Predicting Gene–Disease Associations with GNNs

By Adam Bublil, Shelly Levy, Tom Saacks and Roy Efroni

Abstract

We investigate whether Graph Neural Networks (GNNs), combined with biomedical embeddings, can accurately predict novel gene–disease associations. Using the TBGA Gene–Disease Association dataset, we evaluate multiple embedding strategies and architectures, including GCN, GAT, and Transformer, in both binary and multi-class classification tasks. Our best model is a GNN Transformer with LLM-based embeddings that achieved high performance and uncovered biologically plausible false positives, such as the undocumented yet validated link between DNAJC19 and anemia. These results highlight the potential of GNNs as tools for biomedical discovery. Compared to recent benchmarks, our models achieved similar or better classification scores, highlighting the potential of GNNs as tools for biomedical discovery.

Introduction

Understanding gene–disease relationships is fundamental to biomedical research and drug discovery. Traditional methods rely on curated databases and expert-driven annotations, limiting their scalability. In recent years, **Graph Neural Networks (GNNs)** have emerged as powerful tools for modeling complex biological systems, particularly gene–disease graphs.

This study explores the research question: **Can Graph Neural Networks predict novel gene–disease associations from known biological interactions?**

To address this, we constructed a two-stage GNN pipeline that performs binary classification to identify candidate links, followed by multi-class classification to label their relationship types. We compared various node embedding strategies, including random, PCA, and domain-specific LLM-based embeddings. Our findings show that GNNs not only perform well on known data but also generalize to uncover new, biologically plausible associations. Additionally, Our model achieves performance on par with state-of-the-art GNN benchmarks reported in the literature, reinforcing their effectiveness for biomedical applications.

Literature Review

Our literature review underscores the growing utility of Graph Neural Networks (GNNs) in biological link prediction tasks. Gao et al. (2024) introduced Crop-GPA, a framework that combines graph representations with textual embeddings to identify gene–trait associations in crops. Their findings demonstrated that GCN-based architectures outperformed classical baselines, particularly when leveraging domain-specific embeddings—highlighting the importance of structural and semantic integration in biological graphs. Extending this paradigm to human biomedical data, Kishan KC et al. (2020) proposed the HOGCN model, which incorporates high-order message passing to aggregate multi-hop neighbor information. This approach enhanced classification accuracy and produced semantically coherent clusters of gene–disease associations. Collectively, these studies suggest that both first-order and high-order GNNs are well-suited for modeling complex biomedical knowledge graphs and hold substantial promise for uncovering novel biological relationships.

Recent studies offer valuable benchmarks for gene–trait and gene–disease association tasks. For example, Gao et al. (2024) introduced the Crop-GPA framework for predicting gene–phenotype associations in crops. Their GPA-BERT model combines graph neural networks with contextual embeddings derived from a pre-trained language model, demonstrating the potential of integrating textual and structural information. While their work focuses on crop genomics rather than biomedical data, the underlying task of link prediction is similar, making it a relevant point of comparison for evaluating methodological approaches like

ours. Similarly, Kishan KC et al. (2020) demonstrated the effectiveness of high-order graph convolutional networks (HOGCN) for gene–disease association prediction. Their model achieved strong performance on a gene–disease dataset (GDI), highlighting the benefits of multi-hop information aggregation in capturing complex biological relationships. While their dataset is not identical to ours, it is highly comparable in structure and objective, making the comparison relevant though not exact.
Both studies offer suitable comparison points given their use of graph-based models, biomedical link prediction objectives, and similar evaluation metrics. Our experimental setup and goals closely align with these prior works, allowing for a meaningful contextual evaluation.

