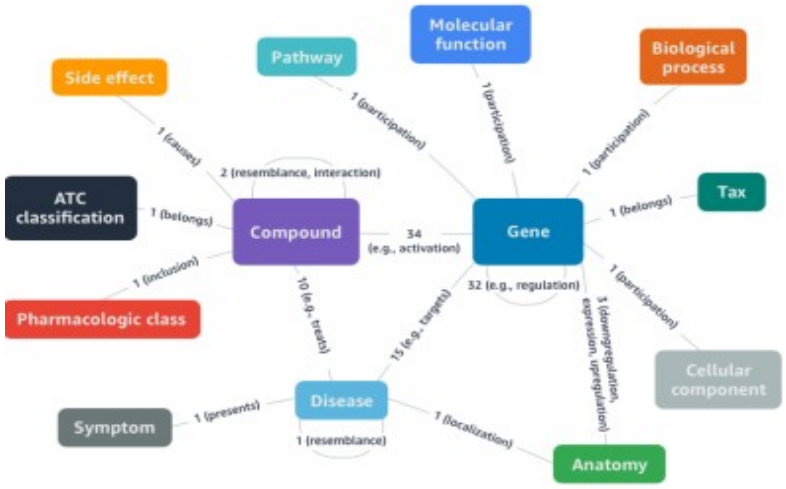

Figure 2.2: Semantic schema of the Drug Repurposing Knowledge Graph (DRKG) illustrating entity-relation interactions.

The diagram depicts the 13 types of biological entities (nodes) and their interconnectedness within the DRKG framework. The central nodes include Compounds, Genes, and Diseases, while peripheral nodes represent biological processes, pathways, and symptoms. The numerical values on the edges indicate the specific number of distinct relationship types existing between the corresponding entity pairs.

- **Technical Methodology**

The study employs Knowledge Graph Embedding (KGE) to project biological entities into a d-dimensional vector space for link prediction:

1. **Model Benchmarking:** Four architectures were evaluated:
    - **TransE:** Translational logic where h + r \approx t.
    - **DistMult:** Bilinear modeling using diagonal matrices.
    - **ComplEx & RotatE:** Complex-space modeling to capture asymmetric, inverse, and rotational biological patterns.
2. **Training:** Utilized **Self-Adversarial Negative Sampling** to distinguish factual triples from "corrupted" associations.
3. **Evaluation:** Performance was measured via **Mean Reciprocal Rank (MRR)** and **Hits@k** (k=1, 3, 10) to rank true therapeutic links.

- **Key Findings and Results**

- **Superiority: RotatE** and **ComplEx** outperformed TransE, proving that capturing asymmetric relations is vital for drug discovery.
- **Graph Density:** Higher edge density and quality node representations were identified as the primary drivers of accuracy.
- **Impact:** DRKG establishes a foundational benchmark for comparing future Graph Neural Network (GNN) architectures against established KGE baselines.

2.2.2 Hao et al. (2023): Deep Learning for Alzheimer's Disease Drug Repurposing using Knowledge Graph and Multi-level Evidence [2]

- **Data Foundation:**

The study constructed a specialized AD-centric heterogeneous knowledge graph by integrating diverse datasets, including DrugBank, CTD, and Agora. The graph architecture consisted of:

1. **Nodes:** 6,543 FDA-approved or experimental drugs, 16,997 genes, and AD as a central disease entity.
2. **Edges:** Relationships including drug-target interactions (DTI), gene-disease associations, and protein-protein interactions (PPI).
3. **Transfer Learning:** Pre-trained embeddings from the **DRKG** (Drug Repurposing Knowledge Graph) were utilized to initialize node features, mitigating the sparsity of AD-specific data.

- **Technical Methodology**

The study implemented a Multi-relational Variational Graph Autoencoder (VGAE) framework.

Figure 2.1: Methodological workflow of the ADKG framework (Hao et al., 2021).

- **Encoder:** A Graph Convolutional Network (GCN) learned the latent space distribution $q(Z|X, A)$ based on graph topology and node features.
- **Multi-task Learning:** Simultaneously optimized for **Link Prediction** (missing drug-gene edges) and **Node Classification** (AD-relevant gene categorization).
- **Clinical Validation:** Integrated transcriptomic evidence with **Real-World Data (RWD)** from insurance claims involving 7 million patients to rank candidates.

