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

This study develops a comprehensive computational framework for predicting drug-gene interactions in Alzheimer's disease by integrating knowledge graphs, graph neural networks, and large language models. We constructed an Alzheimer-centered subgraph from **AlzKB** 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)** for multi-class prediction of drug-gene interaction types (Binds, Increases Expression, Decreases Expression, or No Relation).

Two main models were trained: RGCN, which captures general structural patterns through relational aggregation, and RGAT, which identifies influential neighbors through attention mechanisms for precise contextual representation. The results showed RGCN excels in broad categories like "No relation" and "Binds", while RGAT demonstrates higher sensitivity to subtle relationships like "Increases Expression" and "Decreases Expression".

The hybrid model combines predictions from both architectures using weighted averaging (70% RGCN + 30% RGAT), achieving the best overall performance with **88.9%** accuracy and weighted F1-score of **0.90%**.Systematic ablation studies validated the contribution of feature engineering, loss function selection, and confidence thresholding to model robustness. The system incorporates Large Language Models through Google Generative AI for generating scientific explanations of predicted interactions, enhancing interpretability for researchers. An interactive Streamlit-based web interface was developed to facilitate real-time predictions, comprehensive analysis, and PDF report generation.

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 with enhanced interpretability and user accessibility.

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

Index

Table of Contents

List of Figures

List of Tables

List of abbreviations

Abbreviation	Definition
AlzKB	Alzheimer's Knowledge Base
AUC	Area Under the Curve
CSS	Custom Cascading Style Sheets
EHR	Electronic Health Record
GELU	Gaussian Error Linear Unit
GNNs	Graph Neural Networks
KGs	Knowledge Graphs
LLMs	Large Language Models
MLP	Multi-Layer Perceptron
PDF	Portable Document Format
RGAT	Relational Graph Attention Network
RGCN	Relational Graph Convolutional Network
ROC	Receiver Operating Characteristic
RTL	Right-to-Left
KEGG	Kyoto Encyclopedia of Genes and Genomes

Chapter One

Introduction

1.1 The scientific background of the problem and its importance

Alzheimer's disease (AD) is one of the most prevalent forms of dementia in the world, and is classified as one of the most intractable global health problems in modern times. The disease is characterized by a gradual, irreversible deterioration in cognitive function caused by the accumulation of a range of abnormal proteins such as beta-amyloid plaques and neurofibrillary tangles in the patient's brain, leading to damage to nerve cells, death, and loss of communication between them [1]. Specialized global medical research centers are working to find a cure for this disease. The disease has been around for decades, but the number of medications approved for treatment is still very limited, and these medications rely on relieving the symptoms of Alzheimer's rather than stopping the disease completely [2]. The process of discovering a new drug is very complex, as it requires a long series of clinical trials. The complex work of drug discovery comes with huge economic costs of up to billions of dollars for every single successful drug [3]. The above reasons have led to the emergence of a new strategy based on drug repurposing, where a pre-existing drug is used for new therapeutic uses, as this saves the necessary time for clinical trials as it is a clinically approved and sound drug, and therefore necessarily saves high costs [4]. A new problem has emerged in general and for Alzheimer's disease in particular stemming from the ability to identify drugs capable of influencing the molecular pathways associated with it, due to the biological complexity of this disease, the presence of a large number of proteins and genes that share its causes, in addition to the lack of direct experimental data on the drug–gene interactions in the context of Alzheimer's [2].

1.2 Research Problem

The most important step in discovering drugs that can be reused in the treatment of Alzheimer's is to identify potential drug–gene interactions between drug compounds on the one hand and the genes that affect the disease on the other, but when these drugs are retested using traditional lab experiments, which are time-consuming and costly, traditional computational models lack the ability to represent the true biological complexity of the disease [2]. Intelligent models based on artificial intelligence–graph neural networks provide accurate predictions, but the problem of lack of interpretability and the black box make researchers distrust its results [5]. As a result, there is an urgent need for an intelligent framework for predicting drug–gene interactions capable of predicting specific types of interactions and leveraging structured biological knowledge via Knowledge Graphs [2], as well as providing justified and convincing explanations for these predictions that increase researchers' confidence in this model with the possibility of adding an interactive interface that makes it user-friendly for researchers [6].

1.3 Research Objectives

The aim of this research is to provide an intelligent and interpretable framework for predicting drug–gene interactions in the context of Alzheimer's disease, using a combination of modern artificial intelligence techniques. The main objectives are as follows:

1. Building a custom Alzheimer’s disease–centric Knowledge Graph that integrates reliable biological relationships between drugs and genes, relying on open data sources and peer-reviewed by experts [2].
2. Design and implement a multi-class classification model capable of distinguishing between different types of interactions [4].
3. Integration of relational graph neural networks (RGCN and RGAT) with a weighted ensemble mechanism to improve prediction accuracy and stability [3].
4. Generate understandable scientific explanations for each prediction using a large language model (LLM), while maintaining academic language that enables the distinction between computational prediction and empirical validation [6].
5. Developing an interactive user interface that allows researchers to explore predictions, filter results, and export detailed reports [4].

1.4 Research Methodology

In this paper, an extensive review of peer-reviewed scientific sources on the use of AI technologies in drug repurposing for the treatment of Alzheimer's will be conducted [2], where a computational framework will be developed to predict and interpret drug–gene interactions in the context of Alzheimer's disease, using a Knowledge Graph built from open biological and drug data sources [5]. The methodology includes a set of sequential steps. First, the Knowledge Graph will be built and equipped focused on Alzheimer's disease [2]. Second, a multi-class classification model will be designed and implemented using relational graph neural networks (RGCN and RGAT) [3]. Third, the model will be combined with a large language model to generate scientific explanations [6]. Fourth, create an interactive user interface (Streamlit) to view the results and export reports [4].

The research will be constrained within the ad hoc framework as it will only work on Alzheimer's disease and not on other diseases, nor will it include the collection of new biological data from primary sources or the conduct of any laboratory or clinical experiments to validate the predictions resulting from the proposed model [2] . No new algorithms or LLMs will be developed from scratch.

1.5 Research Structure

This report consists of five main chapters. The first chapter begins with a background on Alzheimer's disease and the challenges of drug discovery, and identifies the research problem, objectives, and scope of work. The second chapter is followed by the technical background necessary to understand the basic concepts used in the project, such as Knowledge Graphs [2], Graph Neural Networks (GNNs) [3], and Large Language Models (LLMs) [6]. The design and implementation of the system is detailed in the third chapter, starting with the construction of the Knowledge Graph dedicated to Alzheimer’s disease [2], from the development of binary and multi-class classification models [3], to the integration of LLMs to generate scientific explanations and build an interactive user interface [4]. Chapter IV presents empirical results, including performance measures, a comparison between the two models, and examples of predictions and explanations produced. Finally, chapter five concludes the report by summarizing the achievements, discussing limitations, and proposing recommendations for future research.

Chapter Two

Technical background of the research

2.1 Knowledge Graphs

Knowledge graphs are graphs that are used to represent complex data that contain many interlocking links such as the relationship between drugs, genes, and diseases, as these relationships are formed in the form of a network of nodes and edges. KGs provide a unified framework for integrating diverse sources into a single data structure, including molecular databases (such as DrugBank and DisGeNET), scientific articles, and biological ontologies (e.g., Gene Ontology), while maintaining the semantic context of the relationships [1]. KGs have proven their ability to support drug discovery and repurposing, through which it is possible to infer new relationships and associations that allow for drug repurposing without direct clinical and experimental data. The AlzKB database is one of the KGs for Alzheimer's disease developed by Romano et al. (2024), which integrates more than 30 biological sources of gene-binding drugs through qualitative relationships such as direct binding and gene expression modulation [2]. The study showed that the use of focused KG on a single disease improved prediction accuracy by up to 18% compared to general graphs. Dobreva et al. (2025) proposed a unified framework for building KGs in neurological contexts, with a focus on scalability and integration with clinical data, opening the door to translating computational results into actual clinical applications [1] .

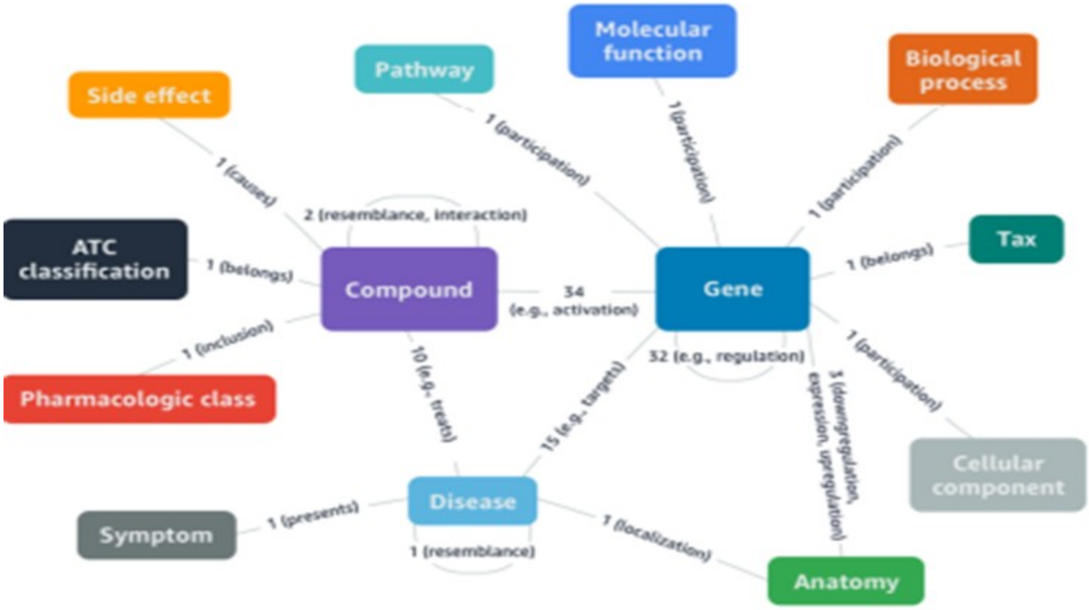


Figure 2.1 - Knowledge Graphs Structure Diagram

Figure 2.1 shows the general structure of the Knowledge Graphs, showing the multiple relationships and the existence of interconnected, node-and-edge relationships between diseases, drugs, and genes.

2.2 Neural Graph Networks (GNNs)

Graph Neural Networks (GNNs) are a class of deep learning models specifically designed to process structural data such as graphs. GNNs generate deep representations (embeddings) of each node by aggregating information from its neighbors, allowing the structural and relational context of the data to be captured [3].

Experiments have shown that relational GNNs provide predictions of drug–gene interactions with high efficiency, especially when the graphs contain multiple types of edges, where RGCN (Relational Graph Convolutional Network) is one of the most commonly used types. A separate weight matrix for each relationship type, allowing for an accurate representation of different biological interactions [3]. As for the use of attention mechanisms to enhance interpretability in relational GNNs, Loesch et al. (2024) employed a dynamic attention approach to weigh the importance of each neighbor based on relationship type and local context, enabling the generation of explanatory attention maps [5]. It was used by Loesch et al. (2024) to improve prediction accuracy as well as to generate explanatory attention maps showing which relationships contributed significantly to the model's decision, enhancing researchers' confidence in the results [5].

When evaluating the Relational Graph Neural Network (RGCN) and the Relational Graph Attention Network (RGAT), there is a fundamental difference in the mechanism by which information is collected from adjacent nodes. The following table summarizes the most prominent differences between these two models in terms of weighting, interpretability, and computational properties [7].

Table 2.1 - Comparison of the RGCN and RGAT models.

Standard	RGCN (Relational Graph Convolutional Network)	RGAT (Relational Graph Attention Network)
Weighting Mechanism	Fixed weights assigned to each relationship type	Dynamic weights calculated by the attention mechanism according to the local context
Interpretability	Low (does not show the importance of individual neighbors)	High (produces analyzable attention weights for each neighbor)
Mathematical complexity	Less (simple linear array operations)	Higher (calculating attention scores requires additional operations)
Fit for our project	Suitable for basic forecasting	Preferred because it supports the project's goal of providing scientific explanations

Based on Table 2.1, the RGAT model was selected as a key component of the proposed framework, due to its ability to generate dynamic attention weights that reflect the importance of each neighbor in the forecasting process. This characteristic enables scientific explanations based on the most influential relationships and the RGCN has been retained as a baseline model for assessing the true benefit that the attention mechanism adds [7].

2.3 Large Language Models (LLMs) in Scientific Interpretation

In recent years, Large Language Models (LLMs) have emerged as powerful tools for understanding and generating natural language in scientific and medical contexts. These models can generate research hypotheses, summarize studies, and interpret computational results in a language that is understandable to researchers [6]. Yan et al. (2024) used the GPT-4 model to generate clinical explanations for predictions of drug repurposing in Alzheimer's disease, and then validated these interpretations using real electronic health record (EHR) data, demonstrating that LLMs can be a bridge between computing and clinical medicine [6]. However, the main challenge lies in the phenomenon of "hallucination", where the model may generate incorrect or data-unsupported information. Wang et al. (2025) developed a framework called ESCARGOT, which integrates the Knowledge Graph with the LLM via a "Graph-of-Thoughts" mechanism, providing the model with only the content extracted from the graph (e.g., drug classes, molecular functions of the gene), and preventing the use of its internal knowledge, ensuring that interpretations are based exclusively on the input data [4].