Data Description

The dataset used in this study represents a biomedical knowledge graph constructed from curated gene–disease relationships. Each data instance corresponds to an edge between a gene and a disease, annotated with one of several relation types. The graph comprises two node types (info from Kaggle):

- **Genes** (9,569 unique entities)
- **Diseases** (7,499 unique entities)

yielding a total of 49,214 **edges** (gene–disease pairs), including both labeled and unlabeled associations. Among the labeled edges, each relationship is categorized as one of three biological types:

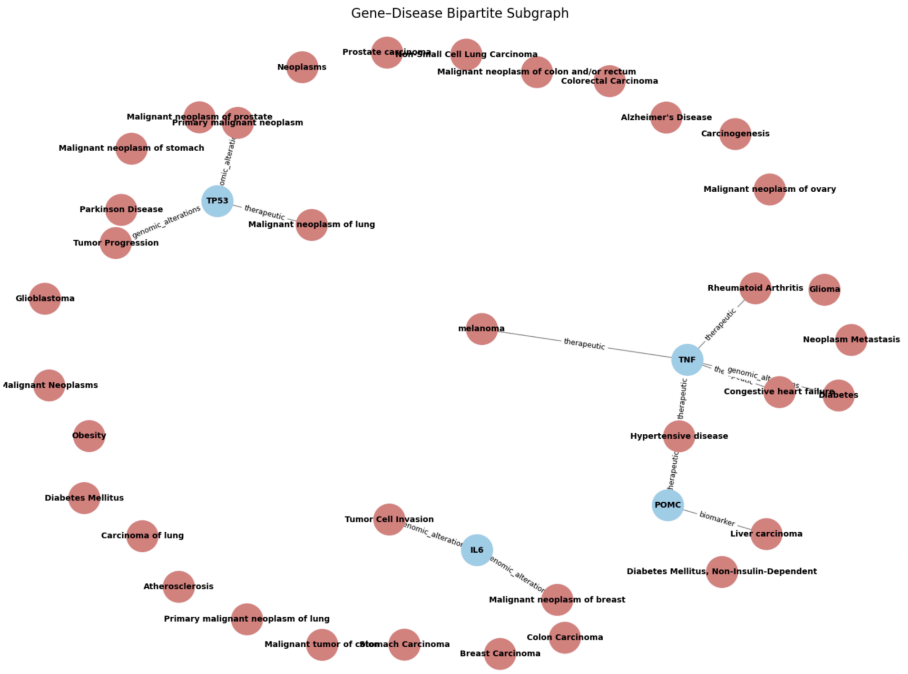
- **Therapeutic** - The gene is a target for treatment or its alteration can be directly addressed by a therapy.
- **Biomarker** - The gene is used as a diagnostic or prognostic indicator for the presence of the disease.
- **Genomic Alteration** - The gene contains mutations or structural changes that are linked to the disease.

Edges with no known association are labeled as NA, and constitute a majority of the dataset.

The dataset, as downloaded from Kaggle, was already pre-split into training, validation, and test sets, ensuring no overlap between them. This separation is particularly important given the nature of the data—it's not standard tabular data, but a graph-based structure with interconnected nodes and edges. In such cases, a naive random split could lead to information leakage due to the inherent dependencies between nodes. The dataset was provided as three separate JSON-formatted text files: TBGA_train.txt, TBGA_val.txt, and TBGA_test.txt.

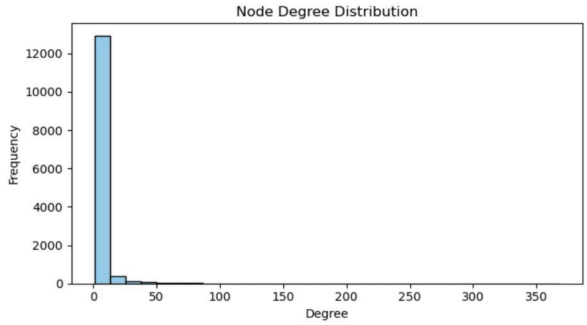
Each line in the file is a JSON object that contains:

- "text": a short biomedical sentence describing the relationship between a gene and a disease.
- "h": the head entity (gene), including an "id" field.
- "t": the tail entity (disease), including an "id" field.
- "relation": the type of connection ("Therapeutic", "Biomarker", "Genomic Alteration").