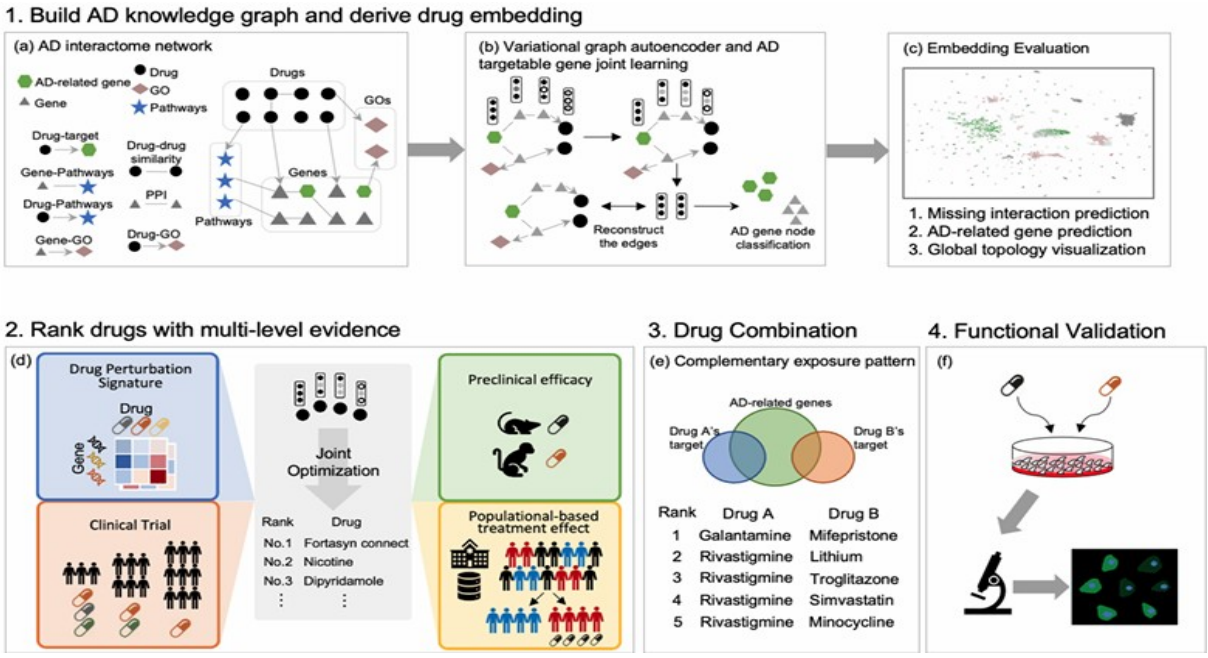

Figure 2.1:Methodological workflow of the ADKG framework, illustrating the integration of knowledge graph embeddings with multi-level clinical evidence for drug repurposing

The figure illustrates the four-stage pipeline developed by Hao et al. (2021): (1) Graph Construction and Embedding: Building an AD-centric interactome and utilizing a Variational Graph Autoencoder (VGAE) for joint learning of drug and gene representations. (2) Multi-level Evidence Ranking: Drugs are prioritized by integrating drug perturbation signatures, clinical trial data, and population-based treatment effects. (3) Drug Combination Identification: Selection of synergistic drug pairs based on complementary exposure patterns in the protein interactome. (4) Functional Validation: Experimental verification of top candidates (e.g., Galantamine and Mifepristone) using in vitro neuronal cell models.

- **Key Findings and Results:**

➢ The model achieved a high predictive accuracy with an **AUROC of 0.991** for the link prediction task.

➢ Several high-priority drug candidates were identified, most notably the combination of **Galantamine** and **Mifepristone**.

➢ **In vitro validation:** Testing on HT22 mouse hippocampal neurons demonstrated that this combination significantly reduced glutamate-induced cell death, providing biological evidence for the computational predictions.

2.2.3 Ratajczak et al. (2022): Task-driven Knowledge Graph Filtering Improves Prioritizing Drugs for Repurposing [3]

- **Data Foundation:** The study utilized two large biomedical knowledge graphs: Hetionet (**45,158** entities, **2,249,807** triples) and DRKG (**96,121** entities, **5,874,261** triples). These graphs incorporate diverse data from sources such as **DrugBank** and **STRING** but often include "noise" irrelevant to the drug repurposing task (e.g., relations distant from Compound-Disease).
- **Technical Methodology:** The researchers proposed a task-driven filtering method using metapaths (predefined sequences of relation types, e.g., Compound → Gene → Disease). Random walks are performed along these metapaths starting from Compound or Disease nodes, and complete walks are concatenated to retain only relevant entities and relations. Knowledge Graph Embedding (KGE) models (TransE, RESCAL, DistMult, ComplEx, ConvE) were trained on the filtered and original graphs for "treats" link prediction.
- **Key Findings:** Filtering reduced Hetionet by **60%** in entities and **33%** in triples, and DRKG by **26%** and **19%**, respectively, while increasing neighborhood density for target nodes. Performance improved significantly: average MRR increase of **20.6%** in Hetionet (up to **40.8%**) and **8.9%** in DRKG. ComplEx was the best-performing model. Case studies on SARS-CoV-2 and cancer types showed better ranking of repurposed drugs, demonstrating that "more data is not always better" in specific tasks like drug repurposing for neurodegenerative diseases.

2.2.4 Bang et al. (2023): Biomedical Knowledge Graph Learning for Drug Repurposing by Extending Guilt-by-Association to Multiple Layers [4]

- **Data Foundation:** Multiple multi-scale biomedical knowledge graphs were used, including **MSI**, **HetioNet**, and **KEGG**, focusing on nodes for drugs, diseases, genes, and pathways, with integration of semantic information (similarity) for drugs and diseases.
- **Technical Methodology:** The DREAMwalk model employs semantically guided random walks with a "teleport" process; upon reaching a drug or disease node, the walker transitions to semantically similar nodes based on a similarity matrix. Balanced node sequences are generated and embedded using a heterogeneous Skip-gram model, followed by drug-disease association classification with XGBoost.
- **Key Findings:** DREAMwalk outperformed state-of-the-art models by up to **16.8%** in accuracy, with average AUROC of **0.938** and AUPR of **0.939** across the graphs. In the Alzheimer's Disease case, it proposed candidates such as levetiracetam (an antiepileptic that improves spatial memory), fluoxetine, and sertraline (antidepressants linked to AD), supported by literature and phase 3 clinical trials (e.g., caffeine and escitalopram). The method provides interpretability through gene pathways, enhancing confidence in repurposing.