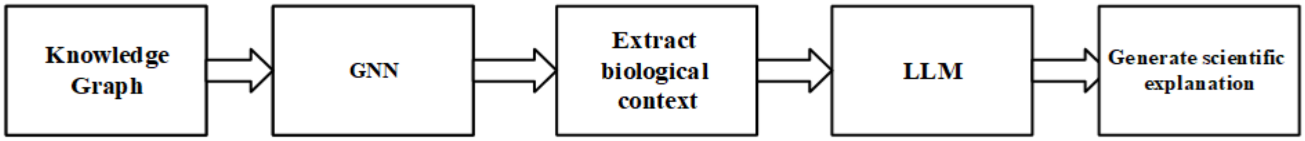


Figure 2.2 - Diagram showing the data flow for scientific analysis.

Figure 2.2 shows the overall structure flow of the proposed system, where the system starts by extracting data from the Knowledge Graph, then passes it to the GNN model for predicting the interaction, and uses the model output as input to a large language model (LLM) to generate an understandable scientific explanation, which is presented to the user as a final report.

2.4 Key Concepts

Link prediction is a primary task in Knowledge Graphs analysis, and is to predict the probability of a relationship between a pair of nodes that are not currently associated in the graph. This prediction translates to determining whether a particular drug may interact with a target gene [3]. This task has been treated as a binary classification problem, where each drug–gene pair is classified as either an interaction or no interaction, resulting in negligence of differences in biological mechanisms, as the effect of a drug on a gene may be stimulating (increased expression) or inhibitory (decreased expression), each with different therapeutic effects. The trend towards multi-class classification, which distinguishes between specific types of interactions, was therefore an inevitable necessity. Studies by Selote & Makhijani (2025) and Wang et al. (2025) have confirmed that this approach increases the clinical value of predictions, as it enables researchers to design laboratory experiments that target a specific mechanism rather than testing the reaction in general [3,4].

2.5 Implementation tools

- This project was implemented using a set of modern software tools that support the development of AI models and graph analysis.
- Python 3.10: The basic programming language, thanks to its rich libraries in the field of data science and machine learning.
- PyTorch Geometric: A specialized library for implementing graph neural networks, which has been used to build and train RGCN and RGAT models across the RGCNConv and RGATConv layers [3].
- Gemini API: Used to call a large language model (LLM) to generate scientific explanations, with deterministic decoding enabled to ensure the stability of the results [4].
- Memgraph: A high-performance graph database, used to store the Knowledge Graph for Alzheimer's disease, with retrieval queries performed via Cypher [1].
- Streamlit: A framework for rapidly developing interactive user interfaces, and it has enabled the construction of a web interface that allows researchers to explore predictions and generate reports.

Chapter Three

system design

3.1 Database

3.1.1 Data Source

AlzKB (Alzheimer's Knowledge Base) version 2.0.0 is an open-source graphical base developed specifically for AI research in Alzheimer's disease by a joint research team between the University of Pennsylvania (UPenn) and Cedars-Sinai Medical Center [1]. The AlzKB database is described as follows:

- Official source: https://github.com/EpistasisLab/AlzKB.
- License: MIT License (open source for academic research purposes).
- Format: alzkb-v2-0-0_memgraph.cypher, a CYPHERL dump compatible with graphical rule systems such as Memgraph.
- Content: AlzKB integrates data from multiple biological and pharmacological sources, including DisGeNET, DrugBank, Gene Ontology, ClinicalTrials.gov, and OWL Rules for Neurological Diseases.
- Structure: Contains a comprehensive network of biological entities covering Drugs, Genes, Diseases, Biological Pathways, Molecular Functions, and Side Effects.

This resource was chosen because it provides a standardized structure specifically designed for Alzheimer's disease, which reduces integration errors and provides a reliable basis for predictions of drug-gene interactions. The file was imported directly into Memgraph 2.10 via the Import & Export → Import Data interface, without the need to rebuild the base from the primary sources.

3.1.2 Explore the database

After the alzkb-v2-0-0_memgraph. cypherl file was imported into the Memgraph environment, a comprehensive exploratory analysis was performed that showed that the AlzKB (Alzheimer's Knowledge Base) expresses a rich and interconnected biological representation, as it was specifically designed for AI research in Alzheimer's disease by a joint research team between the University of Pennsylvania and the Cedars-Sinai Center [1]. AlzKB contains 234,037 nodes and 1,668,487 edges, but the distribution of these elements is not random; rather, it reflects a clear scientific focus. Genes form the backbone of the graph, accounting for 193,279 nodes (more than 82% of the total), with drugs coming in second with 16,581 nodes, providing a broad base for drug repurposing. The rest of the nodes are distributed among supporting entities that enhance the functional context: more than 12,000 biological processes, and 4,516 molecular pathways (from sources such as Reactome and KEGG), as well as thousands of nodes representing molecular functions and cellular components. One of the most notable indicators of personalization is that there are only 34 disease nodes, all of which are dedicated exclusively to Alzheimer's disease and its clinical derivatives — from genetic forms such as Familial Alzheimer's Disease to age-specific forms such as Late Onset, to forms associated with complex disease pathways such as the Lewy Body Variant. This leads to AlzKB being an engineering research tool designed to study this complex neurological disease.

The semantic structure shows the dominance of general biological relationships (e.g., GENEPARTICIPATESINBIOLOGICALPROCESS), which make up the bulk of the edges. However, the focus of this research will be limited to a smaller, more significant group with 65,490 direct relationships between drugs and genes, classified into three distinct mechanistic categories: CHEMICALBINDSGENE, CHEMICALINCREASESEXPRESSION, and decreased expression (CHEMICALDECREASESEXPRESSION). These three relationships represent medicinal mechanisms of action that can be experimentally tested. According to the official AlzKB documentation, the relationship DRUGCAUSESEFFECT refers to side effects and not to the mechanisms of therapeutic action, and is therefore excluded.

When the research was extended to include all nodes within a maximum distance of two nodes of any Alzheimer's disease, more than 20,000 additional nodes (drugs, genes, pathways) were discovered, reflecting the enormous biological complexity surrounding the disease, and confirming that any effective predictive model must take into account this expanded network context.

3.2 Engineering Features

3.2.1 Alzheimer-Centered Subgraph Construction

The rich graph which contains text nodes and semantic edges has been converted into a structured numerical representation that the machine can learn, in order to enable deep learning models to understand the biological structure of Alzheimer's disease as documented in AlzKB. This process began by building a concentrated subgraph focused on entities directly or indirectly related to Alzheimer's disease.

The subgraph extraction followed a systematic three-step approach designed to capture the most relevant biological entities:

Step 1: Disease Node Identification We first identified all disease nodes containing "alzheimer" in their commonName property, which yielded **34 disease nodes** representing various forms and subtypes of Alzheimer's disease.

Step 2: Gene Association Extraction Next, we extracted genes directly associated with these Alzheimer's disease nodes through established disease-gene relationships. This identified **103 genes** with documented connections to Alzheimer's disease, including well-known genes like APOE, PSEN1, PSEN2, and ABCA7.

Step 3: Drug-Gene Interaction Mapping Finally, we extracted drugs that interact with the Alzheimer-associated genes through three key interaction types: chemical binding, increased expression, and decreased expression. This process identified **861 drugs** with documented interactions with Alzheimer-related genes.

Selection Rationale and Connectivity Analysis The selection criteria prioritized entities with strong connectivity to Alzheimer's disease. While we found 20,552 connected nodes within 1-2 hops from Alzheimer's disease nodes across 190,757 paths, we focused on the most directly connected entities to reduce noise and create a more focused dataset. This approach ensured that our subgraph contained the most biologically relevant relationships while excluding weakly connected entities that could introduce noise into the machine learning models.

Final Subgraph Statistics The resulting Alzheimer-centered subgraph contains:

- **998 total nodes** (34 diseases + 103 genes + 861 drugs)
- **2,111 edges** representing various biological relationships

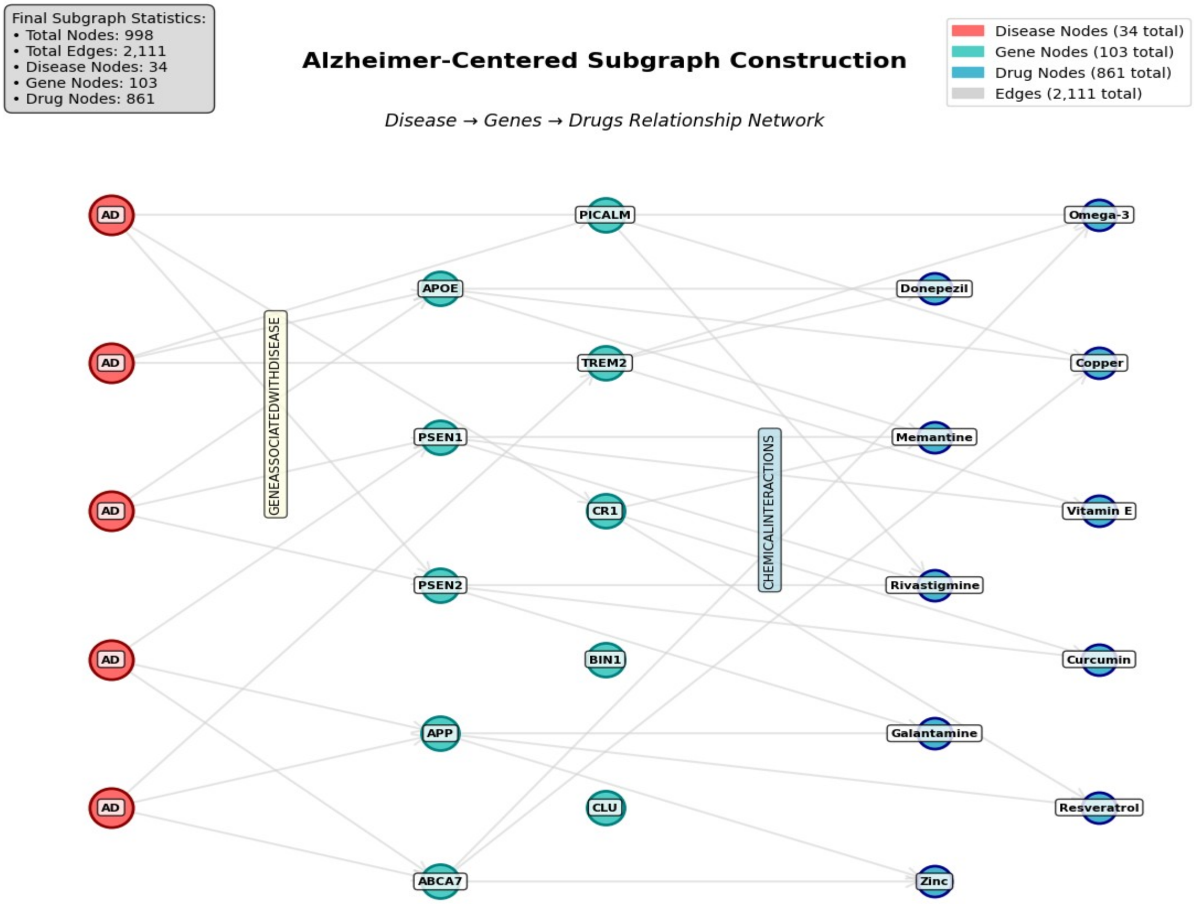


Figure 3.1 Alzheimer-Centered Subgraph Construction showing the Disease → Genes → Drugs relationship network

Figure 3.1: Alzheimer-Centered Subgraph Construction showing the Disease → Genes → Drugs relationship network with 34 disease nodes (red hexagons), 103 gene nodes (teal hexagons), and 861 drug nodes (blue hexagons), connected across 2,111 edges.

This filtering aims to reduce noise generated by unrelated entities and focus the model's capability on the real pathological context, creating a concentrated dataset that captures the essential biological relationships surrounding Alzheimer's disease.

3.2.2 Feature Extraction and Engineering

A set of features that capture different aspects of its location and function were extracted for each node. Initially, simple structural indicators such as degree (number of direct links) and PageRank (relative importance in the network) were calculated, and because these indicators were insufficient to capture the semantic complexity of biological relationships, deep representations were generated using the DeepWalk algorithm[10], which simulates a random walk through the network and then trains a language model (Word2Vec) [11] on these paths to learn a numerical representation of each node. The result was a 128-dimensional vector for each node, reflecting their direct links and broader functional context_for example, two genes involved in the same biological pathway would have similar representations even if they were not directly related. These 128-dimensional DeepWalk embeddings were combined with a 1-dimensional node type label (distinguishing between disease, drug, or gene entities) to form a 129-dimensional primary feature matrix. Since the project's goal is to predict drug-gene interactions, i.e., classify edges rather than nodes, each edge in the subgraph was converted into a graded training sample. three positive categories were defined based on the type of biological relationship:

1. CHEMICALBINDSGENE.

2. CHEMICALINCREASESEXPRESSION.

3. CHEMICALDECREASESEXPRESSION.

To ensure directional consistency—which is crucial in biological prediction—all relationships were standardized so that the drug always refers to the gene, regardless of the orientation of the original edges in the graph. Then, to enable the model to distinguish between real interactions and hypothetical ones, approximately equivalent negative samples were generated by taking random pairs of drugs and genes that were not associated with the graph. Finally, all of this structured structure has been converted to the format required by the PyTorch Geometric framework. Sequential numerical identifiers have been assigned to each node, the feature matrix has been converted to a float tensor, and the edge index has been converted to a long tensor. The feature matrix has also been expanded to include the degree and PageRank structural indicators, resulting in a 131-dimensional expanded feature matrix (129 primary dimensions + 1 degree + 1 PageRank). The final object has been saved as an alz_raw_tensors.pt file, to be ready for the training phase. This comprehensive methodology ensures that the model learns from the raw data as well as from the full semantic structure of Alzheimer's disease as understood in the scientific literature.