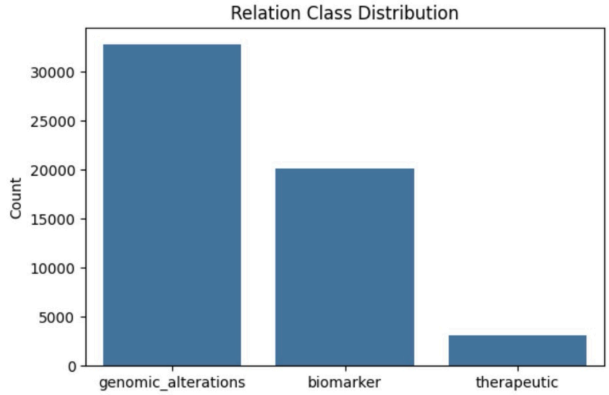


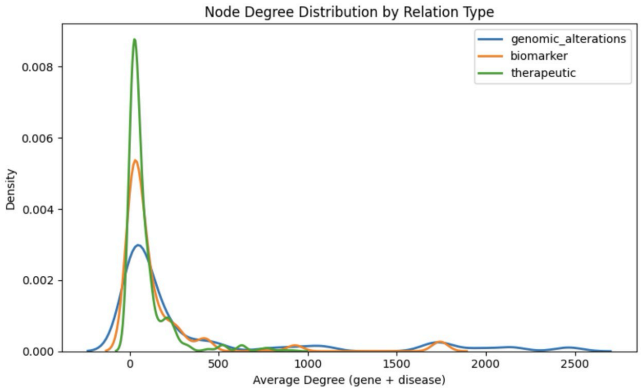
EDA

The network exhibits a skewed degree distribution, consistent with prior observations in biological systems. A small number of genes and diseases function as hubs, participating in numerous associations, while the majority have relatively few connections.

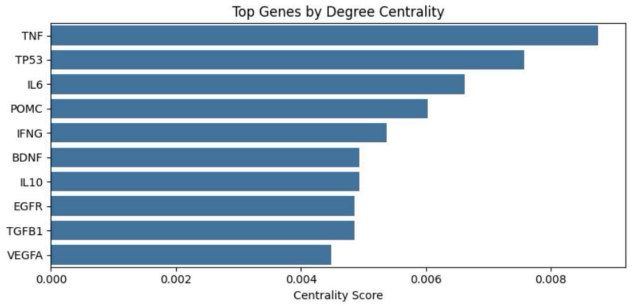


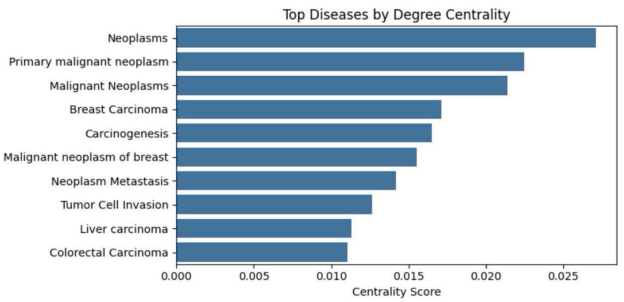
The dataset is imbalanced - genomic alterations dominate, followed by biomarker and fewer therapeutic links. Genomic alterations span a wider degree range, while others concentrate around low-degree nodes. These structural differences can affect GNN message passing and embedding quality.