2.2.5 Romano et al. (2024): The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research [5]

- **Data Foundation:** AlzKB is a large-scale, heterogeneous knowledge graph specifically tailored for Alzheimer's disease (AD), comprising **118,902** distinct entities (e.g., **62,407** genes, **35,063** drugs, **11,381** biological processes, **4,570** pathways) and **1,309,527** relationships. It integrates data from **22** diverse public sources, including DrugBank, DisGeNET, Hetionet, Reactome, Gene Ontology, and GWAS Catalog. A Web Ontology Language 2 (OWL 2) ontology enforces semantic consistency, enabling ontological inference and validation via reasoners like **FaCT++**.
- **Technical Methodology:** The graph is implemented in Neo4j using neosemantics for RDF/XML import. For machine learning applications, topological features (e.g., common neighbors, Adamic-Adar) are extracted for a random forest classifier in genetic target prediction (**80/20** train-test split, 3-fold cross-validation). Drug repurposing employs knowledge graph embedding models (TransE, RotatE, DistMult, ComplEx, ConvE) via PyKEEN, trained with 256-dimensional embeddings, **100** epochs, and early stopping on an **80/10/10** split, evaluated by **Hits@k** and Mean Reciprocal Rank (**MRR**).
- **Key Findings:** The random forest achieved **96.2%** balanced accuracy (precision **0.88**, recall **0.98**) in predicting shared AD-PD genes, identifying **8** novel candidates (e.g., SNCA, FYN, SYNJ1). RotatE outperformed others (**Hits@10=0.358**, **MRR=0.202**), proposing top repurposing candidates including sumatriptan, nicotine, pimozide, risperidone, flurbiprofen, and sertraline; **30%** of these (nicotine, risperidone, sertraline) have been investigated in AD clinical trials. Mechanistic explanations via spanning trees link drugs to AD through known gene pathways, demonstrating AlzKB's utility in generating biologically plausible, replicable therapeutic hypotheses.

2.2.6 Loesch et al. (2024): Explaining Graph Neural Network Predictions for Drug Repurposing [6]

- **Data Foundation:** The study utilizes the Drug Repurposing Knowledge Graph (DRKG), containing approximately **97,000** entities (compounds, genes, pathways, diseases) and over 5 million edges across **107** relation types, sourced from databases like STRING, DrugBank, Hetionet, and IntAct. For the AD case study, a disease-filtered subgraph retains only triples involving Alzheimer's disease.
- **Technical Methodology:** Three Graph Neural Network variants (GCN, GraphSAGE, GAT) are trained for link prediction (Compound treats Disease) with 100-dimensional embeddings, Adam optimizer, and **500** epochs on an **80%** train-test split. Interpretability is achieved via gradient-based Saliency Maps, computing absolute gradients of output with respect to input embeddings to rank node importance. Top-k ranked triples form explanatory subgraphs, focusing on influential genes and pathways.
- **Key Findings:** GAT achieved the highest Hits@5 (**0.451**) and **Hits@10** (**0.672**), while GraphSAGE excelled in recall (**0.992**), minimizing false negatives crucial for repurposing. In the AD case study using GraphSAGE, explanations for Donepezil highlighted bindings to acetylcholinesterase (AChE) and butyrylcholinesterase (BChE), aligning with its role in enhancing cholinergic function by inhibiting acetylcholine breakdown. For Memantine, subgraphs emphasized glutamate receptors (e.g., GRIN1), consistent with its NMDA antagonism reducing excitotoxicity. These literature-supported explanations enhance transparency and clinical trust in GNN-based predictions.

2.2.7 Leveraging Generative AI to Prioritize Drug Repurposing Candidates for Alzheimer's Disease with Real-World Clinical Validation [7]

- **Data Foundation:** Generative AI (GPT-4, trained up to September 2021) synthesized candidates from vast literature via iterative prompting, yielding **59** unique drugs. Validation used de-identified electronic health records from Vanderbilt University Medical Center (VUMC; >3 million patients) and the All of Us Research Program (>**235,000** participants ≥**65** years), standardized to OMOP Common Data Model, with AD diagnosis via ICD codes (**94%** positive predictive value in VUMC).
- Technical Methodology: Ten independent GPT-4 sessions generated ranked lists of **20** promising non-AD-specific drugs in JSON format, excluding approved AD treatments. Top 10 frequent candidates (≥**7** mentions) underwent retrospective cohort analysis: exposure defined by ≥**1** record before age **65**; propensity score matching (**2:1**, caliper **0.1**) on demographics and comorbidities; Cox regression for AD risk over **10**-year follow-up (censored at age **75**).
- Key Findings: Meta-analysis identified metformin (HR=**0.67**, **95%** CI **0.55–0.81**), simvastatin (HR=**0.84**, **95%** CI **0.73–0.98**), and losartan (HR=**0.76**, **95%** CI **0.60–0.95**) with significantly reduced AD risk. Directional benefits were observed for other candidates like pioglitazone. This low-cost approach demonstrates generative AI's ability to prioritize repurposing candidates from scientific literature, bridging computational hypotheses with real-world evidence for further clinical trials.