3.3 Model Training and Inference

The hybrid model is designed to be based on neural graph networks (GNNs): Advanced_RGCN and Advanced_RGAT. Both networks in the proposed model follow the same general two-stage architecture: encode and decode. In the encoding phase, the array of initial features of the nodes which is 131 dimensions long, is transformed as a result of the integration of DeepWalk representations (128 dimensions) with the degree pointer (one dimension), the PageRank (one dimension), and the node type information (one dimension) into deep representations of 256 dimensions. This dimension (256 instead of 128) was chosen to enable the model to capture the complex nonlinear patterns that characterize biological data. In the decoding phase, the source and target node representations of each drug-gene pair are combined to form a complex feature vector, which is passed through a multi-layer MLP network to generate probabilities of belonging to the four classes.

3.3.1 Model Advanced_RGCN*:* Capturing Static Structural Patterns

The Advanced_RGCN model is built on the basis of the RGCNConv layer of PyTorch Geometric [7], a layer specifically designed for multi-relationship graphs. The use of RGCN ensures that each relationship is handled separately in the event of future system expansion. The model consists of two convolutional layers: one converts the 131-dimension input into a 256-dimensional hidden representation, and the second maintains this dimension. To enhance numerical stability and prevent overfitting, the model includes Layer Normalization after each layer, plus 30% dropout after the first layer. A residual connection is also included between the two layers to facilitate the flow of gradients in deep networks. Paired node representations are integrated via a three-layer serial function (MLP) in the decoding phase, which uses GELU [12] as an activation function to improve the ability to represent nonlinear patterns.

```
class Advanced_RGCN(nn.Module):
    def __init__(self, in_dim, hidden_dim, num_classes, num_rel):
        super().__init__()
        self.conv1 = RGCNConv(in_dim, hidden_dim, num_rel)
        self.conv2 = RGCNConv(hidden_dim, hidden_dim, num_rel)
        self.ln1 = nn.LayerNorm(hidden_dim)
        self.ln2 = nn.LayerNorm(hidden_dim)
        self.dropout = nn.Dropout(0.3)
        self.edge_mlp = nn.Sequential(
            nn.Linear(2 * hidden_dim, hidden_dim),
            nn.GELU(),
            nn.Dropout(0.3),
            nn.Linear(hidden_dim, hidden_dim // 2),
            nn.GELU(),
            nn.Linear(hidden_dim // 2, num_classes)
        )
```

3.3.2 Model Advanced_RGAT:

Modeling Dynamic Contexts via Attention

The Advanced_RGAT model differs in its aggregation mechanism, as it relies on the RGATConv layer [9], which integrates the multi-head attention mechanism, where the model learns the dynamic weight of each neighbor based on its local context. The first layer is configured to receive 131 dimensions and produce 64 dimensions for each head (since hidden_dim // heads = 256 // 4 = 64), and then the four heads are combined to form a final representation of 256 dimensions. The second layer follows the same logic. Choosing the number of heads to be 4 balances expressive power with computational costs, as each head learns a different part of the local context (such as the type of relationship or gene function), and when combined, produces a rich representation that dynamically reflects the importance of each neighbor. This configuration follows established practices in graph attention architectures to stabilize the learning process and capture diverse relational features[8]. Like RGCN, RGAT has Layer Normalization, dropout, and residual connections, but it excels at capturing precise contextual interactions that models based on fixed averages might miss.

```
class Advanced_RGAT(nn.Module):
    def __init__(self, in_dim, hidden_dim, num_classes, num_rel, heads=4):
        super().__init__()
        self.conv1 = RGATConv(in_dim, hidden_dim // heads, num_rel, heads=heads)
        self.conv2 = RGATConv(hidden_dim, hidden_dim // heads, num_rel, heads=heads)
        self.ln1 = nn.LayerNorm(hidden_dim)
        self.ln2 = nn.LayerNorm(hidden_dim)
        self.dropout = nn.Dropout(0.3)
        self.edge_mlp = nn.Sequential(
            nn.Linear(2 * hidden_dim, hidden_dim),
            nn.GELU(),
            nn.Dropout(0.3),
            nn.Linear(hidden_dim, num_classes)
)
```

3.3.3 Addressing class imbalances

Due to the significant imbalance in the distribution of categories — negative samples (class 0) make up about 50% of the data, while the expression reduction category (category 3) accounts for less than 5% ,Focal Loss has been adopted as a key loss function. This function, developed by Lin et al. (2017), reduces the effect of easy-to-classify samples (e.g., class 0) and increases the model's focus on difficult samples (e.g., category 3). Category weights were automatically calculated using scikit-learn's compute_class_weight, resulting in weights of [0.50, 0.67, 2.81, 5.88], which exactly reflected the rarity of each category.

3.3.4 Training and assessment settings

Both models were trained with an Adam optimizer with a learning rate (lr) of 0.001 and a weight decay of 1e-4, benchmark values that have proven effective in GNNs tasks. The number of epochs was set to 300, with Early Stopping enabled after 50 epochs of F1 macro impairment on the validation set. The learning curves showed that the two models reached a good balance between overfitting and generalization, where the loss of training and validation converged without significant separation,As shown in (*Figure 4.4 , Figure 4.6*).

3.3.5 Inference and Synthesis

A late fusion strategy is implemented through a weighted average mechanism of the model outputs. After the graph is processed in parallel by Advanced_RGCN and Advanced_RGAT, each model generates independent logits for the edge classification task. These logits are then combined using a weighted averaging approach with fixed weights (70% for RGCN and 30% for RGAT), followed by softmax activation to produce the final probability distribution. This fusion strategy allows the hybrid model to leverage RGCN's strength in capturing static structural patterns and RGAT's ability to model dynamic contexts through the attention mechanism. The predetermined weighting scheme gives greater emphasis to the RGCN model based on its demonstrated superior performance, while still maintaining the complementary contributions from the RGAT model. This approach maintains the interpretability of individual model outputs while achieving improved overall accuracy through strategic combination of their respective strengths.

3.4 Integration with LLM for Scientific Interpretation

3.4.1 Linking Prediction and Interpretation*:* Purpose and Design

The success of the model in drug discovery research is measured by its ability to generate testable hypotheses, and in order to achieve this goal, an integrated interpretation system has been designed that links the prediction phase (based on GNNs) and the interpretation phase (based on LLMs). The basic idea is that every numerical prediction, no matter how accurate, remains a black box unless translated into an understandable biological language. The system extracts the complete biological context of each drug-gene pair from AlzKB, and then uses this context to generate a textual interpretation. Large language models (LLMs) were selected for this role because they have the unique ability to integrate fragmented information (e.g., drug class, gene function, biological process) into a coherent and logical narrative. Although current implementations use the Gemini API for its proven efficiency in scientific tasks, the architecture is designed to be model-independent, allowing it to be replaced by any other model (such as GPT-4 or Llama 3) without the need to re-engineer the system.

3.4.2 Prompt Engineering and Assurance of Repeatability

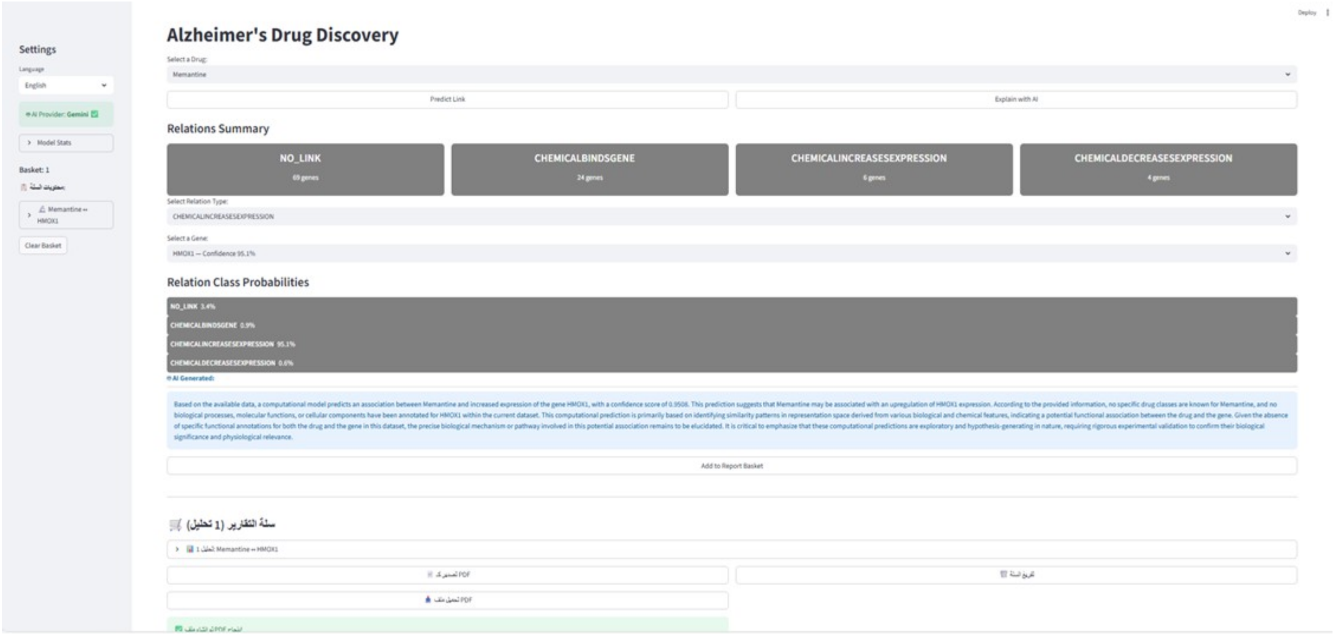
At the heart of this system is an engineered prompt, designed to transform an LLM from a text generator into a trusted research assistant. The prompt includes strict instructions on several levels:

- **First**, the methodological framework: The model is required to always begin by asserting that the result is a computational prediction and not a proven biological fact.
- **Second**, cautious language: it prohibits the use of categorical phrases such as causes or inhibits, and requires the use of phrases such as the model suggests or the computational signal suggest.
- **Third**, strict reliance on the context provided: Include in the prompt explicit warnings not to use any outside knowledge, emphasizing that the interpretation should be limited to data extracted from AlzKB.
- **Fourth**, experimental conclusion: The model is always required to conclude by indicating that these results are preliminary and need to be empirically validated.

These written instructions provide systematic safeguards that make results reproducible and verifiable. Since each interpretation is based on a specific and finite context (extracted from the graph), restarting the system for the same pair will produce almost the same interpretation, especially when adjusting the model settings (temperature=0). The model thus transforms from a black box to an interpretable system, where the researcher can track every step: from numerical representation, to biological context, to textual interpretation. This transparency is what makes the system a real research tool, not just a prediction engine.

3.5 Interactive User Interface

An interactive user interface is designed using the ***Streamlit*** framework. This interface aims to enable the user to explore predictions, understand them via text interpretations, and export the results in the form of reports ready to be shared or published.


Settings

Language: English

AI Provider: Gemini ✅

Basket: 1

معلومات السلة:

Memantine ↔ HMOX1

Alzheimer's Drug Discovery

Select a Drug: Memantine

Relations Summary

NO_LINK 69 genes

CHEMICALBINDSGENE 24 genes

CHEMICALINCREASESEXPRESSION 6 genes

CHEMICALDECREASESEXPRESSION 4 genes

Select Relation Type: CHEMICALINCREASESEXPRESSION

Select a Gene: HMOX1 — Confidence 95.1%

Relation Class Probabilities

NO_LINK 3.4%

CHEMICALBINDSGENE 0.9%

CHEMICALINCREASESEXPRESSION 95.1%

CHEMICALDECREASESEXPRESSION 0.6%

AI Generated:

Based on the available data, a computational model predicts an association between Memantine and increased expression of the gene HMOX1, with a confidence score of 0.9508. This prediction suggests that Memantine may be associated with an upregulation of HMOX1 expression. According to the provided information, no specific drug classes are known for Memantine, and no biological processes, molecular functions, or cellular components have been annotated for HMOX1 within the current dataset. This computational prediction is primarily based on identifying similarity patterns in representation space derived from various biological and chemical features, indicating a potential functional association between the drug and the gene. Given the absence of specific functional annotations for both the drug and the gene in this dataset, the precise biological mechanism or pathway involved in this potential association remains to be elucidated. It is critical to emphasize that these computational predictions are exploratory and hypothesis-generating in nature, requiring rigorous experimental validation to confirm their biological significance and physiological relevance.

سلة التقارير (1 تحليل)

تحليل 1: Memantine ↔ HMOX1

اكتمل إنشاء ملف PDF!

Figure 3.2 Interactive User Interface

The interface shown in Figure 3.2 has several key features:

- **First**, the interactive prediction system: allows the user to select a drug from a list of 861 Alzheimer's-related drugs, and then view all the genes associated with it with the classification of each relationship according to the four categories. The results are displayed in the form of a summary panel showing the number of genes for each relationship type, helping the researcher identify promising biological patterns. At the top of the interface, the title Alzheimer's Drug Discovery appears, with an option to select the language (English/Arabic) and the type of AI provider (Gemini). Below it, there is a relations summary panel showing four boxes showing the number of genes for each category:

  NO_LINK (69 GENES),
  CHEMICALBINDSGENE (24 GENES),
  CHEMICALINCREASESEXPRESSION (5 GENES),
  CHEMICALDECREASESEXPRESSION (4 GENES).

