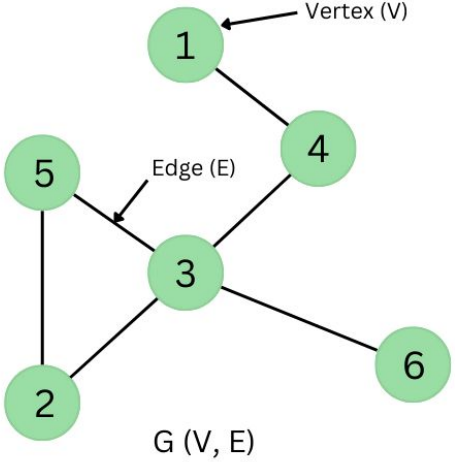
Predicting Gene-Disease Associations with GNNs

Shelly Levy, Adam Bublil, Tom Saacks and Roy Efroni

GNN Recap – Core Concept

Graph Components:

- **Nodes**: Entities (e.g., genes, diseases)
  - Nodes can be typed (heterogeneous): e.g., Gene vs. Disease
  - Each node includes attributes: textual descriptions, ontology tags, etc.
- **Edges**: Relationships (e.g., gene–disease associations)
  - Undirected / Directed
  - Weighted (e.g., strength of association)
  - Attributed (contain metadata like relation type)

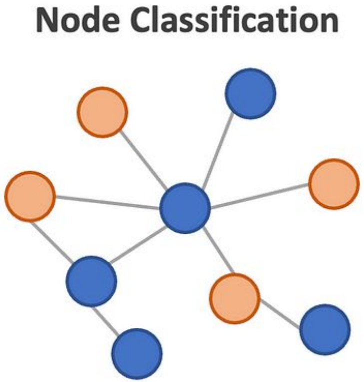


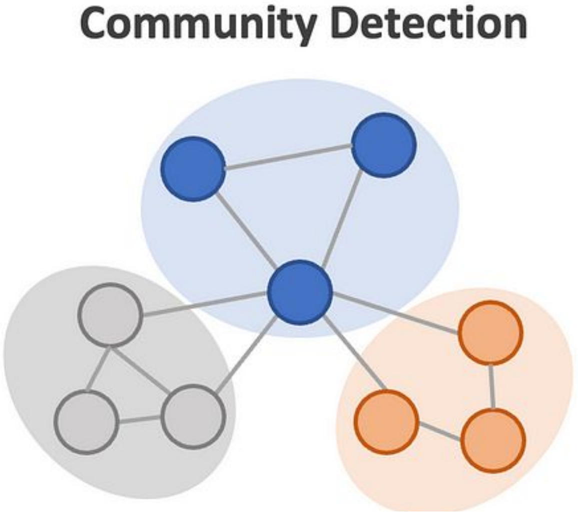
Why GNNs for Biology?

- GNNs learn from both structure & node features
- Capture complex, non-Euclidean dependencies

GNN Recap – Common Node-Level Tasks

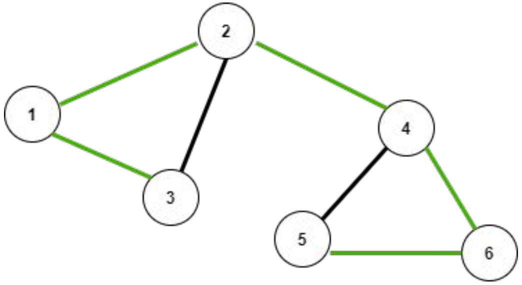
- Node Classification
  - Predict category/label for each node
  - Example: Is this user a bot?
- Node Regression
  - Predict a continuous value per node
  - Example: Estimate air quality at each sensor node
- Node Clustering
  - Group nodes based on structure/features
  - Example: Detect social communities



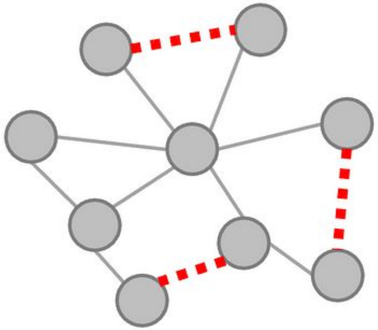


GNN Recap – Common Edge-Level Tasks

- Edge Classification
  - Classify relationships
  - Example: Is this relationship strong or weak?
- Link Prediction
  - Predict if an edge should exist
  - Example: Recommend a new friend



Link Prediction



Our Research Question

Can GNNs predict novel gene–disease associations from known biological interactions?