2.2.8 Dobreva et al. (2025): A Unified Framework for Alzheimer's Disease Knowledge Graphs: Architectures, Principles, and Clinical Translation [8]

- **Data Foundation:** This systematic review synthesizes AD knowledge graphs integrating multimodal sources: literature (via LLM-augmented NLP), structured databases (e.g., STRING, DrugBank), neuroimaging (ADNI/OASIS), clinical records, and omics data.
- **Technical Methodology:** The proposed **AD-KG 2.0** is a modular, adaptive four-layer framework: Data Layer (tailored ingestion pipelines), Integration Layer (ontology-driven and LLM-based semantic alignment), Embedding & Reasoning Layer (complex embeddings, multimodal/self-explainable GNNs), and Application Layer (customized for repurposing, stratification, progression modeling). Features include decision-tree-guided adaptation, multi-tiered terminology harmonization (**87–94%** alignment), and periodic updates (automated surveillance, expert curation).
- **Key Findings:** Reviewed methods achieve high performance, e.g., LLM extraction F1 **0.78–0.89**; RotatE embeddings AUROC **0.97**; multimodal GNNs **89.6%** accuracy in classification; clinical decision support ≥**81.7%** accuracy. AD-KG 2.0 addresses heterogeneity, interpretability (e.g., **7.3%** gain via self-explainable GNNs), and translation challenges (privacy via federated learning), supporting extensible applications across neurodegenerative diseases while advocating interdisciplinary collaboration for regulatory compliance.

2.2.9 Selote and Makhijani (2025): A Knowledge Graph Approach to Drug Repurposing for Alzheimer's, Parkinson's and Glioma using Drug–Disease–Gene Associations [9]

- **Data Foundation:** The study constructs a heterogeneous, multi-relational knowledge graph integrating drug–disease–gene association data from public resources, focusing on neurodegenerative diseases (Alzheimer's disease [AD], Parkinson's disease [PD]) and glioma. The graph features nodes representing drugs, diseases, and genes, connected by four distinct relationship types, emphasizing shared pathophysiological mechanisms such as neuron loss, motor decline, and neuroglial stem cell involvement.
- **Technical Methodology:** The knowledge graph is implemented in Neo4j for efficient storage and querying. Node embeddings are generated using the scalable feature learning algorithm node2vec, with random walk parameters optimized (return parameter $\mathbf{p \geq 2}$ to encourage outward exploration from the source node, and in-out parameter $\mathbf{q \leq 1}$ to balance depth-first and breadth-first sampling). Unknown drug-disease associations are predicted by computing cosine similarity between disease and drug node embeddings, prioritizing high-similarity pairs as repurposing candidates.
- **Key Findings:** The approach yielded definitive sets of candidate drugs for repurposing in AD, PD, and glioma, leveraging the graph's ability to capture indirect correlations through gene-mediated links. Candidates were rigorously validated against existing literature, ranked using the CodReS online tool (which assesses repurposing potential based on semantic similarity and evidence), and cross-verified in pharmaceutical knowledge databases for clinical status and biological significance. This in silico method highlights the efficiency of graph-based feature learning in addressing data challenges for complex, multi-indication neurodegenerative disorders, offering a simple yet scalable pipeline for hypothesis generation.

2.2.10 Wang et al. (2025): Drug Repurposing for Alzheimer's Disease using a Graph-of-Thoughts based Large Language Model to Infer Drug-Disease Relationships in a Comprehensive Knowledge Graph [10]

• **Data Foundation:** The study compares a general drug repurposing knowledge graph (utilized via TxGNN) with the Alzheimer's KnowledgeBase (AlzKB), an AD-specific ontology-based knowledge graph encompassing over **234,000** entities (including genes, pathways, drugs, diseases, body parts, drug classes, and transcription factors) and extensive multi-hop relationships curated for AD relevance.

• Technical Methodology: Five distinct strategies were evaluated: DR1 (machine learning with TxGNN on general KG), DR2 (RotatE embeddings on AlzKB; hyperparameters: **40** epochs, batch size **256**, embedding dimension **960**), DR3 (LLM-based chatbot translating natural language to Cypher queries on AlzKB), DR4 (ESCARGOT framework: Graph-of-Thoughts [GoT]-enhanced LLM generating executable Python workflows integrated with AlzKB, incorporating multi-step filtering for 2-hop/3-hop paths), and DR5 (iterative prompting with general LLM, e.g., GPT-4). ESCARGOT's filtering includes: **(1)** pathway or body part connections to AD, **(2)** gene-drug interaction thresholds exceeding known AD drugs, **(3)** links via transcription factors or drug classes; 3-hop paths restrict genes to those associated with approved AD treatments.

• Key Findings: Evaluated against a benchmark of **573** previously reported AD repurposing candidates, AlzKB-based methods outperformed general KGs. ESCARGOT (DR4) achieved the highest overlap coefficients (**0.846** for intersection of **13** drugs, **0.783** for union of **46** drugs in 2-hop paths; **~3** min computation on standard hardware). It rediscovered known drugs like donepezil (acetylcholinesterase inhibition for synaptic enhancement), minocycline (anti-inflammatory, reduces Aβ/tau pathology), retinoids (neuroprotection against plaques/phosphorylation), and tamoxifen (tau regulation via CDK5 inhibition). Novel proposals included epirubicin, vemurafenib, fulvestrant (antineoplastics targeting neuroinflammation via microglia), and vitamin A (modulates Aβ metabolism, inflammation, and cognition). This accessible, expertise-reducing framework accelerates AD drug discovery by bridging LLMs and disease-specific KGs, with broad adaptability to other indications.

2.3 Comparative Analysis and Synthesis

This table provides a concise comparison of the key methodologies, data reliance, and outcomes of the reviewed literature, establishing the context for this thesis's unique approach.