- **Second**, the dual interpretation system: When selecting a specific drug-gene pair, the user can request an explanation of the expected relationship. Here, the system first tries to use the Gemini API to generate a rich and detailed explanation. If the connection fails (due to key or network issues), the system activates a local mode that generates a modular interpretation based on the pre-prepared academic template. This hybrid design ensures that the interface remains usable even during a cloud service outage. At the bottom of the interface, the Relation Class Probabilities section appears which shows the probabilities of the four categories, highlighting the most likely category (in this example: CHEMICALINCREASESEXPRESSION with 95.1% probability). Underneath it, there is an AI Generated box containing the full textual interpretation, which always begins with Based on the available data..., emphasizing the exploratory nature of the result.
- **Third**, bilingual support: The interface is designed to support both Arabic and English, taking into account the technical challenges specific to the Arabic language, such as text orientation (RTL). To achieve this, custom CSS has been used to dynamically adjust the orientation of text elements based on the user's chosen language, ensuring a smooth and clear user experience.
- **Finally**, the export and reporting system: The system has a report basket that allows the user to collect several interesting analyses, and then export them as a unified PDF file. This file is created using the ReportLab library, and contains a summary page that shows the number of predictions by type, followed by separate pages for each analysis, showing the basic details and scientific explanation. The PDF file has been designed to be in English only, it is the standard option for scientific reports, ensuring that all Arabic interpretations are accurately translated into English using a dedicated scientific glossary. At the bottom of the interface is the Report Basket section that displays the current analysis (example: Analysis 1: Memantine = HMOX1), with buttons to export it as a PDF or add it to the cart.

This interface provides an interactive gateway that connects AI to biological research, enabling the researcher to turn computational predictions into testable hypotheses in the laboratory.

Chapter Four

Results and evaluation

4.1 Introduction

The phase began with the construction of a binary classification model aimed at distinguishing between the presence or absence of a drug-gene interaction. This model demonstrated acceptable performance under traditional statistical metrics, such as accuracy and area under the ROC curve (AUC), confirming its ability to capture general patterns in the data.

However, its conceptual limitations soon emerged: while the model could answer the question "Is there an interaction?", it is completely incapable of providing any information about the nature of this interaction — the most important information from the perspective of molecular biology and drug repurposing. Knowing only that there is an interaction is not enough to guide laboratory experiments, nor does it help the researcher understand the potential drug mechanism.

Based on this shortcoming, a more advanced model was developed: the multi-class classification model. This model classifies interactions into four clear categories that reflect realistic biological mechanisms:

[1] No interaction.

[2] Direct binding (CHEMICALBINDSGENE).

[3] Increased gene expression (CHEMICALINCREASESEXPRESSION).

[4] Decreased gene expression (CHEMICALDECREASESEXPRESSION).

The model has proven superior not only in terms of statistical performance, but more importantly in terms of practical scientific value, producing rich predictions that can be directly converted into testable hypotheses in the laboratory.

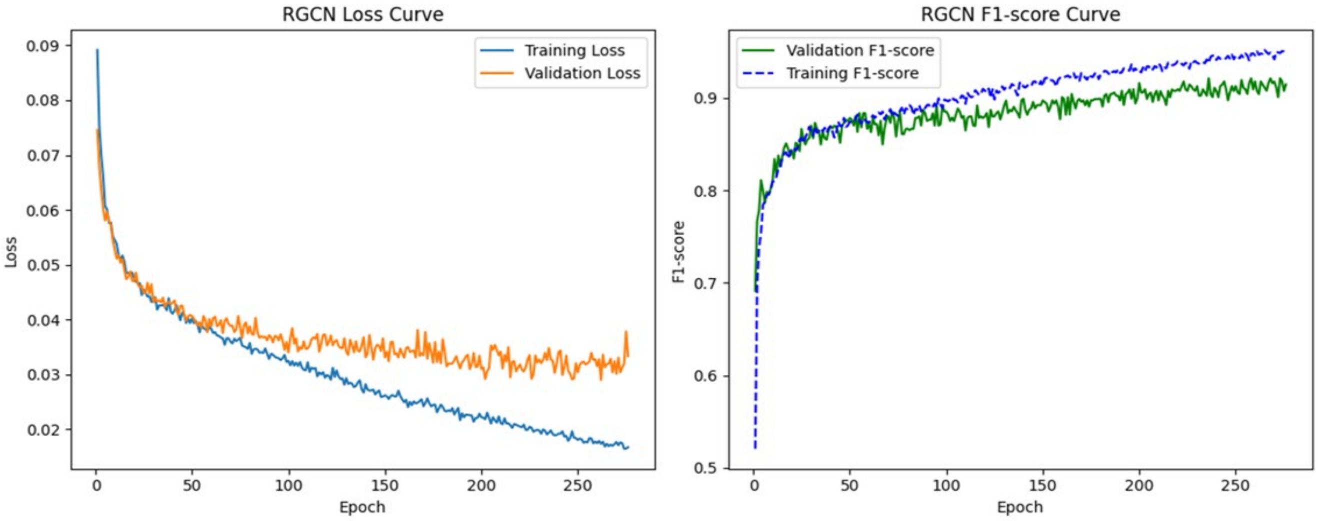
4.2 Binary classification

In the first phase, a binary classification approach was adopted to assess the ability of models to predict whether or not a drug-gene interaction was present. This phase involved three main steps: building an Advanced_RGCN-based model and an Advanced_RGAT-based model, and then integrating them into an ensemble model via weighted average (0.7 * RGCN + 0.3 * RGAT).

Although this approach does not meet the full biological requirements it does not provide information about the nature of the interaction it is a critical methodological step. It enables the evaluation of the feasibility of using graph neural networks (GNNs) to capture structural patterns of data, and provides a necessary basis for the development of a more advanced model capable of distinguishing between different drug mechanisms.

4.2.1 RGCN Model for Binary Classification

The first model relied on a two-layer Advanced_RGCN network, which used 131-dimensional nodal representations (including DeepWalk representations and structural indicators such as degree and PageRank).



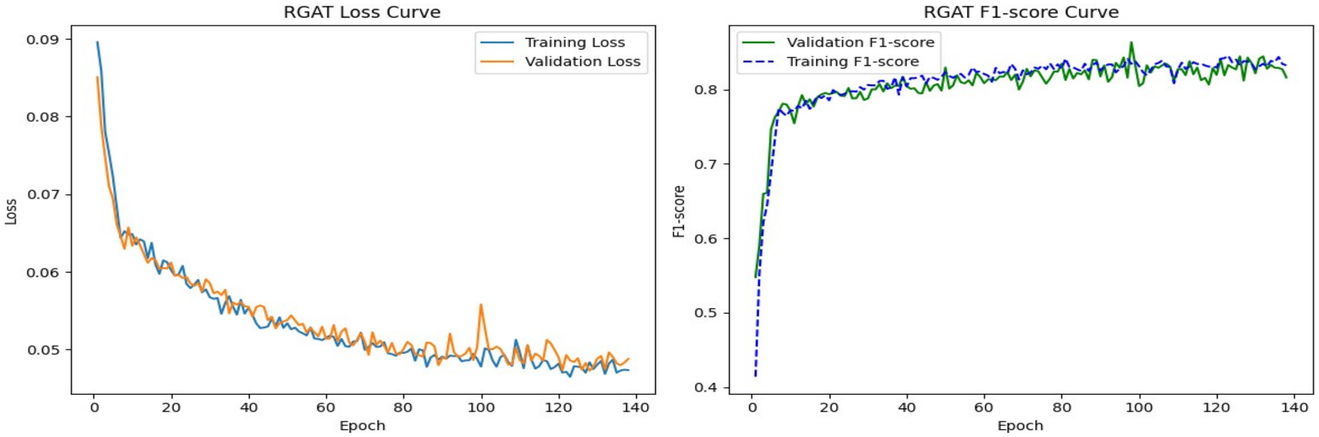
Fi

Figure 4.1: Loss curves and F1-score of the RGCN binary classification model

The model’s loss curve (Figure 4.1) showed a gradual decrease in both training and validation loss, reflecting a continuous improvement in model performance as the training epochs progressed. In contrast, the F1-score curve (Figure 4.1-b) shows a clear increase in model performance on the validation set until stabilization was reached, with marked convergence between training and validation F1-scores. This convergence indicates the absence of overfitting, which was confirmed by the Early Stopping mechanism, terminating the training process at epoch 276 after the validation loss ceased improving for 40 consecutive epochs.

4.2.2 RGAT Model for Binary Classification

The second model relied on an Advanced_RGAT network that relies on the multi-head attention mechanism to capture the local context of each drug-gene relationship. The advantage of this mechanism is its ability to distribute dynamic weights to neighbors based on their local context, allowing the model to focus attention on the most biologically relevant relationships.



h

Figure 4.2: Loss curves and F1-score for the binary classification RGAT model

The performance curves (Figure 4.2) showed similar behavior in terms of low loss and high F1-score, with a key note: the model reached a state of stability in a much lower number of training epochs compared to RGCN, where training ended at epoch 138 (vs. 276 for RGCN), due to the activation of the early stopping mechanism after the validation loss stopped improving for 50 consecutive epochs. This reflects a greater speed of learning and more efficient exploitation of contextual information in the graph, which corresponds to the nature of attention that improves the flow of gradients and accelerates convergence.

In terms of performance, the model achieved an F1-score on the validation set of ~0.85, which is slightly higher than that achieved by the RGCN (~0.82) at the same stage, confirming the theoretical hypothesis that the attention mechanism is able to distinguish the most important neighbors in the prediction process, thereby improving the accuracy of classification—especially in the presence of complex or nonlinear relationships.

4.2.3 Ensemble RGCN + RGAT for Binary Classification

A hybrid model was built by integrating the outputs of Advanced_RGCN and Advanced_RGAT to leverage the strengths of both models. In this experiment, the graph was passed through both models in parallel, and then the weighted average (0.7 × RGCN + 0.3 × RGAT) of the decoding phase outputs for each drug–gene pair was calculated; this average was used to make the final decision. The classification report (Table 4.1) shows a good balance between the two categories.

Table 4.1: Classification Report for Model Ensemble RGCN+RGAT

support	F1-score	recall	Precision	
590	0.91	0.87	0.96	0
582	0.92	0.96	0.88	1
1172	0.91			Accuracy
1172	0.91	0.91	0.92	Macro avg
1172	0.91	0.91	0.92	Weighted avg

The test results showed that the hybrid model performed better than the two individual models, with an accuracy of about 91.5%, an overall average F1 value of about 0.91, and an area under the ROC curve (AUC) of about 0.97, reflecting a high ability to distinguish between interacting and non-interacting pairs. The classification report (Table 4.1) shows that the model maintained a good balance between the two categories, with high values for both Precision and Recall. The confusion matrix (Figure 4.3) also shows that the number of false positive and false negative errors remained limited compared to the number of correct predictions.

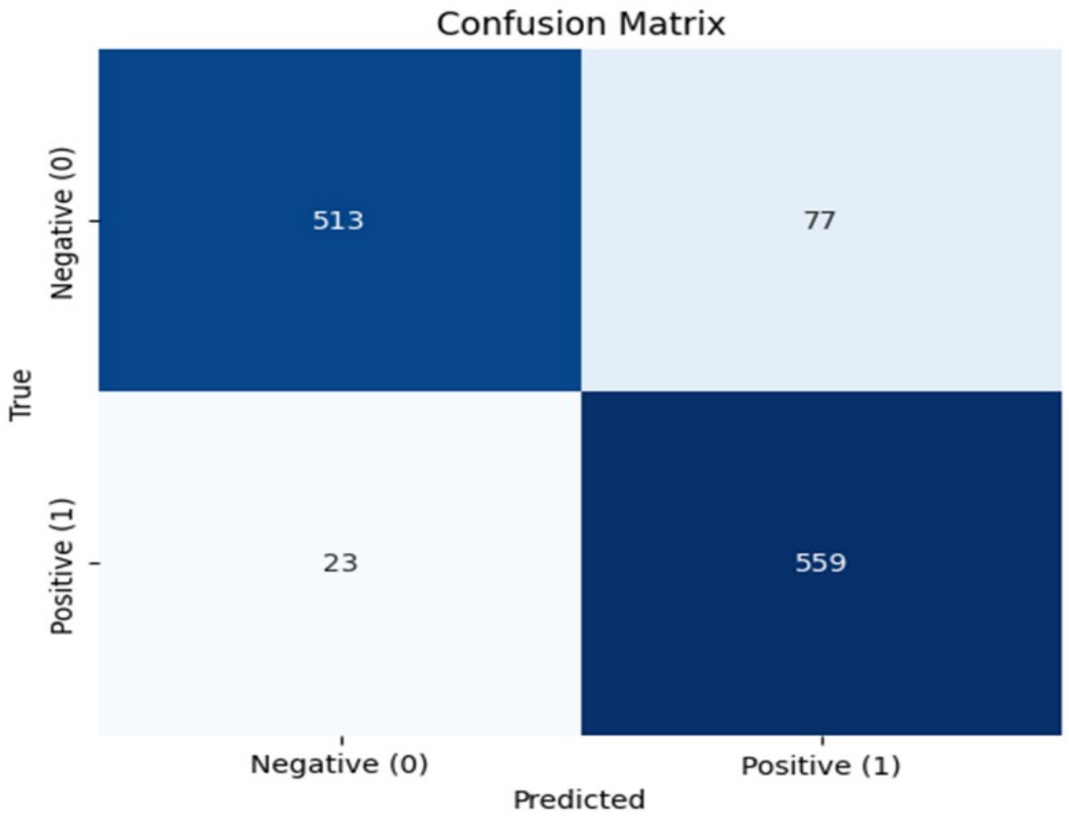


Figure 4.3: Confusion Matrix for Ensemble RGCN + RGAT Binary Classification

The confusion matrix (Figure 4.3) shows the exact distribution of errors:

- Correct predictions:
- Negative → Negative: 513
- Positive → Positive: 559
- Errors:
- False →Positive: 77
- False →Negative: 23

This means that the model was able to classify 96% of the true positive samples (Recall = 0.96) and 96% of the negative predictions were correct (Precision = 0.96), confirming its ability to generalize without being too biased to any category.