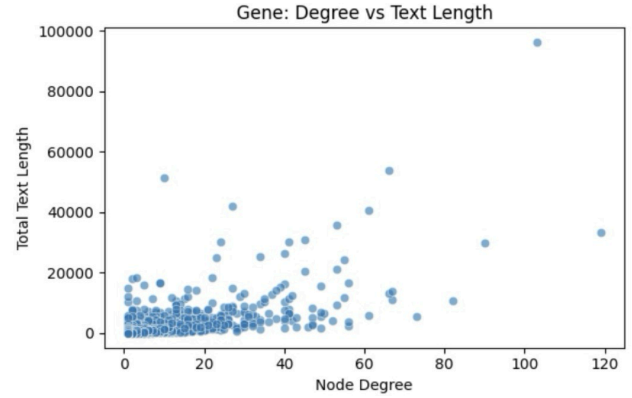


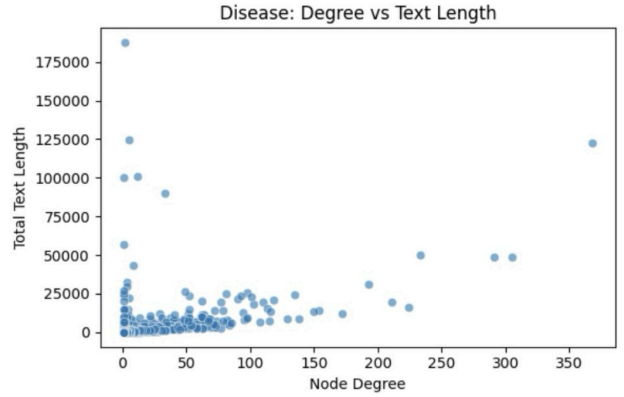
To further explore centrality, we examined the most connected nodes. This indicates that diseases such as breast cancer and colorectal cancer, and genes such as TP53 and EGFR, dominate the graph. These entities are well-studied and likely contribute to research bias and label imbalance.



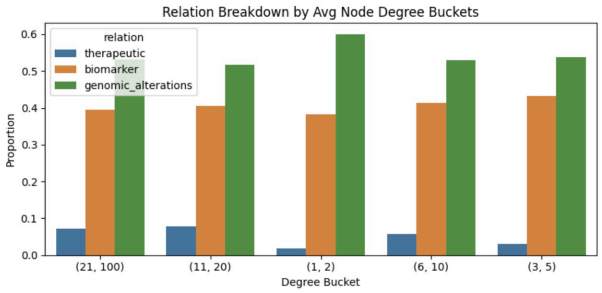


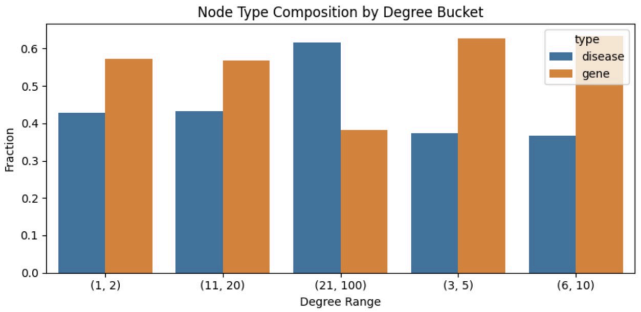
Each node is associated with a short textual description derived from biomedical literature. These descriptions were embedded using various strategies, including pretrained language models. Gene and disease descriptions confirm that most descriptions are concise, with the majority under 200 tokens, making them well-suited for transformer-based encoding without truncation.





Relation Breakdown by Degree Buckets shows genomic alterations are consistently dominant, regardless of node degree. This confirms their structural prevalence and the imbalance GNNs must address. Node Type Composition by Degree Bucket reveals that high-degree nodes are mostly diseases. This suggests that GNNs might benefit from learning asymmetric patterns between gene and disease nodes.





Methodology

We addressed two core tasks in this study: binary classification to determine whether a gene–disease relationship exists, and multi-class classification to predict the type of relationship for known connections. To support both tasks, we evaluated four types of node embeddings—random initialization, LLM-based embeddings, PCA-reduced vectors on LLM embeddings, and LLM embeddings trained on a reduced graph excluding unlabeled connections. These embeddings were tested on three architectures: GCN, GAT, and Transformer. The multi-class classification task was used as the initial benchmark for comparing these embedding-model combinations. The best-performing embedding strategy was selected for all subsequent stages.