Study (Year)	Primary Computational Method	KG Data Scope / Source	Key Result / Finding	Comparative Focus
Gogineni (2021)	Knowledge Graph Embeddings (KGE: RESCAL, RotatE, TransE, DistMult, ComplEx, CompGCN)	DRKG (97,238 entities, multiple biomedical databases)	RESCAL best performance (MRR 0.66); identified candidates like Colchicine with clinical trial support.	Benchmarking KGE models on large-scale KG for repurposing in sparse data scenarios.
Hao et al. (2023)	Variational Graph Autoencoder (VGAE) with GCN	AD-centric heterogeneous KG (DrugBank, CTD, etc.) + RWD	AUROC 0.991; Galantamine-Mifepristone combination validated in vitro.	Multi-task deep learning on disease-specific KGs with real-world evidence integration.
Ratajczak et al. (2022)	Task-driven metapath filtering + KGE (TransE, ComplEx, etc.)	Hetionet and DRKG	Up to 40.8% MRR improvement after filtering; better prioritization for antivirals.	Reducing KG noise through task-specific subgraph extraction.
Bang et al. (2023)	DREAMwalk (semantically guided random walks + XGBoost)	Multi-scale KGs (MSI, HetioNet, KEGG)	Up to 16.8% accuracy gain (AUROC 0.938); AD candidates like levetiracetam with literature support.	Extending guilt-by-association with semantic teleport for interpretable embeddings.
Romano et al. (2024)	Random Forest + KGE (RotatE best)	AlzKB (118,902 entities from 22 sources)	96.2% accuracy in gene prediction; repurposing candidates (e.g., nicotine) with mechanistic paths.	Disease-specific KG construction and ML for hypothesis generation.
Loesch et al. (2024)	Graph Neural Networks (GAT, GraphSAGE) + Saliency Maps	DRKG	GAT Hits@10 0.672; biologically plausible explanations for AD drugs (e.g., Donepezil-AChE).	Post-hoc interpretability in GNN predictions for clinical trust.
Yan et al. (2024)	Generative AI (GPT-4) + retrospective cohort analysis	Literature-synthesized candidates + EHR (VUMC, All of Us)	Reduced AD risk for metformin (HR 0.67), simvastatin, losartan.	Bridging generative AI hypotheses with large-scale real-world validation.
Dobreva et al. (2025)	Proposed unified framework (AD-KG 2.0: LLM, GNNs, embeddings)	Multimodal synthesis (ADNI, literature, databases)	AUROC up to 0.97; addresses heterogeneity and translation challenges.	Systematic review and adaptive architecture for AD KG applications (review).
Selote and Makhijani (2025)	node2vec embeddings + cosine similarity	Drug-disease-gene KG (public associations, Neo4j)	Definitive repurposed candidate sets for AD/PD/glioma, validated via literature/CodReS.	Scalable graph embedding for multi-disease neurodegenerative repurposing.
Wang et al. (2025)	Graph-of-Thoughts LLM (ESCARGOT)	AlzKB (AD-specific) vs. general KGs	Highest overlap 0.846; rediscovered known (donepezil) and novel (epirubicin) candidates.	LLM-KG integration with multi-hop filtering to reduce expertise barriers.

2.4 Summary of Research Gaps

Despite significant advances in graph-based drug repurposing for Alzheimer's disease, several critical gaps remain unaddressed in the current literature:

1. Multi-Class Interaction Type Modeling: Most existing studies focus on binary link prediction (treats/does-not-treat), lacking the granular ability to distinguish between specific drug-gene interaction mechanisms such as binding, expression upregulation, and expression downregulation. This limitation restricts the biological interpretability and clinical applicability of predictions.

2. Class Imbalance in Biological Networks: Few approaches adequately address the inherent class imbalance in multi-class drug-gene interaction prediction, where certain interaction types (e.g., expression modulation) are significantly underrepresented compared to binding interactions, potentially biasing model performance toward majority classes.

3. Ensemble Learning for Heterogeneous GNN Architectures: While individual GNN models (RGCN, RGAT) have been explored, there is limited research on ensemble approaches that combine complementary strengths of different relational graph neural network architectures with optimized weighting strategies for drug repurposing tasks.

4. Advanced Loss Functions for Biomedical Applications: The application of focal loss and class weighting strategies specifically designed for imbalanced multi-class biological interaction prediction remains underexplored, despite their proven effectiveness in addressing rare class detection.

5. Real-Time LLM-Enhanced Interpretability: Current explainability approaches rely primarily on post-hoc analysis (saliency maps, attention weights) without integrating large language models for generating contextual, metadata-enriched explanations that incorporate domain-specific biological knowledge from knowledge graphs.

6. Interactive Clinical Decision Support Systems: Most computational frameworks lack user-friendly, real-time interfaces that enable researchers and clinicians to explore predictions, receive AI-generated explanations, and generate comprehensive reports for further experimental validation.

7. Confidence-Based Conservative Prediction: Limited attention has been given to implementing confidence thresholding mechanisms that prioritize precision over recall for sensitive biological predictions, particularly important when false positives could mislead expensive experimental validation efforts.

This research addresses these gaps through a unified framework that combines ensemble relational graph neural networks (RGCN + RGAT) with multi-class focal loss, LLM-based explanations using Gemini models, and an interactive Streamlit interface, specifically designed for multi-class drug-gene interaction prediction in Alzheimer's disease with enhanced interpretability and clinical translation potential.

Chapter Three

Methodology and Implementation

3.1 Data Source

3.1.1 AlzKB Knowledge Graph

This study is based on the AlzKB biomedical knowledge graph, which integrates heterogeneous evidence relevant to Alzheimer's disease, including disease–gene associations, drug–gene relations, functional gene annotations, and additional biomedical relations. The graph contains **234,037 nodes** and **1,668,487 relationships** distributed across **19 relation types**.