Motivation

- **Why Predict Gene–Disease Associations?**
  - Manual discovery is slow, expert-driven, and incomplete
  - Novel associations critical for rare diseases and drug repurposing
  - Biomedical data is graph-structured: genes, diseases, interactions
  - GNNs excel at relational learning on graphs
  - Goal: Automate and scale discovery using graph-based learning

Dataset overview – TBGA structure

Source:

- Text-Based Gene–Disease Association (TBGA) dataset
- Derived from curated biomedical literature (DisGeNET-like schema)

Graph Composition:

- **Nodes**:
  - **Genes**: 9,569 unique entities
  - **Diseases**: 7,499 unique entities
- **Edges** (Gene–Disease pairs):
  - 49,214 total associations

Relation Types (Labeled Edges):

- Therapeutic
- Biomarker
- Genomic Alteration

Dataset overview – TBGA structure

Node Features:

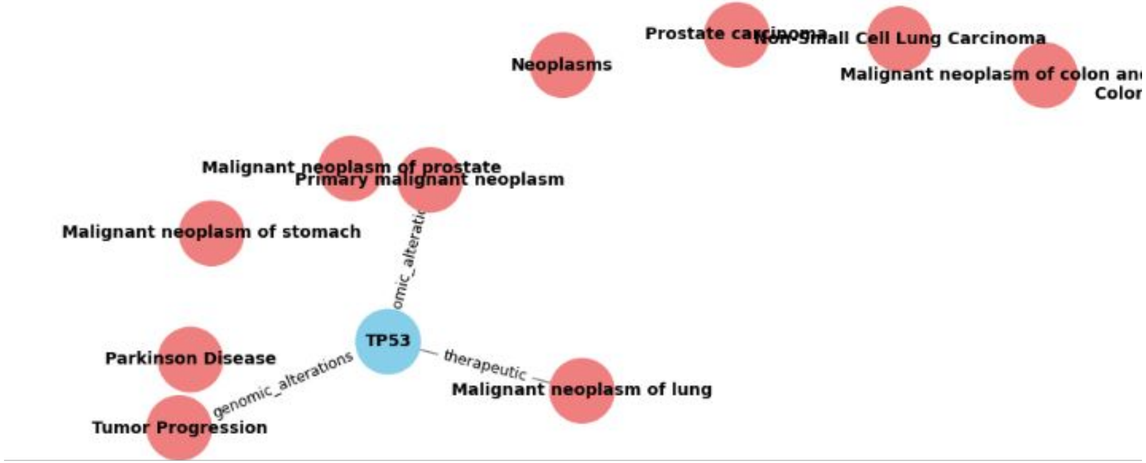
- Short **textual descriptions** derived from biomedical abstracts
- Used as input to various **embedding strategies**

Graph Characteristics:

- **Undirected & Attributed**
- **Heterogeneous**, **sparse**, **imbalanced**
- **Hub nodes** dominate (e.g., TP53, breast cancer) – biological realism

Edge Type Distribution:

- **Biomarker –** Indicates the gene is used to diagnose/track disease
- **Therapeutic –** Gene targeted for treatment/intervention
- **Genomic Alteration –** Mutation or expression change linked to disease


Prostate carcinoma
Non-Small Cell Lung Carcinoma
Neoplasms
Malignant neoplasm of colon and
Color
Malignant neoplasm of prostate
Primary malignant neoplasm
Malignant neoplasm of stomach
omic_alteratic
TP53
Parkinson Disease
genomic_alterations
therapeutic
Malignant neoplasm of lung
Tumor Progression

Problem Definition & Objectives

- Can we predict new gene–disease associations using GNNs?
- Can we also classify the biological nature of those relationships?
- How good will our model be?