4.3 Results of Multi-class Classification of Drug-Gene Relationships

The transition from simple binary to multi-class classification was made, with the aim of distinguishing four different types of drug-gene relationships: no relationship, direct binding, increased gene expression, and decreased gene expression. This transition more accurately reflects biological reality, transforming the model from a mere tool for detecting the "existence" of a relationship into an intelligent system capable of characterizing its mechanistic nature — critical information to guide laboratory experiments. Three modelling approaches were adopted:

- An Advanced_RGCN-based model to capture static structural patterns in the graph.

- An Advanced_RGAT-based model for exploiting the multi-head attention mechanism in modeling dynamic contexts.

- A hybrid model that combines the prediction outputs from both models (RGCN + RGAT) through a weighted averaging mechanism, where each model utilizes its own MLP-based decoder for final edge classification.

These models were evaluated using a comprehensive set of metrics, including training curves (loss and F1-score), confusion matrices, and classification reports, to ensure a fair and comprehensive evaluation of their performance.

4.3.1 RGCN Model

The Advanced_RGCN model exhibits mature and stable training behavior, which is clearly evident in its training curves (Figure 4.4). In the loss curve, the training loss gradually and smoothly decreases until it reaches the final value of 0.0544, while the validation loss decreases to 0.1809 before stabilizing. This relative convergence between the training and validation curves, with the validation loss remaining within a low range, indicates that the model learns a good general representation of the data without being limited to patterns specific to the training set — that is, it is not overfitting.

In the F1-score curve (Figure 4.4), a gradual increase in F1 can be observed on the training set until it reaches 0.8928, while the F1 on the validation set reaches 0.8763, with a clear convergence between the two curves in the later stages of training. This behavior reflects consistency between the model's performance on the data it learned from and unseen data.

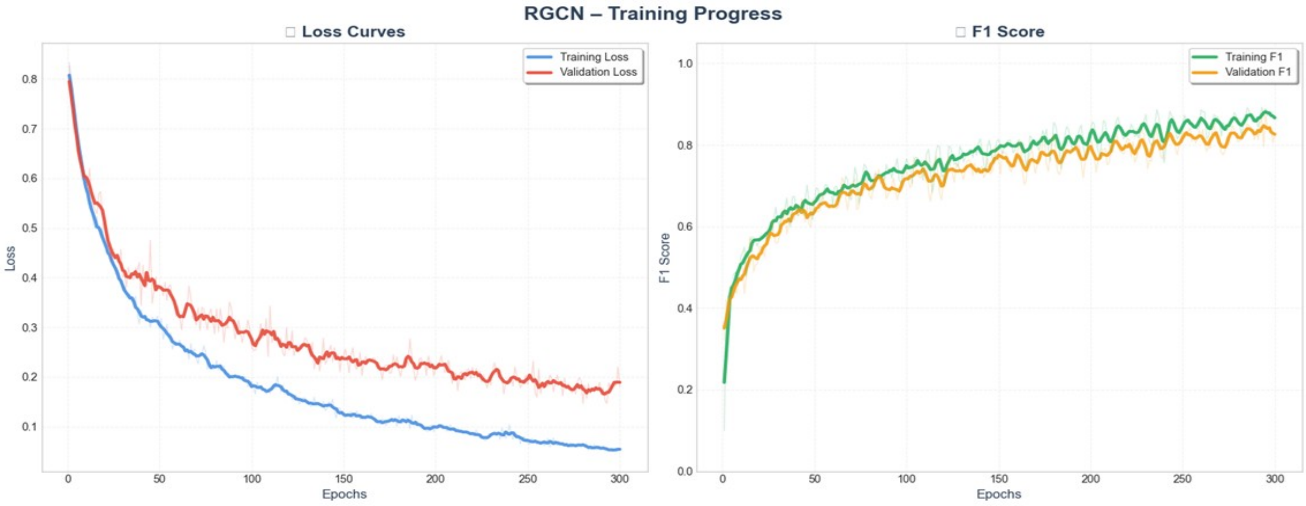


Figure 4.4: Loss curves and F1-score of the RGCN multi-class classification model

The model’s confusion matrix supports this quantitative reading. The matrix shows that the model achieves a large number of correct predictions in the "No relation" category, where 473 samples were correctly classified out of 576 (per-class accuracy = 82.1%), and in the "Binds" category, where 412 out of 433 samples were correctly classified (per-class accuracy = 95.2%). Importantly, the model does not collapse in the less-represented contextual categories: in the "Increases Expression" category, 96 out of 103 samples were correctly classified (per-class accuracy = 93.2%), and in the "Decreases Expression" category, 45 out of 50 samples were correctly classified (per-class accuracy = 90.0%). This distribution shows that the model retains good sensitivity to relatively rare classes, which is crucial in biological applications that are typically characterized by data imbalance.

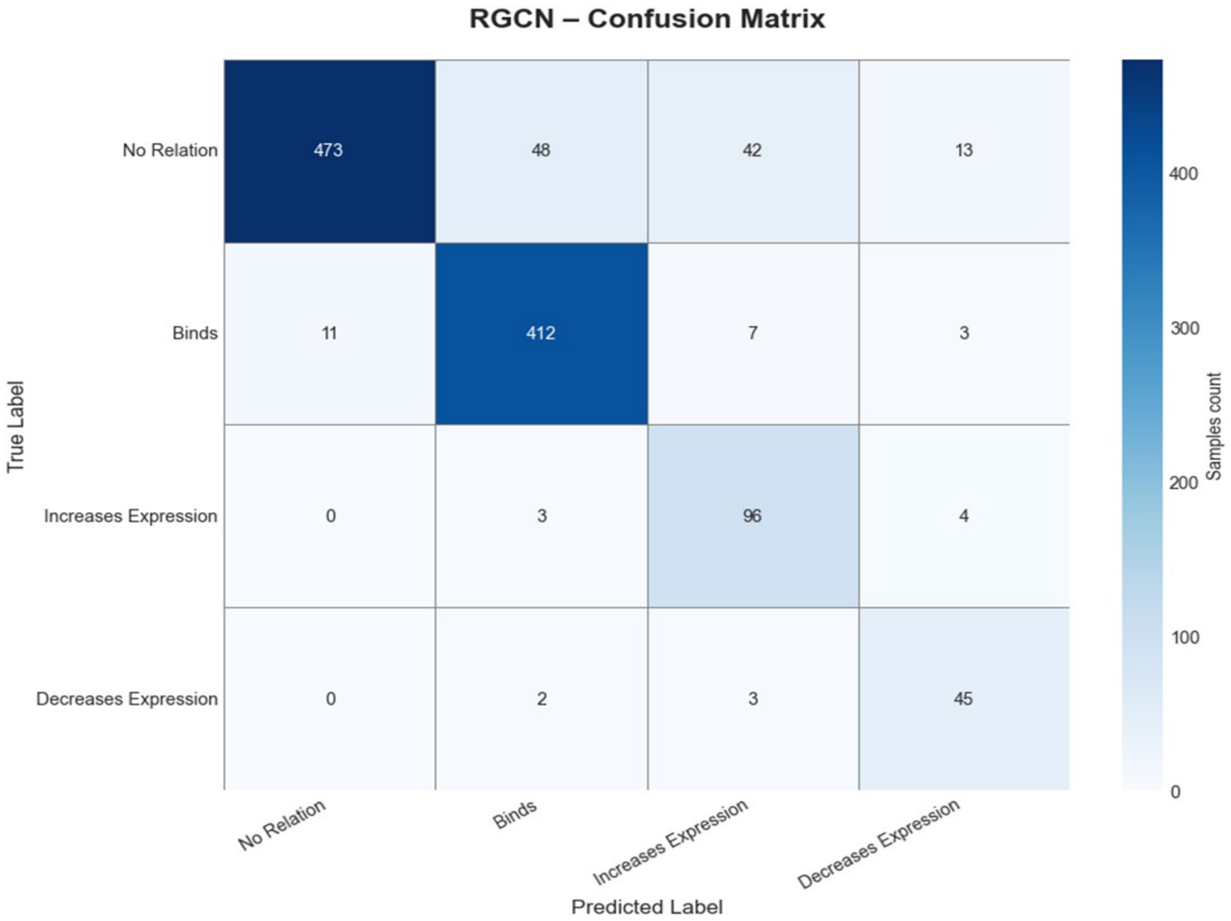


Figure 4.5: Confusion Matrix for the RGCN Multi-class Classification Model

The classification report provides a more detailed quantitative picture of the model's performance at the level of each class. High Precision values in the "No relation" category (about 0.9773) show that the model rarely predicts the existence of a relationship when it does not actually exist, reducing false positives in this category. In contrast, high Recall in the "Binds" category (about 0.9515) shows that the model is able to detect most cases of true binding, which is important in the context of drug interaction discovery. In the contextual categories "Increases Expression" and "Decreases Expression", although Precision is relatively lower, F1-scores remain good (about 0.7649 and 0.7826), reflecting an acceptable balance between precision and recall in classes with limited support.

Table 4.2: Classification Report for the RGCN Model for Multi-class Classification

support	F1-score	recall	Precision	
576	0.8925	0.8212	0.9773	No relation
433	0.9176	0.9515	0.8860	Binds
103	0.7649	0.9320	0.6486	Increase Expression
50	0.7826	0.9000	0.6923	Decrease Expression
1162	0.8830			Accuracy
1162	0.8394	0.9012	0.8011	Macro avg
1162	0.8858	0.8830	0.9019	Weighted avg

RGCN proves to be a powerful model in representing the overall structure of a graph, with a non-negligible ability to handle delicate relationships, which makes it a solid foundation for any subsequent hybrid structure.

4.3.2 RGAT Model

The Advanced_RGAT model is based on a multi-head attention mechanism, which allows it to identify the most important neighbors for each node rather than applying a uniform aggregation to all neighbors. This structural difference is clearly reflected in its training curves. In the figure showing the loss curves, it can be seen that the training loss gradually decreases until it reaches around 0.1787, while the validation loss stabilizes at a relatively higher value of approximately 0.2462. This is the largest gap between the training and validation curves compared to the RGCN, suggesting that the model is more sensitive to data distribution and perhaps less stable in generalizing to unseen data. In the F1-score curves, the model's performance gradually increases until F1 on the training set reaches around 0.7504, and on the validation set to 0.7545, which are lower values than those achieved by RGCN, but remain acceptable in the context of a model that focuses more on fine-grained relationships than on general patterns.

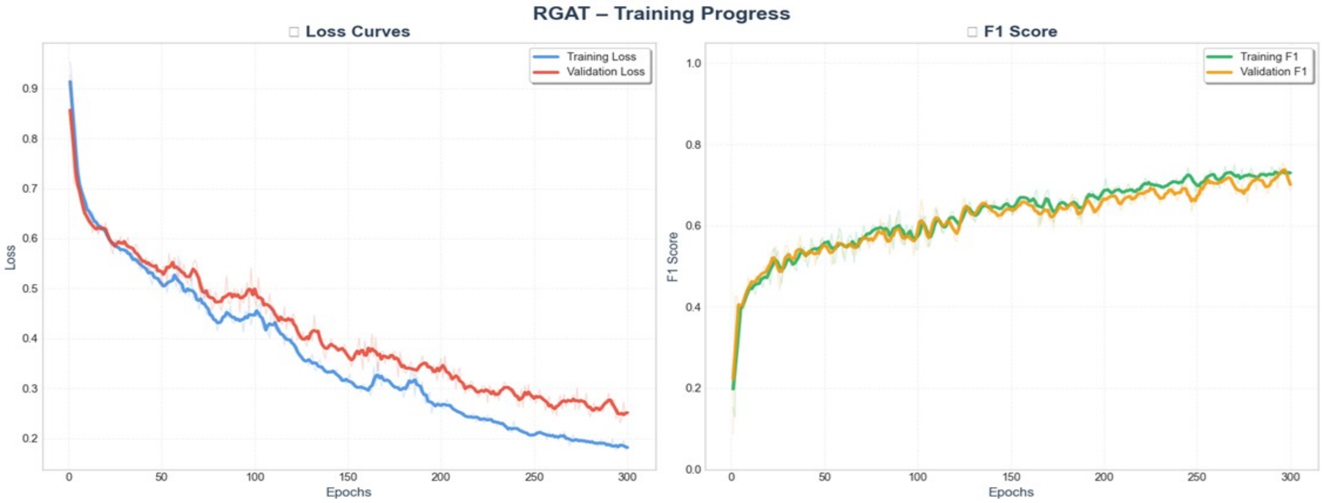


Figure 4.6: Loss curves and F1-score for the multi-class classification RGAT model

The RGAT's confusion matrix clearly shows this trend. In the "No relation" category, the model achieves 415 correct predictions out of 576, which is lower than that achieved by the RGCN, indicating that the model is less rigorous in distinguishing between the presence and absence of a relationship. In contrast, in the contextual categories "Increases Expression" and "Decreases Expression," the model performs strongly, with 90 out of 103 samples correctly classified in the "Increases Expression" category, and 43 out of 50 in the "Decreases Expression" category. This suggests that the attention mechanism enables the model to pick up precise signals associated with gene expression changes, even at the expense of some accuracy in the dominant categories.