Node Embedding Process:

Extract Text and Embed with an LLM

For all entries in the training dataset, we extracted the "text" field—each one describing a specific gene–disease association. These sentences were passed through a pre-trained biomedical language model (PubMedBERT), which transformed each sentence into a 768-dimensional vector using the contextual [CLS] token representation. Each vector was stored alongside the corresponding gene and disease IDs.

Group Embeddings by Node

Since the same gene or disease can appear in multiple associations, we grouped all the sentence embeddings that referenced the same gene or disease. This resulted in a collection of embedding vectors for each individual node (gene or disease).

Average Across Occurrences

To obtain a single embedding per node, we computed the average of all the sentence vectors associated with that node. This step ensures that the final node representation captures diverse contexts in which the entity appeared across the dataset.

Build Final Node Feature Dictionary

Finally, we combined the averaged vectors into one comprehensive dictionary containing a unique 768-dimensional feature vector for every node in the graph. This dictionary served as the input node features for the GNN, enabling it to leverage both graph structure and semantic content.

We applied the same embedding process after removing the "NA" (which means no connection) edges and also experimented with reducing the embedding dimensionality to 128 using PCA. However, as discussed earlier, the best performance was achieved when using the full 768-dimensional embeddings generated by the LLM on the complete set of connections.

Using the optimal embeddings, we proceeded to train all three architectures on the binary classification task. To account for the heavily skewed nature of the data, we constructed two datasets: one balanced through undersampling of negative edges, and one using the original (unbalanced) distribution. Each architecture was trained and evaluated on both datasets to identify the best-performing configuration. After selecting the optimal model, embedding, and dataset combination, we further investigated the impact of training duration by increasing the number of epochs. In parallel, we conducted a threshold analysis on the model's probabilistic outputs to identify the optimal decision boundary.

Final models were retrained on the combined training and validation data sets, and evaluated on a held-out test set to assess generalization. The final system was implemented as a two-stage pipeline: the binary model was first used to predict whether a connection exists, then only the false positive cases—instances incorrectly predicted as positive by the binary classifier—were then passed to the multi-class classifier to infer the specific relationship type. To deepen our understanding of model behavior, we conducted a manual review of the false positives identified by the binary model.
In addition, we aimed to benchmark our best-performing models against results reported in prior literature, mainly HOGCN (Kishan KC et al., 2020) and Crop-GPA (Gao et al., 2024) to a lesser degree, to assess whether our approach achieves competitive or superior performance in terms of predictive precision.

Results & Discussion

Embedding Strategies: Multi-Class Classification

We compared four types of node embeddings- random initialization, PCA on LLM-based (lower dimension to a vector of 128 components instead of 768), LLM-based, and LLM trained on a graph without NA connections across three architectures. The model used a pre-trained BiomedNLP language model, which has been trained on biomedical literature. The results are on the validation set:

Embedding Type	Model	Therapeutic	Biomarker	Genomic Alterations	Accuracy	F1
Random Initialization	**GCN**	0.428	0.444	0.546	0.489	0.476
	GAT	0.388	0.563	0.582	0.551	0.550
	Transformer	0.371	0.517	0.527	0.504	0.504
PCA on LLM Embeddings	**GCN**	0.334	0.458	0.552	0.491	0.475
	GAT	0.340	0.542	0.601	0.546	0.539
	Transformer	0.416	0.560	0.567	0.547	0.546
LLM Embeddings	**GCN**	0.272	0.601	0.554	0.556	0.548
	GAT	0.211	0.591	0.582	0.555	0.545
	Transformer	0.181	0.602	0.626	**0.582**	**0.563**
LLM w/o NA Connections	**GCN**	0.347	0.276	0.510	0.408	0.361
	GAT	0.367	0.516	0.575	0.527	0.519
	Transformer	0.321	0.547	0.586	0.541	0.534

LLM embeddings generally outperformed other types across models, with the Transformer architecture achieving the highest overall weighted F1-score (0.563) and accuracy. This suggests that LLM-based representations capture rich semantic relationships crucial for multi-class prediction. The random and PCA embeddings performed reasonably well but lacked the expressiveness of pretrained embeddings. The embeddings trained on graphs excluding NA connections underperformed - likely due to loss of structural information, confirming the value of preserving the graph's global context.