Problem Definition & Objectives

- Can we predict new gene–disease associations using GNNs?
  Link Prediction

- Can we also classify the biological nature of those relationships?
  Node Classification

- Requires models that leverage both graph topology and semantic node features

Binary Link Prediction

- **Task:** Predict if a gene–disease edge **should exist**
- **Input:** Graph with known positive (labeled) and NA (unlabeled) edges
- **Output:** Binary label – *Associated (1)* or *Not Associated (0)*
- Addresses the **discovery** challenge (find novel pairs)
- Evaluated using **AUC, Accuracy, Precision, Recall, F1**, and **Confidence thresholding**

Multi-Class Relation Classification

- **Task:** Given an edge, classify its **biological role**
- **Classes:**
  - *Therapeutic*
  - *Biomarker*
  - *Genomic Alteration*
- **Input:** Only labeled edges used
- Supports **interpretability** and **functional understanding**
- Evaluated via **Accuracy,** and **class-wise Precision, Recall, F1**

Two-Stage Prediction Pipeline

- **Stage 1:**
  - Binary GNN predicts candidate gene–disease links
- **Stage 2:**
  - Multi-class GNN assigns biological relation types
- **Benefits:**
  - Allows high-precision link discovery
  - Adds fine-grained interpretability via relation classification
- Modular and **extendable to new biological contexts**

Literature Review – Key Insights

- **GCNs demonstrate high accuracy** in predicting gene–trait associations in crops, outperforming classical models (Crop-GPA, Gao et al., 2024).

- **Text + graph fusion (GPA-GCN)** enables discovery of latent biological links using semantic similarity and network proximity.

- **HOGCN models outperform GCNs** by leveraging high-order connections, capturing indirect relationships in biomedical graphs.

- **Deeper GNN architectures (e.g., MDA-HOGCN/ Transformes)** improve AUC and clustering, especially in complex biological networks.

- **Evidence supports GNN applicability** in gene–disease prediction, motivating our use of GCN and GAT on biomedical data.

What Success Looks Like

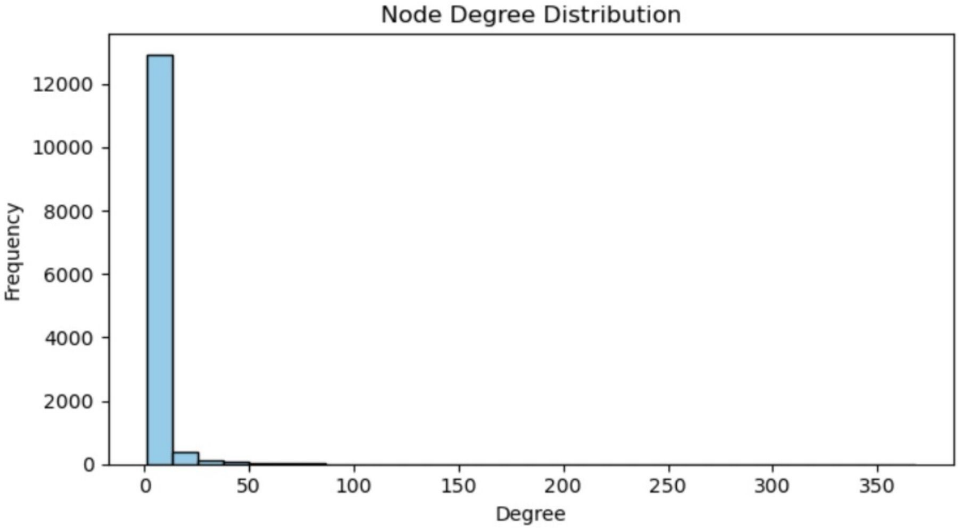
- **Similar to Prior Benchmarks (from literature)**
- **Generalization to Novel Associations**

  Correctly predict biologically plausible links not included in the labeled dataset
- **Precision–Recall Tradeoff Optimization**

  Enable threshold tuning to adjust the balance between recall and precision depending on downstream needs.