The drug–gene relations of primary interest in this work are:

•**CHEMICALBINDSGENE**: 25,726 edges

•**CHEMICALINCREASESEXPRESSION**: 18,713 edges

•**CHEMICALDECREASESEXPRESSION**: 21,051 edges

These relations provide the ground-truth positive evidence used to construct supervised learning datasets for **drug–gene** relationship prediction.

3.1.2 Graph Storage and Querying

The knowledge graph was hosted in **Memgraph** and accessed through the **Cypher** query language. Cypher queries were used to:

•Retrieve relation statistics (counts per relation type)

•Identify Alzheimer-relevant disease nodes

•Extract Alzheimer-associated genes

•Retrieve drugs connected to those genes through drug–gene relations

•Collect metadata (drug classes and gene functional annotations) to support interpretation

3.1.3 Computational Environment

All implementation was performed in **Python 3.10.11**. Deep learning components were implemented using **PyTorch 2.5.1 (CUDA 12.1)** and **PyTorch Geometric** for graph neural network operations. Data handling and evaluation utilities used standard scientific Python libraries (e.g., NumPy, pandas, scikit-learn). The graph database layer was managed with Memgraph (Docker-based deployment) and queried programmatically from Python.

3.2 Alzheimer-Centered Subgraph Construction and Representation Learning

Both methodological pipelines rely on an Alzheimer-focused subgraph extracted from the full knowledge graph. The objective of this step is to restrict learning and inference to a domain-relevant region while preserving graph structure that captures biological context.

3.2.1 Entity Filtering Strategy

A targeted filtering approach was used to construct an Alzheimer-centered node set:

1. **Disease selection**

Disease nodes were selected by matching textual attributes containing the term "Alzheimer" (case-insensitive), yielding **34 disease nodes**.

2. **Gene selection**

Genes connected to these diseases through disease–gene association relations were collected, yielding **103 Alzheimer-associated genes**.

3. **Drug selection**

Drugs connected to the selected genes through drug–gene relations (binding or expression modulation) were extracted, yielding **861 drugs**.

The resulting Alzheimer-centered node set contains **998 nodes** in total (34 diseases + 103 genes + 861 drugs).

3.2.2 Subgraph Synthesis and Edge Filtering

After selecting the Alzheimer-centered node set, graph edges were filtered to retain only edges whose endpoints lie within the selected node set. This produced an Alzheimer-centered subgraph with:

- **998 nodes**
- A filtered edge list among these nodes
- A simplified undirected NetworkX representation with **2,111 unique edges**

This subgraph is used as the basis for random-walk embedding learning and as the structural backbone for the graph neural models.

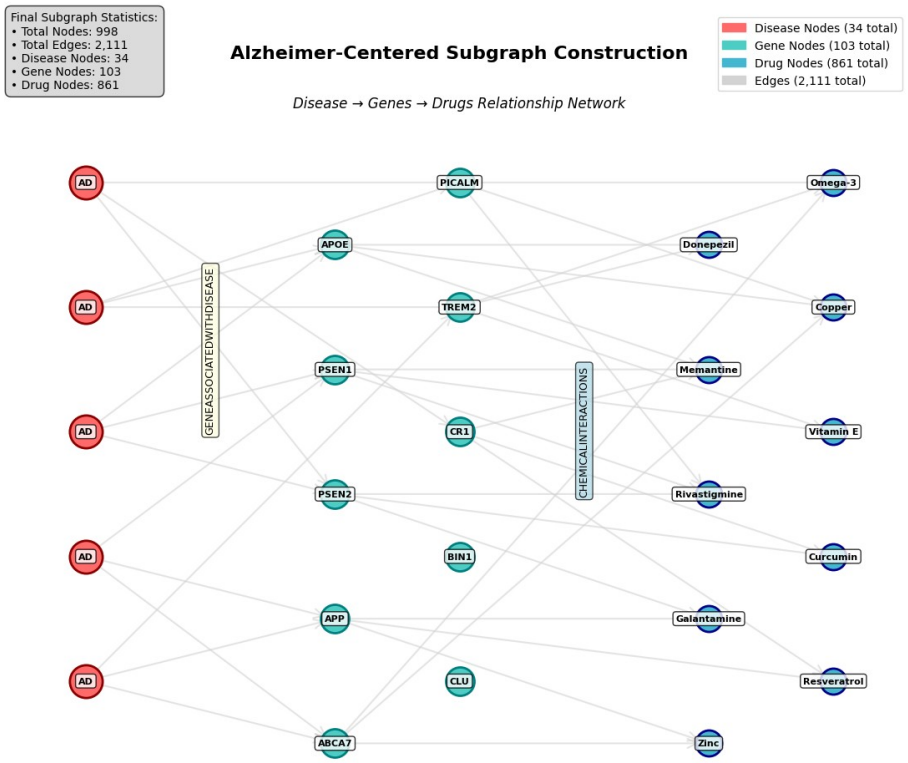

Figure 3.1: Schematic summary of the Alzheimer-centered subgraph construction showing disease-gene-drug relationships and final node/edge counts

Schematic representation of the Alzheimer-centered subgraph architecture.
The visualization demonstrates the three-tier structure: disease nodes (red), gene nodes (teal), and drug nodes (blue), with their respective connection patterns. Edge types include GENEASSOCIATEDWITHDISEASE (disease-gene) and CHEMICALINTERACTIONS (drug-gene). This subgraph serves as the foundation for both embedding generation and graph neural network modeling in our approach.

3.2.3 Structural Node Embeddings

To learn vector representations of graph nodes that capture structural context, **DeepWalk** was applied to the Alzheimer-centered subgraph. DeepWalk generates random walks and learns embeddings using a skip-gram objective over the walk sequences.