Binary Classification: Original vs Balanced Data

We evaluated the binary link prediction task using both the original (imbalanced) dataset and a balanced dataset created through undersampling. Each architecture was tested under both conditions.

Dataset	Model	Precision (Class 1)	Recall (Class 1)	F1 (Class 1)	Accuracy	Macro F1
Original	GCN	0.575	0.213	0.311	0.809	0.600
	GAT	0.551	0.306	0.393	0.809	0.640
	Transformer	**0.729**	0.289	0.413	**0.834**	0.659
Balanced	GCN	0.435	0.616	0.510	0.766	0.678
	GAT	0.473	0.741	0.578	0.786	0.717
	Transformer	0.499	**0.711**	**0.587**	0.802	**0.728**

Balancing the dataset improved recall, but the original (unbalanced) dataset achieved higher accuracy and the best precision for class 1 with the Transformer (0.729, 83.4% accuracy), we will explain about the importance of precision later on. Based on this, we selected the original dataset for all subsequent experiments, prioritizing overall performance and precision.

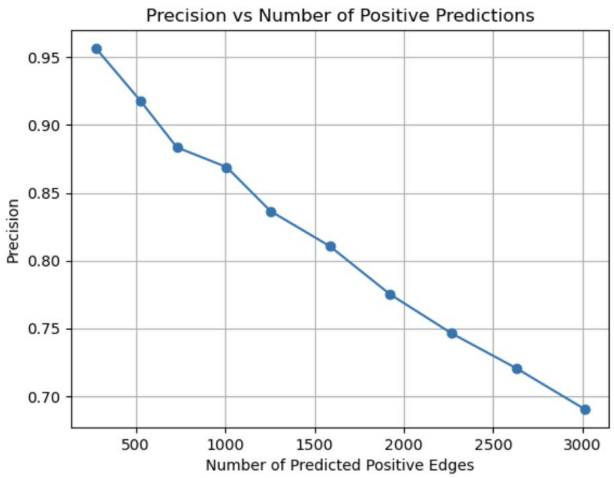
Final Architecture & Embedding Selection

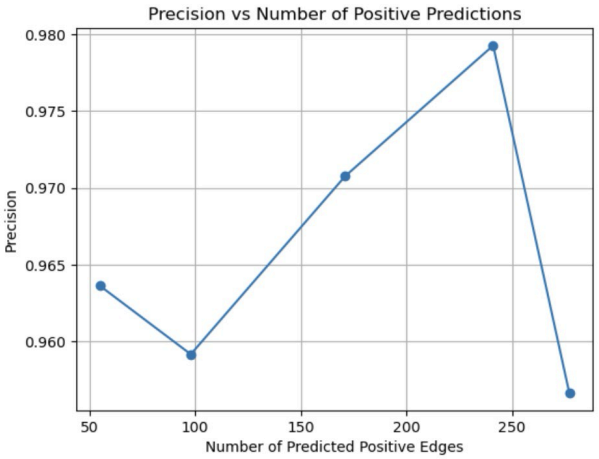
Based on validation results, the best-performing configuration across both the binary and multi-class classification tasks was the Transformer-based GNN model using LLM-derived embeddings. This setup consistently outperformed alternative architectures, including GCN, which served as our basic baseline. In the binary classification task, the Transformer model achieved a 27% improvement in precision and a 3% gain in accuracy over the GCN. To further enhance generalization, we extended the training duration to 200 epochs, which led to additional performance improvements and further solidified this model as our final choice.

Threshold Analysis - Validation Set

We prioritize high precision because we want the model to be confident when predicting a connection, even if it means missing some true positives. By focusing on the few cases it predicts as positive—especially those it misclassifies despite a high decision threshold—we can identify potential real-world associations that may be missing from the labeled data.