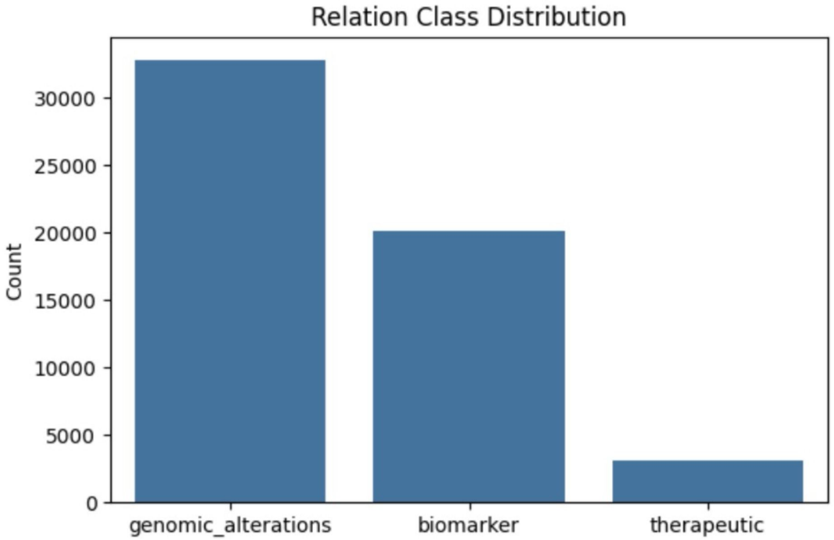
EDA - Degree Distribution

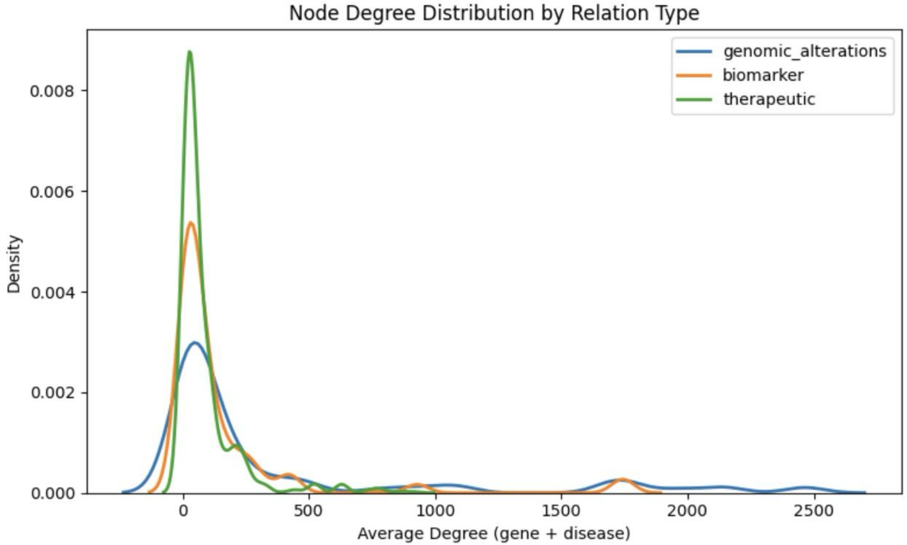
- Long-tailed distribution
- Most nodes have few connections
- Small number of high-degree hubs
- Typical in biological networks (power law)



EDA - Relation Type Distribution by Node Degree

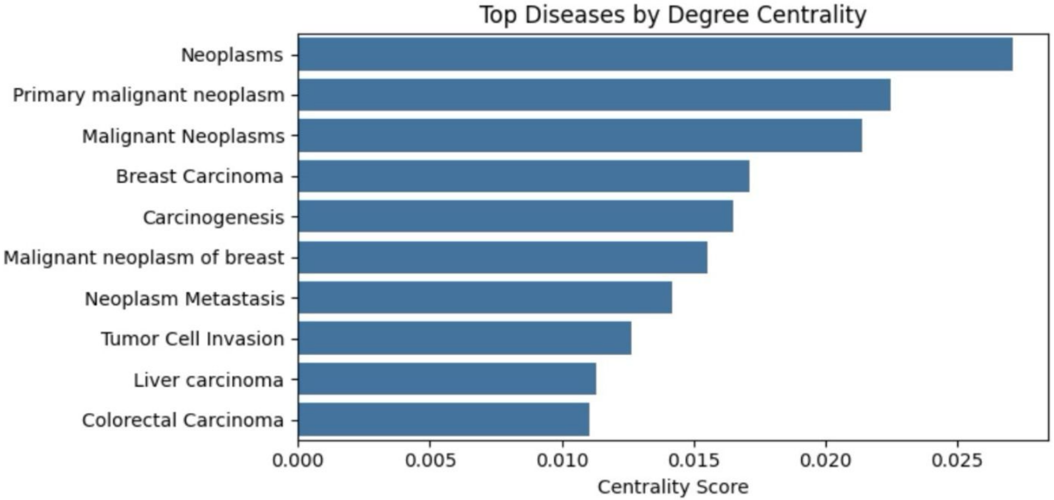
- Genomic Alterations dominate across all degree buckets
- Biomarker and Therapeutic relations are rarer
- Higher-degree nodes tend to concentrate Genomic Alterations