Configuration used:

•Embedding size: **128**

•Walk length: **80**

•Number of walks per node: **10**

•Context window size: **10**

•Skip-gram training objective

The resulting **128-dimensional** embedding is learned for each node and serves as a general-purpose structural representation used by downstream predictors.

3.2.4 Topological Features (Degree and PageRank)

Topological features were computed to complement learned representations:

•**Degree**: used as a structural indicator of local connectivity and hubness. Degree was computed and normalized and can be integrated as an additional feature.

•**PageRank**: computed and integrated specifically in **Methodology II** (the final GNN pipeline) as a global importance feature.

This separation is intentional: the final pipeline integrates PageRank into the node feature matrix to enrich the GNN encoder with a global centrality signal.

3.3 Methodology I: Binary Link Prediction Framework

Methodology I establishes a foundational approach for drug-gene interaction prediction by formulating the problem as binary classification. This methodology focuses on determining the existence of any interaction between drug-gene pairs without distinguishing between specific interaction mechanisms.

3.3.1 Problem Formulation

The binary classification framework categorizes each drug-gene pair into one of two classes:

• Class 0: No significant interaction

• Class 1: Interaction exists (encompassing binding, expression increase, or expression decrease)

This simplified formulation provides a robust baseline for evaluating the feasibility of graph-based drug-gene interaction prediction while reducing the complexity inherent in multi-class classification tasks.

3.3.2 Dataset Construction Strategy

All observed drug-gene relationships from the three primary interaction types (CHEMICALBINDSGENE, CHEMICALINCREASESEXPRESSION, CHEMICALDECREASESEXPRESSION) were consolidated into a single positive class. This approach maximizes the available positive evidence while maintaining a clear distinction between interacting and non-interacting pairs.

Negative samples were generated through systematic random sampling of drug-gene pairs with no documented interactions in the AlzKB knowledge graph, ensuring balanced representation across both classes.

3.3.3 Graph Neural Network Architecture

The binary classification employs advanced graph neural network architectures, specifically Relational Graph Convolutional Networks (RGCN) and Relational Graph Attention Networks (RGAT). Both models utilize:

• Node feature representations combining DeepWalk embeddings with topological features
• Standard architectural components including dropout regularization and layer normalization
• Residual connections for improved gradient flow

3.3.4 Training Objective and Loss Function

The binary classification employs cross-entropy loss optimization:

$$L = -\sum [\, y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \,]$$

Where:

• y_i represents the true binary label

• p_i represents the predicted probability

Training utilizes the Adam optimizer with early stopping mechanisms based on validation performance to ensure optimal generalization.

3.3.5 Ensemble Strategy

The methodology combines predictions from both RGCN and RGAT architectures through ensemble averaging, leveraging the complementary strengths of both approaches to improve prediction robustness.

3.4 Methodology II: Multi-Class Interaction Type Prediction Framework

Building upon the binary baseline, Methodology II addresses the clinically critical task of predicting specific drug-gene interaction mechanisms. This approach provides detailed mechanistic insights essential for drug discovery and therapeutic development applications.

3.4.1 Enhanced Problem Formulation

The multi-class framework extends the prediction task to four distinct categories:

• Class 0: No significant interaction

• Class 1: Direct binding interaction (CHEMICALBINDSGENE)

• Class 2: Expression upregulation (CHEMICALINCREASESEXPRESSION)

• Class 3: Expression downregulation (CHEMICALDECREASESEXPRESSION)

This formulation preserves the mechanistic specificity of drug-gene interactions, enabling more precise therapeutic predictions.

3.4.2 Dataset Characteristics and Class Distribution

The multi-class dataset comprises 5,809 labeled drug-gene pairs distributed across training (4,066 samples), validation (581 samples), and test (1,162 samples) sets. The dataset exhibits severe class imbalance, with expression-direction classes representing minority categories, reflecting the natural distribution of documented drug-gene interactions in biomedical literature.

3.4.3 Enhanced Feature Engineering

Multi-class prediction necessitates enhanced feature construction through edge-level representation learning. Each drug-gene pair is represented by concatenating the 128-dimensional DeepWalk embeddings of both entities, resulting in 256-dimensional feature vectors. Additional topological features including normalized degree centrality and PageRank scores are integrated to capture network-level importance.

3.4.4 Advanced Graph Neural Network Architecture

Methodology II employs sophisticated GNN architectures with the following enhancements:

Advanced RGCN Implementation:

- Hidden dimensionality of 256 units across two convolutional layers
- Residual connections for improved gradient flow
- Layer normalization for training stability
- Dropout regularization (p=0.3) to prevent overfitting
- GELU activation functions for enhanced non-linearity

Advanced RGAT Implementation:

- Multi-head attention mechanism with four attention heads
- Dynamic neighbor weighting through learned attention coefficients
- Similar architectural enhancements to RGCN with attention-based aggregation

3.4.5 Class Imbalance Mitigation Strategy

To address the severe class imbalance, Methodology II implements Multi-Class Focal Loss with balanced class weighting. The focal loss formulation addresses class imbalance through:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

Where:

- p_t is the predicted probability of the true class
- α_t represents the class weight for balanced training
- γ is the focusing parameter (set to 2)

The focal loss mechanism down-weights easily classified examples while focusing learning on challenging cases, particularly benefiting minority classes. Class weights are computed using balanced weighting strategies to ensure equitable representation during training.