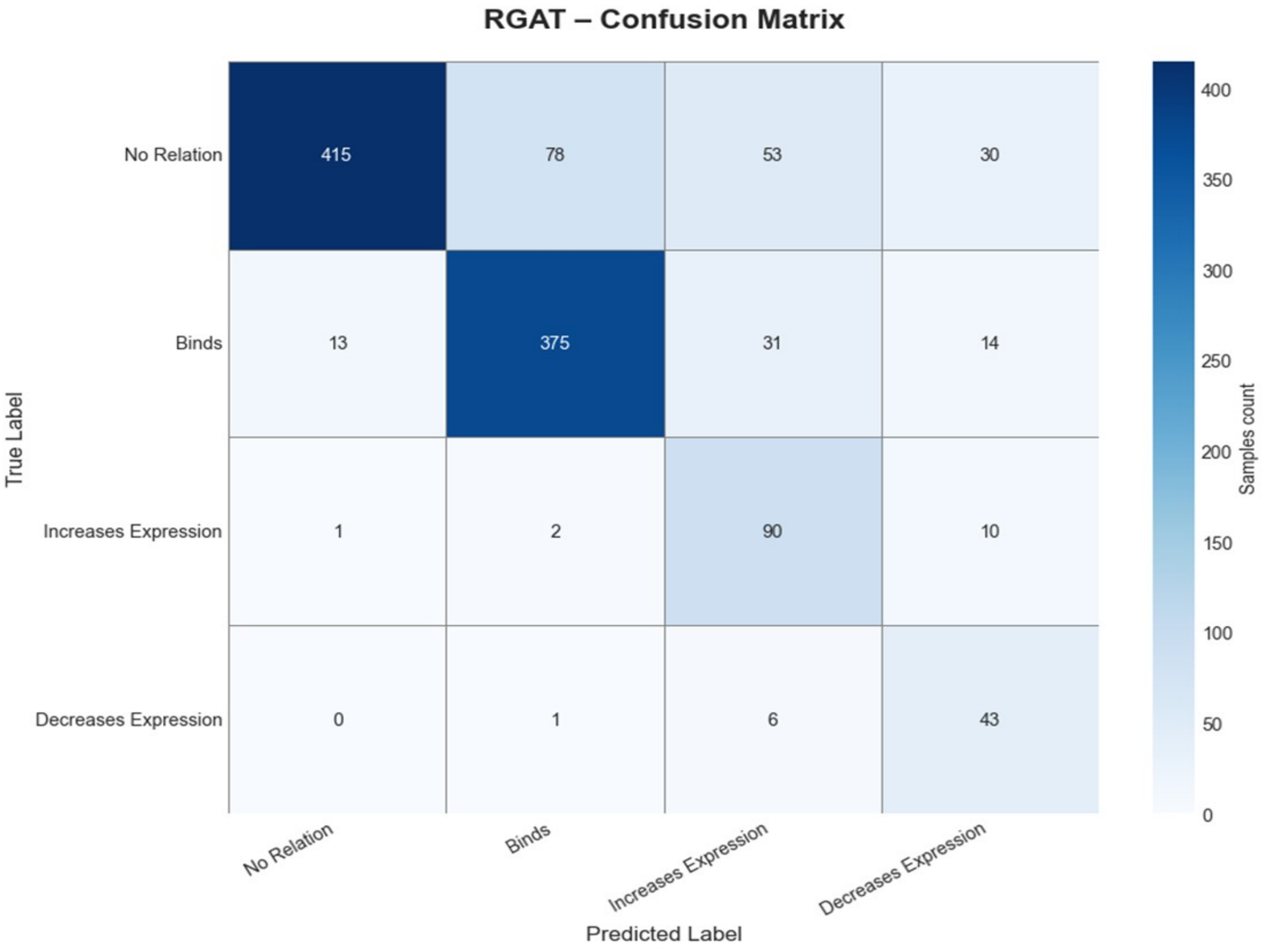


Figure 4.7: Confusion Matrix of the RGAT Multi-class Classification Model

High Recall values in contextual categories (close to 0.874 in "Increases Expression" and 0.860 in "Decreases Expression") show that the model tends to detect most of the true positive cases in these categories, but it does so with lower Precision (about 0.500 and 0.443, respectively), which means that a portion of the positive predictions in these categories are false positives. This pattern reflects a more "sensitive" model than a "rigorous" one—i.e., it prefers to capture as many real cases as possible, even if it leads to an increase in false positives. In dominant categories, such as "No relation" and "Binds," performance remains good but lower than RGCN, confirming that RGAT plays more of a complementary role than a complete alternative in this context.

Table 4.3: Classification Report for the RGAT Multi-class Classification Model

	Precision	recall	F1-score	support
No relation	0.9674	0.7205	0.8259	576
Binds	0.8224	0.8661	0.8436	433
Increase Expression	0.5000	0.8738	0.6360	103
Decrease Expression	0.4433	0.8600	0.5850	50
Accuracy			0.7943	1162
Macro avg	0.6833	0.8301	0.7226	1162
Weighted avg	0.8494	0.7943	0.8053	1162

4.3.3 Hybrid model

The hybrid model that combines Advanced_RGCN and Advanced_RGAT addresses the individual limitations of each, taking advantage of the power of RGCN in representing the overall structure of the graph, and the sensitivity of RGAT to precise contextual relationships. In this model, the graph is passed through both models in parallel to extract independent predictions for each drug-gene pair, and these predictions are then combined using a weighted averaging mechanism (70% RGCN + 30% RGAT) to produce the final classification probabilities. This design allows the hybrid model to leverage both structural and contextual perspectives simultaneously, with greater emphasis given to the RGCN model based on its superior performance.

The classification report for the hybrid model clearly reflects the success of this integration, with the overall accuracy rising to 88.90%, which is higher than the accuracy of both RGCN and RGAT. The Macro F1 and Weighted F1 values also show a strong balance between the categories, with values of approximately 0.87 and 0.90 respectively, which shows that the model does not perform well in one category at the expense of another, but rather maintains a high level of performance across all categories. In dominant categories such as "No relation" and "Binds", the model maintains a high level of Precision and Recall, while the contextual categories "Increases Expression" and "Decreases Expression" show a better balance between precision and recall compared to RGAT, not just in terms of sensitivity.

Table 4.4: Ensemble RGCN + RGAT for Multi-class Classification

	Precision	recall	F1-score	support
No relation	0.94	0.89	0.91	576
Binds	0.89	0.95	0.92	433
Increase Expression	0.83	0.82	0.82	103
Decrease Expression	0.81	0.84	0.82	50
Accuracy			0.90	1162
Macro avg	0.87	0.87	0.87	1162
Weighted avg	0.90	0.90	0.90	1162

The hybrid model's confusion matrix provides strong visual evidence of this balance. The model correctly classified 479 samples out of 576 in the "No relation" category, which is a clear improvement over RGAT and slightly superior to RGCN, meaning that the hybrid model was able to leverage the rigor of RGCN in this class while maintaining some of RGAT's sensitivity. In the "Binds" category, the model achieves 413 correct predictions out of 433, the highest value among all models, reflecting a strong ability to capture direct drug-gene relationships. In the contextual categories, the model retains excellent performance, with 96 out of 103 samples correctly categorized in "Increases Expression" and 45 out of 50 in "Decreases Expression", results that are almost identical to the best achieved by RGCN, with the implicit use of contextual information provided by RGAT.

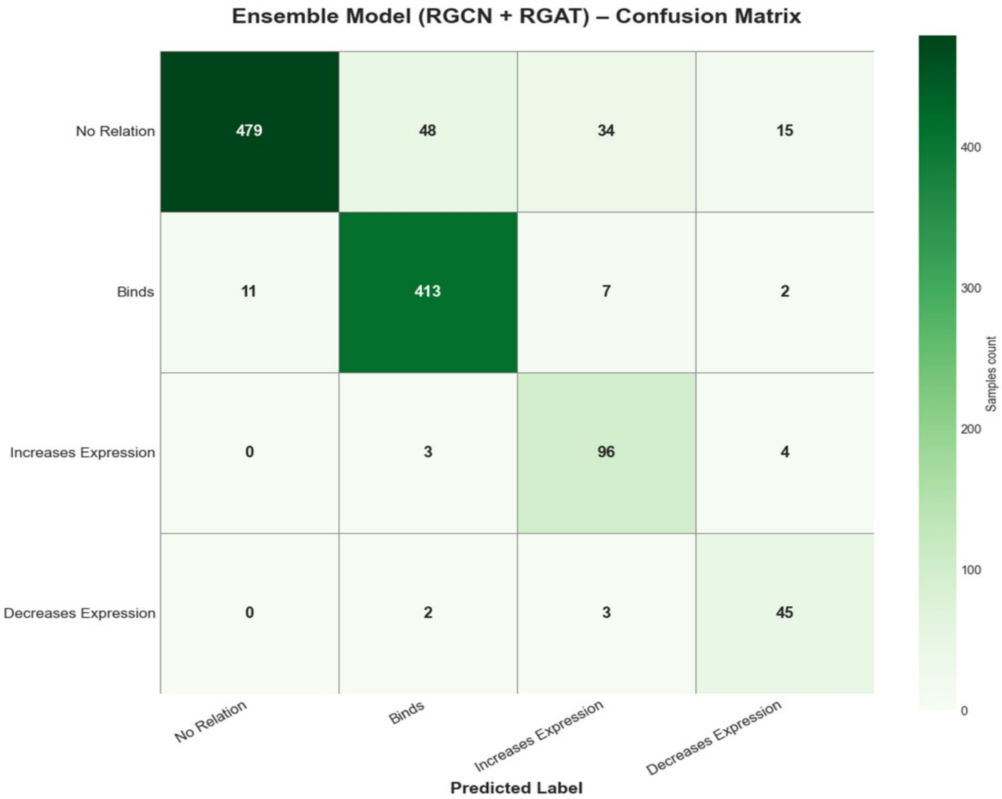


Figure 4.8: Confusion Matrix for Ensemble RGCN+RGAT Model for Multi-class Classification

The hybrid model represents a conceptual integration between a structural representation based on fixed aggregation of relationships, and a contextual representation based on attention directed toward the most important neighbors. The figures and tables associated with this model show that this integration has resulted in a more balanced and generalizable model, capable of dealing with both common and rare categories, without sacrificing accuracy or sensitivity.

4.4 Ablation Analysis of Core Components

4.4.1 Introduction and objectives of the ablation study

The efficiency of intelligent models depends on the final numerical result as well as on the validity and consistency of the internal components and algorithms chosen. The purpose of the Ablation Study is to deconstruct the proposed model and conduct comparative experiments by isolating specific variables (e.g., loss functions, feature dimensions, and architecture) to study the impact of each on performance. The main purpose of using this approach is to demonstrate that the final model is built on the basis of a rigorous scientific trade-off that balances predictability, reliability, and interpretability.

4.4.2 Comprehensive Comparative Trials Table

Three systematic ablation studies were conducted to assess the impact of the core components of the proposed framework:

1. the dimensions of the input features.
2. the loss function and class weights
3. the confidence threshold in the inference phase.

The results show that the final design (131 features + Focal Loss with weights + threshold 0.75) represents the optimal balance between statistical accuracy and biological reliability, as shown in the following table:

Table 4.5 – Detailed results of all Ablation Study Experiments

experience	settings	Overall Accuracy	No Relation	Binds	Increases Expression	Decreases Expression	F1_score	Notes
Features Dimensions	128 features (DeepWalk embeddings only)	94.15%	90.6% (522/576)	98.6% (427/433)	94.2% (97/103)	96.0% (48/50)	0.95	Exceptional performance in the rare category but with the risk of overfitting
	129 features (+ one-hot node type encoding)	94.75%	93.4% (538/576)	98.4% (426/433)	92.2% (95/103)	84.0% (42/50)	0.95	Slight improvement in overall accuracy with deterioration in category balance
	131 features (+ Degree + PageRank)	88.90%	83.2% (479/576)	95.4% (413/433)	93.2% (96/103)	90.0% (45/50)	0.90	High methodological stability across rigorous data splits
Loss functions	Focal Loss + weights [0.50, 0.67, 2.81, 5.88]	88.90%	83.2% (479/576)	95.4% (413/433)	93.2% (96/103)	90.0% (45/50)	0.90	Perfect balance (Macro F1 = 0.87) across all classes
	Cross Entropy + Weights	93.72%	92.0% (531/576)	98.2% (425/433)	90.3% (93/103)	82.0% (41/50)	0.93	Excessive bias toward the dominant class (Binds) with equilibrium breakdown (Macro F1 = 0.69)
	Cross Entropy without weights	92.17%	92.7% (534/576)	93.1% (403/433)	91.3% (94/103)	80.0% (40/50)	0.89	Moderate degradation in rare categories (80.0%) despite high overall accuracy (92.17%), indicating incomplete mitigation of class imbalance
Confidence Threshold	Without threshold	73.19%	67.8% (389/576)	79.0% (342/433)	68.0% (70/103)	96.0% (48/50)	0.86	High classification noise: low-confidence predictions (<0.5) forced as final classifications
	With threshold (0.75)	88.90%	83.2% (479/576)	95.4% (413/433)	93.2% (96/103)	90.0% (45/50)	0.90	Systematic noise filtering: transforming the model into a reliable hypothesis generator

Although models with 128 and 129 features achieved superior numerical performance in some experimental splits, the choice of the final model with 131 features was a strategic decision that transcended the language of abstract numbers. This representation, which integrates the topological context (Degree and PageRank) with the semantic context, enhances the model's ability to generalize when applied to different biological networks, and serves as the essential bridge for the linguistic model (LLM) to generate scientifically grounded explanations reflecting the structural importance of nodes within the knowledge graph. The primary purpose of incorporating these structural indicators was to provide the system with comprehensive contextual information, enabling the language interpreter to deliver a sober scientific analysis commensurate with the structural significance of each node.