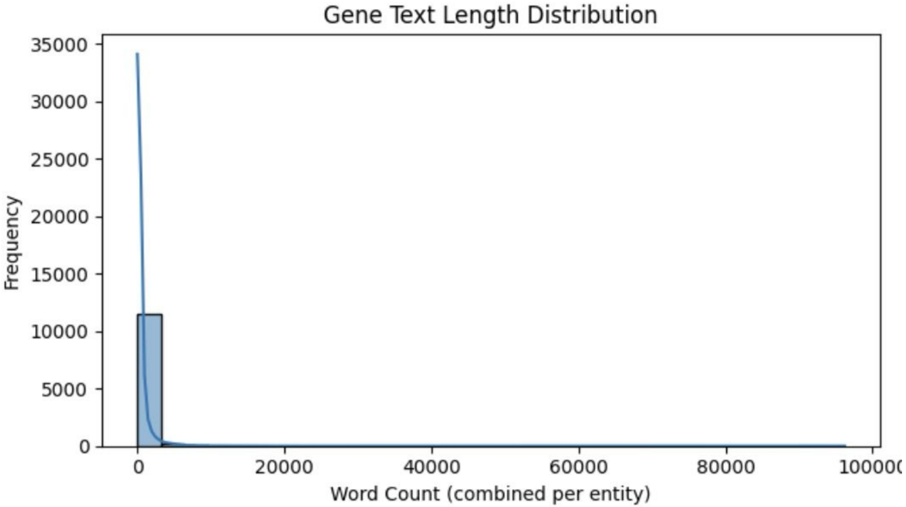
EDA - Hub Nodes – Diseases

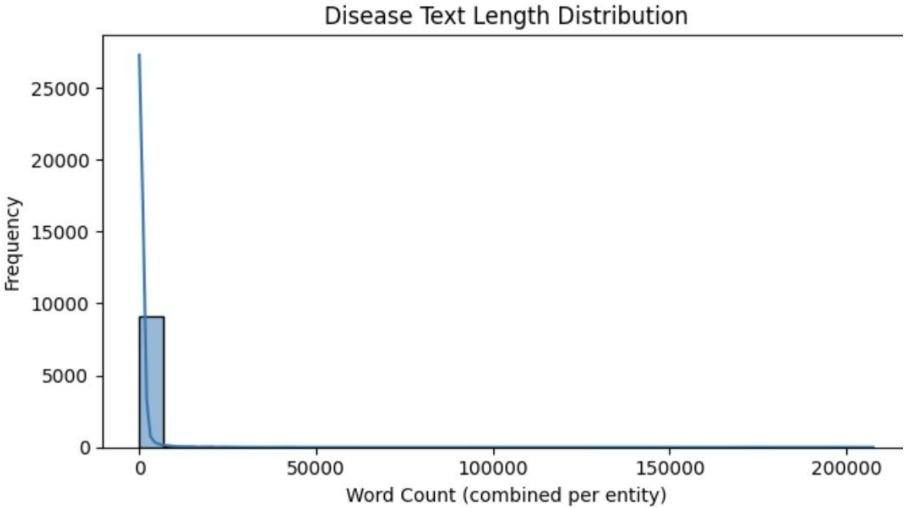
- Top disease hubs (by degree):
  - Breast cancer
  - Colorectal cancer
  - Lung neoplasm
- Likely overrepresented due to **research focus**