To refine the decision boundary in the binary classification task, we analyzed the model's probabilistic outputs under various threshold values. Our goal was to find a threshold that improves precision without losing too many predictions.



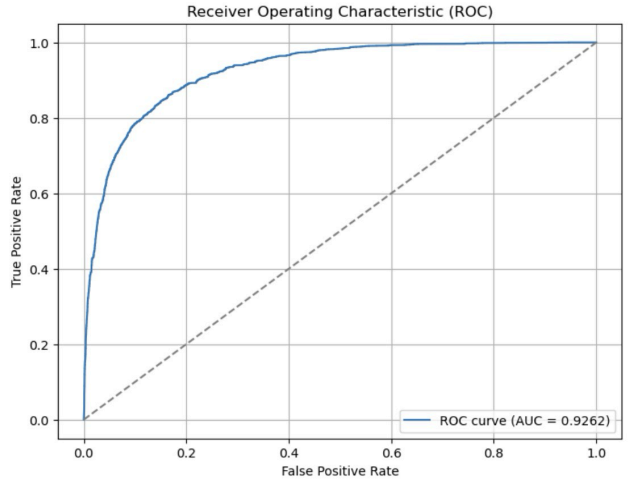


The analysis revealed that increasing the threshold led to fewer predicted positives, but with higher confidence. A threshold of **0.96** produced the **highest precision - nearly 98%**, albeit over a small subset of approximately **240 instances**. This indicates that when the model is highly confident, it is rarely wrong. However, this comes at the cost of recall, since many true positives are no longer detected at such a strict cutoff.

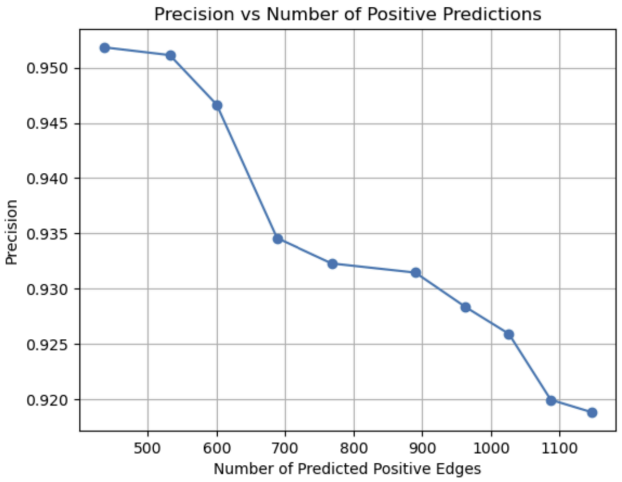
Final Evaluation on Test Set

We retrained the best models (binary and multi-class) using the combined training and validation sets and evaluated them on a held-out test set. High performance on the binary model, and reasonable performance on the multi-class model, which is a tougher task. Below are graphs archived on the test dataset:

Task	Model	Embedding	Accuracy	F1 Score	AUC
Binary	Transformer	LLM	0.89	0.83	0.926
Multi-class	Transformer	LLM	0.68	0.64	



AUC: 0.9262



False Positive Analysis

To better understand the model's behavior and uncover potentially novel associations, we conducted a manual review of the false positives generated by the binary classification model.
Based on validation results, a threshold of approximately 0.96 was found to be optimal. After further evaluation on the test set, we chose a stricter threshold of 0.98. At this level, the model predicted around 451 positive connections, with only 27 false positives—demonstrating a high level of precision (**0.94**).
Out of those **27 false positives**, many predictions were found to be **semantically plausible** and supported by contextual information in the node features. For example, several gene–disease pairs involved genes that appeared frequently in connection with similar diseases elsewhere in the graph, or diseases that shared ontology terms with known associations. Again, all of these false positives had **high predicted probabilities** (above 0.98), indicating strong model confidence.
One specific example is the predicted link between **DNAJC19** and **anemia**. While this association was not labeled in the dataset, mutations in DNAJC19 are known to cause **MLASA (mitochondrial myopathy, lactic acidosis, and sideroblastic anemia)**, confirming the biological plausibility of the model's output.
This outcome underscores the utility of false positive review not merely as an error analysis step, but as a **discovery mechanism**—where confident predictions outside the gold standard may point to novel or under-documented gene–disease links worth further investigation. The second model predicted this type of connection to be a "biomarker" which is wrong in this case.