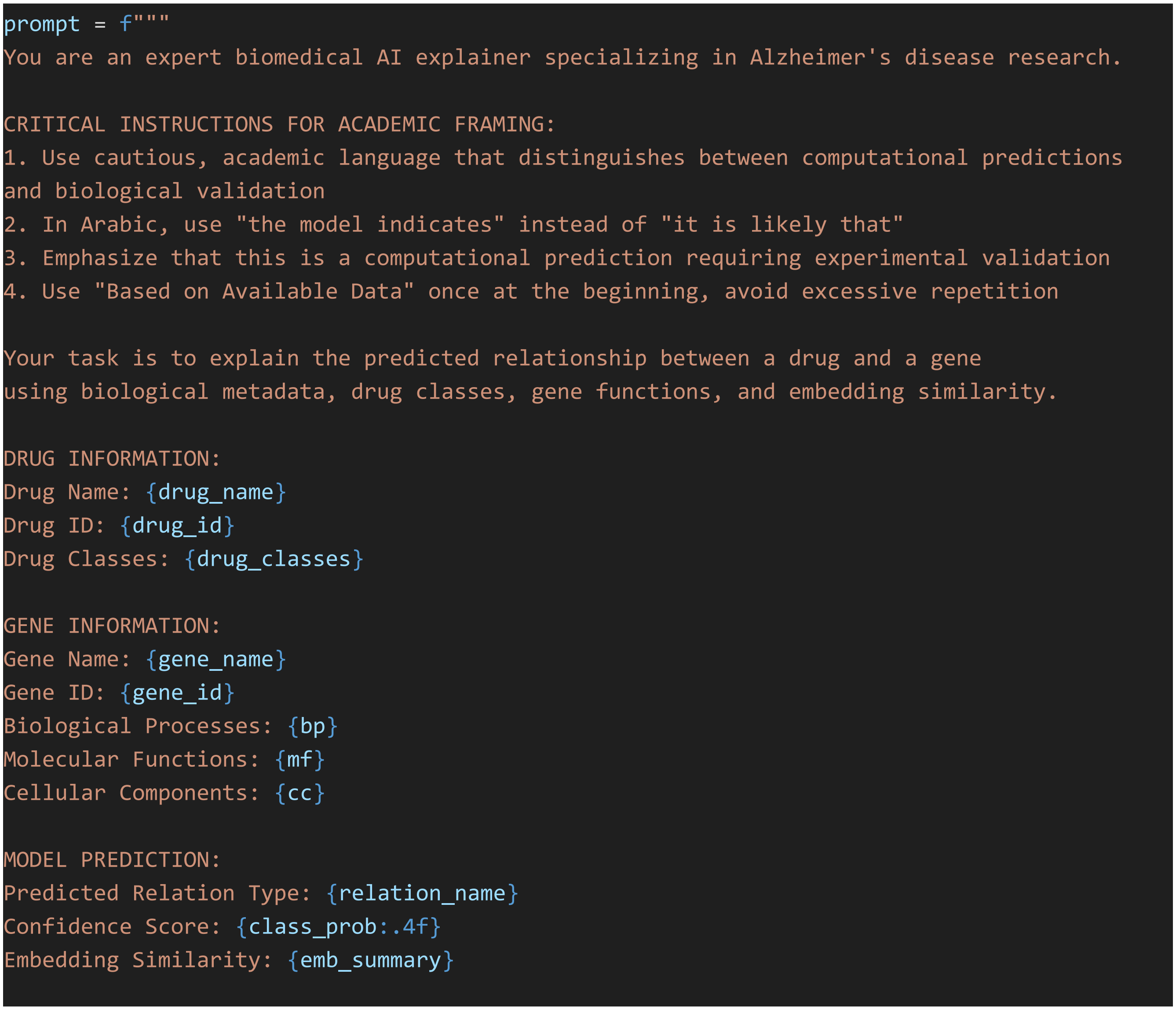
The comparison confirms that the choice of Focal Loss was the optimal methodological choice to address the biological challenge of the Decreases Expression category (<5%). While the traditional Cross Entropy function recorded a significant degradation in the rare category when used without weights (80.0% versus 90.0% for Focal Loss), and the addition of class weights alone led to excessive bias toward dominant categories despite achieving 93.72% overall accuracy (Macro F1 imbalance), Focal Loss was able to focus learning on the hard samples without sacrificing the performance of dominant categories, achieving an ideal balance (Macro F1 = 0.87) that ensures prediction reliability across all interaction types.

The drastic difference between a model with a threshold of 0.75 (88.90%) and without it (73.19%) reveals a methodological dilemma, as mandatory classification generates enormous noise by imposing predictions with weak probability support. The adoption of a threshold of 0.75 reflects the principle of methodological caution used in clinical research, where flimsy hypotheses are rejected to avoid wasting resources in unreliable laboratory experiments. The threshold has proven its ability to filter out noise and transform the model from a classification machine into a high-certainty hypothesis generator.

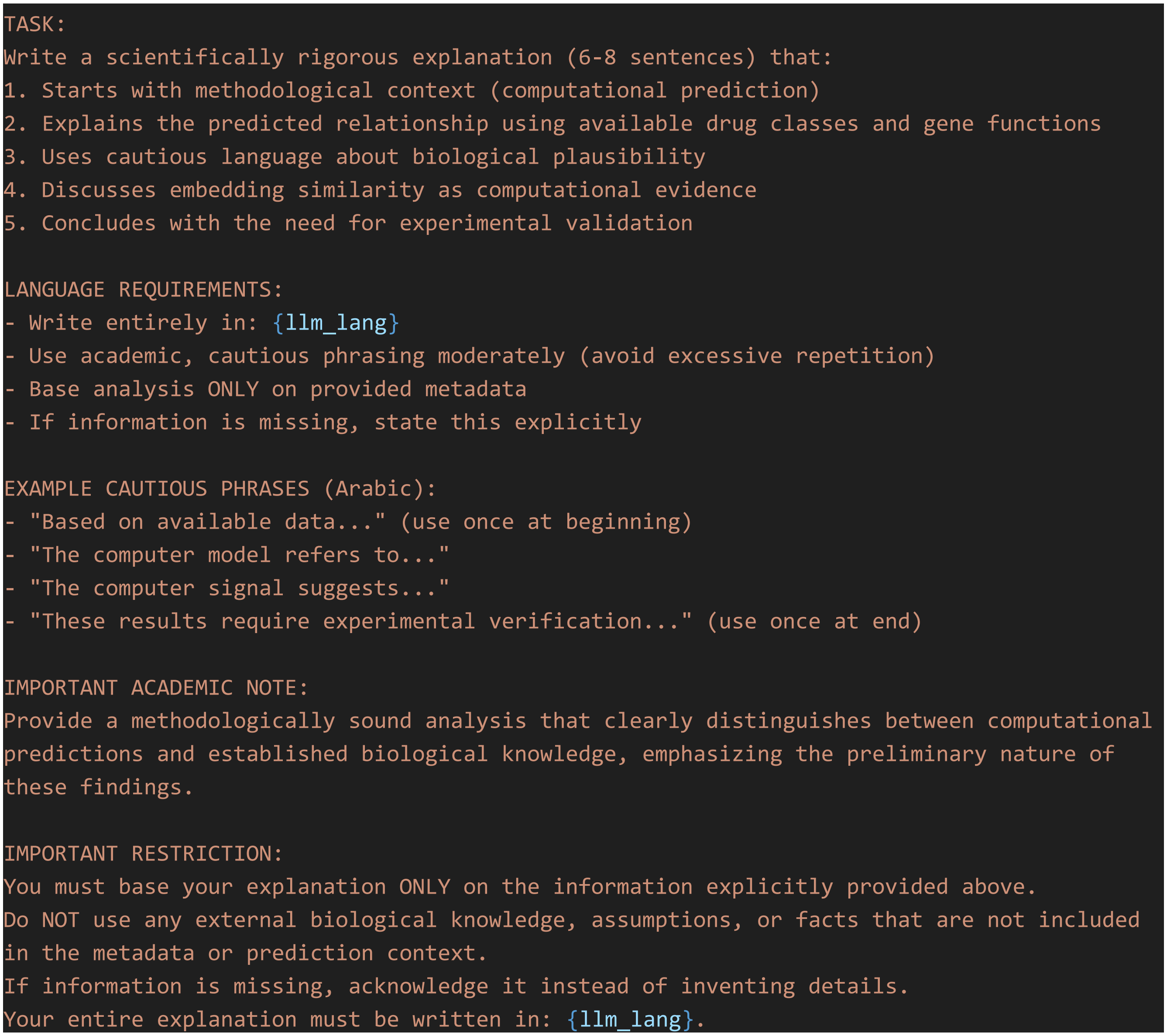
A small margin of maximum accuracy has been sacrificed in exchange for gaining structural stability, biological interpretability, and reducing false positives, in order to conform to best practices in medical AI, as the model becomes an integrated research framework capable of providing solid hypotheses that serve the actual path of drug discovery in Alzheimer's disease.

4.5 Mechanism for Generating Interpretive Reports in a Hybrid System

The explanatory reports produced by the hybrid system are based on a carefully crafted methodological prompt, which directs the large language model towards producing a disciplined academic analysis that reflects the outputs of the graphical model without adding any external knowledge. This prompt is based on a set of strict instructions that oblige the model to use careful scientific language, always beginning with clarifying that the analysis is based on a computational prediction derived from available data, while emphasizing the need to distinguish between mathematical signals and established biological facts. The prompt also forces the model to rely exclusively on the information provided during the prediction—such as the drug name, its drug class, the gene name, its biological processes, molecular functions, cellular components, the predicted relationship type, the confidence score, and the embedding similarity—without allowing it to invoke any external knowledge or assumptions that are not explicitly stated.



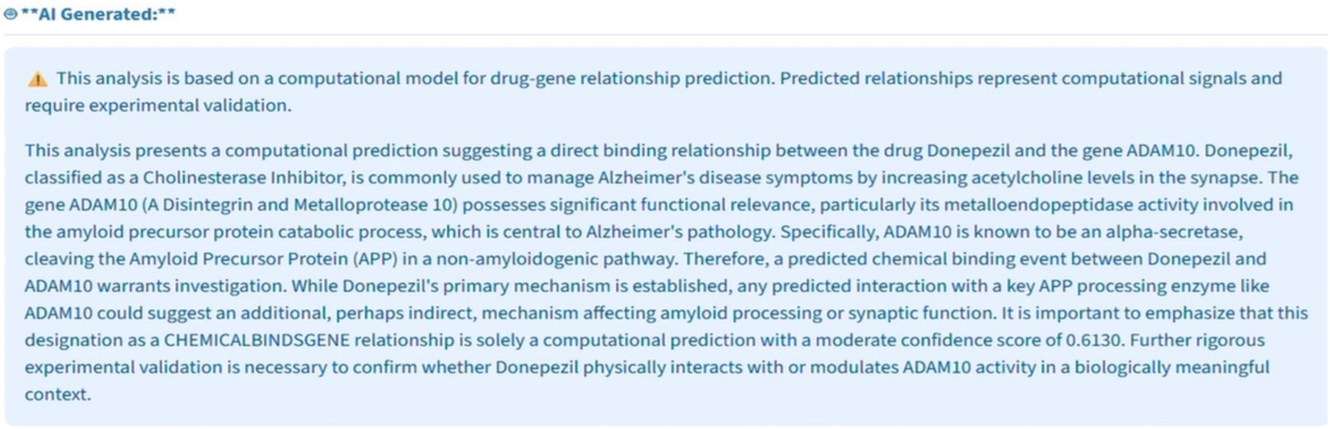
```
prompt = f"""
You are an expert biomedical AI explainer specializing in Alzheimer's disease research.

CRITICAL INSTRUCTIONS FOR ACADEMIC FRAMING:
1. Use cautious, academic language that distinguishes between computational predictions
and biological validation
2. In Arabic, use "the model indicates" instead of "it is likely that"
3. Emphasize that this is a computational prediction requiring experimental validation
4. Use "Based on Available Data" once at the beginning, avoid excessive repetition

Your task is to explain the predicted relationship between a drug and a gene
using biological metadata, drug classes, gene functions, and embedding similarity.

DRUG INFORMATION:
Drug Name: {drug_name}
Drug ID: {drug_id}
Drug Classes: {drug_classes}

GENE INFORMATION:
Gene Name: {gene_name}
Gene ID: {gene_id}
Biological Processes: {bp}
Molecular Functions: {mf}
Cellular Components: {cc}

MODEL PREDICTION:
Predicted Relation Type: {relation_name}
Confidence Score: {class_prob:.4f}
Embedding Similarity: {emb_summary}
```



```
TASK:
Write a scientifically rigorous explanation (6-8 sentences) that:
1. Starts with methodological context (computational prediction)
2. Explains the predicted relationship using available drug classes and gene functions
3. Uses cautious language about biological plausibility
4. Discusses embedding similarity as computational evidence
5. Concludes with the need for experimental validation

LANGUAGE REQUIREMENTS:
- Write entirely in: {llm_lang}
- Use academic, cautious phrasing moderately (avoid excessive repetition)
- Base analysis ONLY on provided metadata
- If information is missing, state this explicitly

EXAMPLE CAUTIOUS PHRASES (Arabic):
- "Based on available data..." (use once at beginning)
- "The computer model refers to..."
- "The computer signal suggests..."
- "These results require experimental verification..." (use once at end)

IMPORTANT ACADEMIC NOTE:
Provide a methodologically sound analysis that clearly distinguishes between computational
predictions and established biological knowledge, emphasizing the preliminary nature of
these findings.

IMPORTANT RESTRICTION:
You must base your explanation ONLY on the information explicitly provided above.
Do NOT use any external biological knowledge, assumptions, or facts that are not included
in the metadata or prediction context.
If information is missing, acknowledge it instead of inventing details.
Your entire explanation must be written in: {llm_lang}.
```