EDA - Node Text Description Lengths

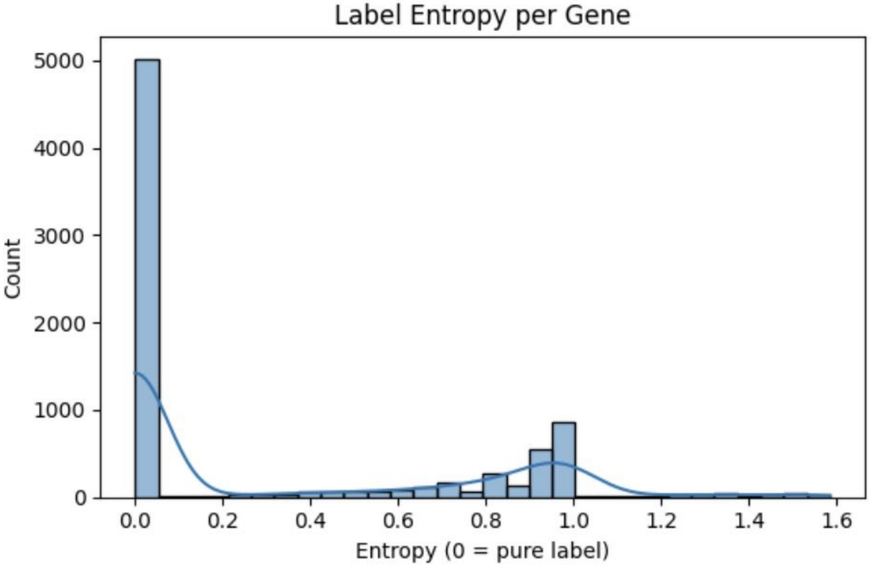
- Each node has a biomedical **text summary**
- Token lengths vary: mostly short (sub-200 tokens)
- Suitable for transformer-based encoders





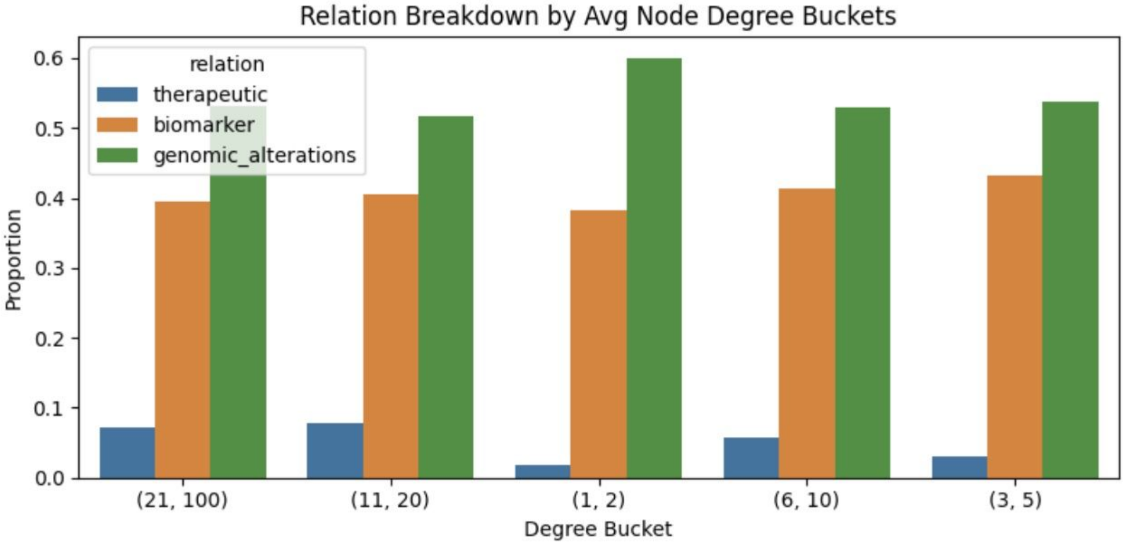
EDA - Label Entropy Across Nodes

- Most genes have low entropy → linked to a single relation type.
- Some genes show high entropy → involved in diverse roles.
- Supports need for expressive models to capture multi-relational behavior.



EDA - Degree Buckets Reveal Structure-Label Imbalance

- Across all degree ranges, **Genomic Alterations** dominate
- **Therapeutic links** are consistently underrepresented
- Suggests **class imbalance** is correlated with node connectivity



Methodology - Two-Stage Framework Overview

- **Stage 1:** Binary classification → Predict whether gene–disease edge exists
- **Stage 2:** Multi-class classification → Predict type of biological relation
- Shared graph + textual embeddings as input for both stages
- Trained and evaluated on TBGA biomedical graph

Methodology - Task Definition & Setup

- **Two main tasks:**
  Binary classification: Is there a gene–disease link?
  Multi-class classification: What is the type of link? (Therapeutic / Biomarker / Genomic alteration)

- **Two-stage pipeline:**
  Stage 1: Binary model → candidate pairs
  Stage 2: Multi-class model → relation labeling