Comparative Performance Against Prior GNN-Based Benchmarks

To contextualize our results, we compared them with recent GNN-based approaches addressing similar tasks. Gao et al. (2024) achieved better results using pre-trained LLM for the initial embeddings, and so did we. In the gene-disease link prediction task, our model achieved a AUC score of **0.926**, reflecting similar capability to Kishan KC (2020) which achieved 0.936 using state of the art model (HOGCN) on a similar dataset (GDI-Gene Disease Interaction) to ours, correctly identifying if there is a binary biological relationship. These results suggest that our approach yields high predictive performance in comparison to established GNN baselines.

Conclusion and Innovation

This study set out to evaluate whether Graph Neural Networks (GNNs), particularly models like GCN, GAT, and Transformer, can accurately predict novel gene–disease associations using biomedical embeddings. Our findings provide a clear affirmative answer. Through rigorous experimentation with different embeddings and architectures, we demonstrated that GNN-based models—especially the Transformer with LLM embeddings—not only classify known relationships effectively but also generalize to previously unseen data.

One of the key innovations in our approach lies in the integration of domain-specific LLM-derived sentence embeddings, aggregated at the node level across multiple textual contexts. This enabled us to infuse each node with rich semantic meaning beyond traditional graph structure. We also introduced a two-stage prediction pipeline that first detects the existence of a link and then classifies its biological type—allowing for more interpretable predictions and targeted discovery.

Most notably, our manual review of high-confidence false positives revealed biologically plausible associations not present in the labeled dataset. For example, the model predicted a connection between DNAJC19 and anemia. While this was marked as a false positive by the gold-standard data, it is well-established in biomedical literature that DNAJC19 mutations cause MLASA, a rare mitochondrial disorder that includes sideroblastic anemia as a clinical feature. This supports the model's capacity to uncover meaningful, novel associations—one of the core innovations demonstrated in this study.

In comparison to prior GNN-based studies, our models performed on par with existing state-of-the-art benchmarks, reinforcing the effectiveness of our architectural choices and demonstrating the value of combining deep learning with domain-aware representations in biomedical discovery.

Key Contributions

- Proved GNNs can predict novel, biologically valid links (e.g., DNAJC19–anemia).
- Introduced a two-stage prediction pipeline (link + relation).
- Highlighted the strength of LLM-based embeddings.
- Showed unbalanced data improves real-world precision while balanced improves F1.
- Offered threshold tuning for precision vs. recall trade-offs.

Future Work

- Extend to larger biomedical graphs (e.g., including proteins, drugs)
- Explore GNN+LLM combinations (e.g., RAG for literature validation)
- Fine-tune BioBERT on domain-specific gene–disease corpora, or get more text/information about them- for better initial node embeddings.
- Fine tune the transformer model even more- hyperparameter grid, different architectures etc.
- Collaborate with medical researchers and scientists to investigate high-confidence false positives that are not yet documented in the scientific literature. This could potentially lead to the discovery of entirely new gene–disease associations, opening the door to impactful biomedical insights and further research.

References

1. Kishan KC, Rui Li, Feng Cui, and Anne R. Haake (2020). HOGCN: Predicting Biomedical Interactions with Higher-Order Graph Convolutional Networks. *arXiv preprint* arXiv:2010.08516.
2. Gao, Y., Zhou, Q., Luo, J., Xia, C., Zhang, Y., & Yue, Z. (2024). Crop-GPA: an integrated platform of crop gene-phenotype associations. NPJ Systems Biology and Applications