Thanks to this design, the resulting reports are fully aligned with the content of the prompt, as the model begins by providing a systematic context that shows the presented relationship is the result of a computational analysis, and then proceeds to interpret the predicted relationship by linking the drug's characteristics to the gene's biological functions as reported in the data. Embedding similarity is used as additional mathematical evidence to support the indication of a possible relationship, while maintaining precise explanatory language that avoids assertiveness or overgeneralization. At the end of the report, the model adheres to an academic conclusion confirming that the results are preliminary and require empirical verification before they can be adopted as confirmed biological knowledge. With this mechanism, the interpretive report becomes a complementary knowledge layer to the hybrid model's outputs, translating numerical predictions into a disciplined scientific interpretation consistent with the system's methodology.

The interpretation shown in Figure 4.7 is a clear practical example of how the hybrid system adheres to the methodological instructions specified in the prompt used to generate the interpretation. The report begins with a cautionary statement that the analysis is based on a computational model, which directly corresponds to the first part of the prompt, which requires the model to use careful academic language that distinguishes between computational prediction and biological facts. It is also noted that the report uses wording such as "this analysis provides a computational prediction" and "requires empirical verification", which are explicitly stated in the prompt within the cautious language instruction.



AI Generated:

⚠ This analysis is based on a computational model for drug-gene relationship prediction. Predicted relationships represent computational signals and require experimental validation.

This analysis presents a computational prediction suggesting a direct binding relationship between the drug Donepezil and the gene ADAM10. Donepezil, classified as a Cholinesterase Inhibitor, is commonly used to manage Alzheimer's disease symptoms by increasing acetylcholine levels in the synapse. The gene ADAM10 (A Disintegrin and Metalloprotease 10) possesses significant functional relevance, particularly its metalloendopeptidase activity involved in the amyloid precursor protein catabolic process, which is central to Alzheimer's pathology. Specifically, ADAM10 is known to be an alpha-secretase, cleaving the Amyloid Precursor Protein (APP) in a non-amyloidogenic pathway. Therefore, a predicted chemical binding event between Donepezil and ADAM10 warrants investigation. While Donepezil's primary mechanism is established, any predicted interaction with a key APP processing enzyme like ADAM10 could suggest an additional, perhaps indirect, mechanism affecting amyloid processing or synaptic function. It is important to emphasize that this designation as a CHEMICALBINDSGENE relationship is solely a computational prediction with a moderate confidence score of 0.6130. Further rigorous experimental validation is necessary to confirm whether Donepezil physically interacts with or modulates ADAM10 activity in a biologically meaningful context.

Figure 4.9: Example of Interpretation

The report relies on the drug class (Cholinesterase Inhibitor) and on the molecular functions and biological processes of the ADAM10 gene as reported in the metadata, without adding any external information, which reflects the model’s adherence to the explicit condition in the prompt that prevents it from using any knowledge that is not present in the inputs. The report also incorporates the predicted relationship type (CHEMICALBINDSGENE) and a confidence score (0.6130) into the interpretation context, just as the prompt requests in the MODEL PREDICTION section.

The reference to embedding similarity as supporting mathematical evidence corresponds to the part of the prompt that asks the model to "discuss similarity as computational evidence" without considering it as biological evidence. The report concludes with a statement emphasizing the need for empirical validation, which is also a mandatory part of the prompt. Thus, it is clear that the report shown in Figure 4.7 is not just a general explanatory text, but rather a precise and direct response to each of the prompt’s instructions, reflecting the system’s consistency and ability to produce systematically disciplined interpretive report.

Chapter Five

Discussion and conclusions

5.1 Interpretation of Results and System Evaluations

Results obtained from the developed graph models reveal a clear ability of the system to represent pharmacogenetic relationships in a way that goes beyond traditional methods based on surface similarity or statistical association. The Advanced_RGCN model, which relies on structural aggregation of relationships, has shown strong performance in categories with a clear structure within the graph, such as "No relation" and "Binds". This performance reflects the model's ability to capture the general structural patterns that bind the nodes together, including bond density, neighbor distribution, and topological characteristics that are important indicators in understanding biological relationships. The model also showed clear stability in the training and validation curves, suggesting that the structural representation used is able to generalize without falling into the problem of overfitting, which is an important indicator of representation quality.

In contrast, the Advanced_RGAT model, which relies on the relational attention mechanism, showed a higher ability to capture precise contextual relationships, especially in categories that require more complex functional discrimination such as "Increases Expression" and "Decreases Expression." This performance reflects the effectiveness of the attention mechanism in identifying the most influential neighbors for each node, allowing the model to focus on subtle cues that may not be obvious in the overall structure of the graph. Although the RGAT's performance was less stable in dominant categories, its high sensitivity to fine-grained relationships is an important strength, especially in biological applications that require precise functional explanation.

The hybrid model, which combines RGCN and RGAT, achieved the best overall performance, integrating structural and contextual representations into a single architecture, allowing the system to have a multidimensional view of biological relationships. This combination was reflected in the highest overall accuracy (88.9%) and the highest Macro F1 value (0.87), with a clear balance across all four categories. This balance is an important indication that the model does not rely on a single category or type of relationship, but rather has a balanced capability to handle both structural and contextual relationships. The hybrid model's confusion matrices also showed a significant reduction in cross-class errors, demonstrating that the combination of the two individual models not only raised performance but also improved the quality of classification distinction.

Systematic ablation studies were conducted to rigorously validate the contribution of each architectural component. These studies examined the impact of feature dimensionality (128 vs. 129 vs. 131 features), loss function selection (Focal Loss versus Cross Entropy variants), and confidence thresholding (0.75 versus no threshold) on model robustness. The results confirmed that the final design choices—131-dimensional feature representation incorporating topological metrics, Focal Loss with class weights to address severe class imbalance, and a 0.75 confidence threshold for noise filtering—collectively constitute a methodologically sound configuration that balances predictive accuracy with biological reliability.

In addition, the interpretive reports generated by the system, which are based on a controlled systematic prompt, have demonstrated a clear ability to translate numerical outputs into a cautious scientific interpretation based on available data. These reports have contributed to enhancing the usability of the system in computational biology, not only providing predictions but also metadata-supported functional interpretations, emphasizing that the results are computational in nature and require experimental validation. This explanatory aspect is a qualitative addition to the system, as it connects numerical predictions to the biological context, helping researchers assess the plausibility of predicted relationships.

Overall, the results show that the developed hybrid system represents an important step towards building a computational framework capable of supporting the discovery of pharmacogenetic relationships, generating testable biological hypotheses, and providing scientific explanations that help researchers understand the functional context of predicted relationships. The results also confirm that the integration of structural and contextual representation is the optimal approach to dealing with complex biological relationships, and that deep graph models are capable of providing scientific value beyond traditional methods of analyzing biological data.

5.2 Research Limitations

The research was constrained within several methodological boundaries that shaped the scope and implementation of this study. These limitations, while defining the research boundaries, also provide clarity on the specific context in which the results should be interpreted.

Disease Scope Limitation: The computational framework was developed specifically for Alzheimer's disease and does not extend to other neurological or medical conditions. This focused approach, while ensuring depth and specificity in Alzheimer's research, limits the generalizability of the findings to broader drug discovery contexts.

Data Source Constraints: The study relied exclusively on existing biological and drug data sources from the AlzKB knowledge graph, without incorporating new primary biological data collection or conducting laboratory/clinical experiments. This constraint ensures that all predictions are computational in nature and require experimental validation for clinical application.

Methodological Boundaries: The research utilized existing algorithms and large language models rather than developing novel ones from scratch. While this approach leverages proven technologies and ensures implementation feasibility, it limits the potential for algorithmic innovations that might address specific challenges in Alzheimer's drug discovery.

Validation Limitation: The absence of experimental validation means that all predicted drug-gene interactions remain computational hypotheses. The system's predictions, despite high confidence scores, require rigorous laboratory confirmation before clinical application.

These constraints collectively define a focused research scope that prioritizes computational rigor and reproducibility while acknowledging the need for future experimental validation and broader application domains.

5.3 Future recommendations

Based on the results achieved by the hybrid system, a set of future recommendations can be proposed aimed at enhancing its capabilities, improving its accuracy, and expanding its application in computational biology.

First, expanding data sources is an essential step to raise the quality of biological representation. Integrating gene expression data, 3D protein structure data, molecular pathway data, as well as clinical or mutation data, can enrich the graph and give models a greater ability to capture complex relationships that are not apparent in structural data alone. Incorporating temporal or contextual data—such as cellular responses under different conditions—may open the door to more dynamic models that can predict changing relationships over time.

Second, more advanced interpretation mechanisms can be developed within the system, so that the reports not only interpret the predicted relationship but also provide an analysis of attention effects or causal pathways within the graph. This type of deep interpretation can help researchers understand the "why" behind the model's decision, not just the "what" it predicted, enhancing confidence in the system and making it more usable in experimental biology. Techniques such as Node Influence Analysis could be integrated to provide more accurate insights into the factors that influenced the prediction.

Third, scaling up the application is a natural step in light of the results achieved. The current system can be upgraded to serve additional tasks such as drug repurposing, side effect prediction, disease network analysis, or even supporting treatment decisions in personalized medicine. The model can also be extended to other types of biological relationships, such as protein-protein interactions or gene-disease relationships, transforming the system into a comprehensive analytical platform.

Finally, it must be emphasized that experimental validation is an indispensable step in any biological application. Computational predictions, no matter how accurate, are preliminary signals that need rigorous laboratory confirmation before they can be adopted as solid biological knowledge. Therefore, it is recommended to establish an integrative framework that links model results to laboratory experiments, so that relationships of high confidence or functional importance are selected for validation, enhancing the scientific value of the system and ensuring that its outputs are applicable in future research.

REFERENCES

[1] Dobreva J, Simjanoska Misheva M, Mishev K, Trajanov D, Mishkovski I. A Unified Framework for Alzheimer's Disease Knowledge Graphs: Architectures, Principles, and Clinical Translation. Brain Sci. 2025 May 19;15(5):523. doi: 10.3390/brainsci15050523. PMID: 40426694; PMCID: PMC12110335.

[2] Romano J, Truong V, Kumar R, Venkatesan M, Graham B, Hao Y, Matsumoto N, Li X, Wang Z, Ritchie M, Shen L, Moore J. The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research. J Med Internet Res 2024;26:e46777. URL: https://www.jmir.org/2024/1/e46777. DOI: 10.2196/46777

[3] 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.

[4] Wang, Z. P., 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. BioData Mining, 18, 51.

[5] Loesch, J., et al. (2024). Explaining graph neural network predictions for drug repurposing. SWAT4HCLS 2024.

[6] Yan, C., et al. (2024). Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. npj Digital Medicine, 7, 46.

[7] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling Relational Data with Graph Convolutional Networks. In The Semantic Web – ESWC 2018 (pp. 593–607). Springer.

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

[9] Busbridge, D., Sherburn, S., Cavallaro, P., & Hammerla, N. Y. (2019). Relational graph attention networks. arXiv preprint arXiv:1904.05813.

[10] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710).

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[12] Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

الجامعة السورية الخاصة

كلية الهندسة المعلوماتية

قسم الذكاء الاصطناعي

وعلوم المعطيات

أُعدّت هذه الأطروحة:

لإتمام مشروع التخرج الأول في مجال الذكاء الاصطناعي وعلوم المعطيات

بعنوان:

التنبؤ بالتفاعل بين الدواء والجين في مرض الزهايمر المعزز برسوم المعرفة البيانية:

منهجية تعتمد على الشبكات العصبونية البيانية(GNN) والنماذج اللغوية الكبيرة(LLM)

إعداد:

مريم عادل عبد العال

إشراف:

د. ميساء أبو القاسم _ م.آية الأسود

الفصل الدراسي الأول

2025/2026