3.4.6 Ensemble Strategy and Conservative Prediction

The final prediction system combines RGCN and RGAT outputs through weighted ensemble averaging:

$$P_{final} = 0.7 \times P_{RGCN} + 0.3 \times P_{RGAT}$$

This weighting reflects RGCN's stability while incorporating RGAT's attention-based refinements. A conservative thresholding mechanism is applied to expression-direction predictions (Classes 2 and 3), requiring confidence scores above **0.75** to prevent false positive predictions in these clinically sensitive categories.

3.4.7 Training Configuration and Optimization

Model training employs the Adam optimizer with a learning rate of **0.001** and weight decay of $\mathbf{1 \times 10^{-4}}$. Early stopping is implemented with patience of **50** epochs, monitoring validation macro-F1 scores to ensure optimal generalization. The maximum training duration is set to 300 epochs to prevent excessive computational overhead.

3.4.8 LLM-Based Interpretation Framework

To support interpretability, drug and gene metadata were extracted from the knowledge graph and used as context for explanation generation:

• Drug metadata (e.g., drug classes)
• Gene functional annotations (biological processes, molecular functions, cellular components)

Large language models generate explanations under two key constraints:

Structured response requirement: Output requested in strict JSON structure for automated parsing

Deterministic decoding: Configuration settings to reduce variability in generated explanations

The explanation text is framed conservatively to distinguish computational prediction from established biological knowledge and to emphasize the need for experimental validation.

Figure 3.3: Example of structured JSON explanation generated by the LLM for predicted drug-gene relations

3.5 Methodological Comparison and Evolution

3.5.1 Comparative Framework Analysis

The two methodologies represent a systematic progression from foundational validation to clinically applicable prediction:

Methodology I (Binary Classification):
• Establishes feasibility of graph-based drug-gene interaction prediction
• Provides robust baseline framework with simplified problem formulation
• Serves as proof-of-concept for GNN applicability in biomedical domains
• Utilizes balanced class distribution for stable training

Methodology II (Multi-Class Classification):
• Addresses clinically relevant mechanistic prediction requirements
• Handles realistic class imbalance present in biomedical datasets
• Incorporates advanced loss functions and ensemble strategies
• Provides actionable insights for drug discovery applications

3.5.2 Technical Innovation and Methodological Contributions

Both methodologies contribute distinct innovations to the field: • Integration of knowledge graph structure with deep learning architectures • Application of focal loss to biomedical class imbalance problems • Conservative prediction strategies for high-stakes medical applications • Ensemble approaches combining complementary GNN architectures • Bilingual explanation generation for diverse research communities

Chapter Four

Experimental Setup and Results

Chapter Five

Conclusion and Future Work

References

[1] Gogineni, A. K. (2021). Analysis of drug repurposing knowledge graphs for COVID-19 (Version 1). arXiv. https://arxiv.org/abs/2212.03911

[2] Hao, M., et al. (2023). Deep learning for Alzheimer's disease drug repurposing using knowledge graph and multi-level evidence. iScience, 26(3), Article 106094. https://doi.org/10.1016/j.isci.2023.106094(Note: The published version of the 2021 preprint; full author list available on the journal site.)

[3] Ratajczak, F., Joblin, M., Ringsquandl, M., & Hildebrandt, M. (2022). Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. BMC Bioinformatics, 23, 84. https://doi.org/10.1186/s12859-022-04608-y

[4] Bang, D., Lim, S., Lee, S., & Kim, S. (2023). Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. Nature Communications, 14, 3570. https://doi.org/10.1038/s41467-023-39301-y

[5] Romano, J. D., Truong, V., Kumar, R., Venkatesan, M., Graham, B. E., Hao, Y., Matsumoto, N., Li, X., Wang, Z., Ritchie, M. D., Shen, L., & Moore, J. H. (2024). The Alzheimer's Knowledge Base: A knowledge graph for Alzheimer disease research. Journal of Medical Internet Research, 26, e46777. https://doi.org/10.2196/46777

[6] Loesch, J., Yang, Y., Ekmekci, P., Dumontier, M., & Celebi, R. (2024). Explaining graph neural network predictions for drug repurposing. In Proceedings of SWAT4HCLS 2024 (CEUR Workshop Proceedings, Vol. 3890). http://ceur-ws.org/Vol-3890/paper-5.pdf

[7] Yan, C., Grabowska, M. E., Dickson, A. L., Li, B., Wen, Z., Roden, D. M., Stein, C. M., Embí, P. J., Peterson, J. F., Feng, Q., Malin, B. A., & Wei, W.-Q. (2024). Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. npj Digital Medicine, 7, 46. https://doi.org/10.1038/s41746-024-01038-3

[8] Dobreva, J., Simjanoska Misheva, M., Mishev, K., Trajanov, D., & Mishkovski, I. (2025). A unified framework for Alzheimer's disease knowledge graphs: Architectures, principles, and clinical translation. Brain Sciences, 15(5), 523. https://doi.org/10.3390/brainsci15050523

[9] Selote, R., & Makhijani, R. (2025). A knowledge graph approach to drug repurposing for Alzheimer's, Parkinson's and glioma using drug–disease–gene associations. Computational Biology and Chemistry, 115, 108302. https://doi.org/10.1016/j.compbiolchem.2024.108302

[10] Wang, Z. P., Li, X., Matsumoto, N., Venkatesan, M., Chang, J.-H., Moran, J., Choi, H., Li, B., Meng, Y., Hernandez, M. E., & Moore, J. H. (2025). Drug repurposing for Alzheimer's disease using a graph-of-thoughts based large language model to infer drug-disease relationships in a comprehensive knowledge graph. BioData Mining, 18, 51. https://doi.org/10.1186/s13040-025-00466-5