- **Why this setup?**
  Mimics real-world use: first find unknown links, then classify them

Methodology - Embeddings & Model Architectures

- **Node representation strategies:**
  Random initialization
  PCA-reduced embeddings
  LLM-based (pretrained on biomedical text)
  LLM-based excluding NA edges

- **GNN architectures tested:**
  GCN, GAT and Transformer

- **Purpose:**
  Identify which embedding + model combo yields best performance for each task

Methodology - Dataset Versions

- **Skewed class distribution → tested two setups:**
  Original (imbalanced)
  Balanced (undersampled negatives)

Methodology - Hyperparameter Tuning

- Epoch tuning
- Threshold analysis (binary classifier)

Comparison to Prior Benchmarks

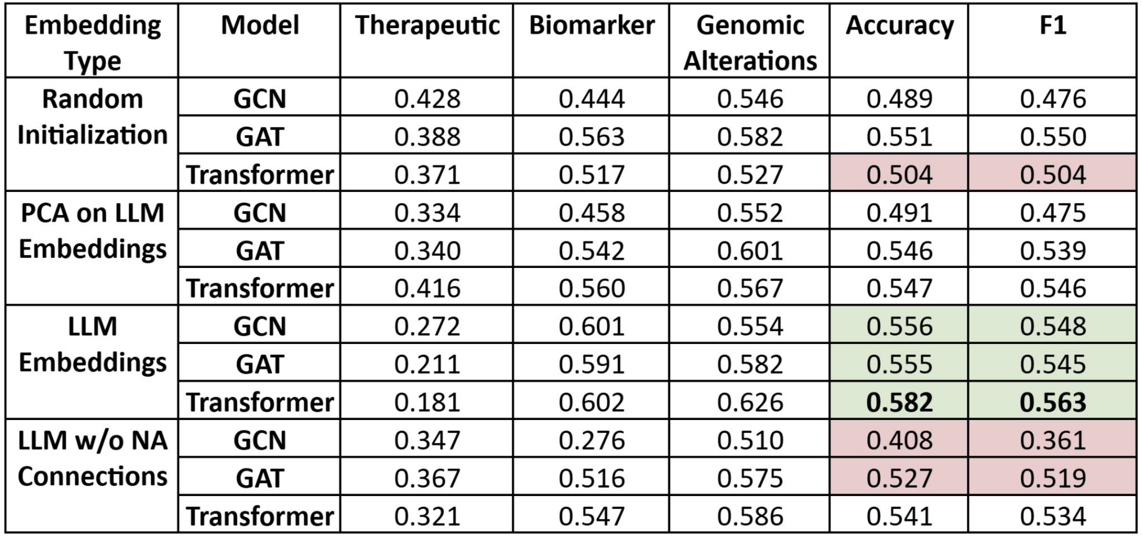
- Benchmarked against HOGCN (Kishan KC et al., 2020) and Crop-GPA (Gao et al., 2024) to assess precision and performance relative to existing GNN-based methods

False Positive Analysis

- Manually reviewed top high-confidence false positives

LLM Embeddings Boost Relation Classification

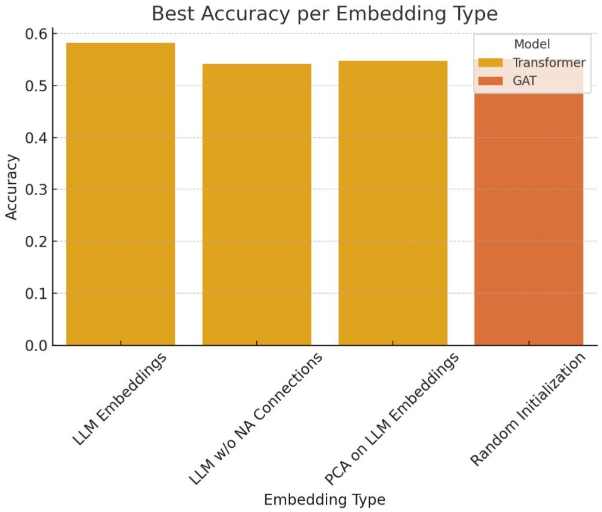
- We tested 4 embedding strategies across 3 GNN models
- Best performing setup: Transformer + LLM embeddings, with an F1 score of 0.563
- Simpler embeddings (random, PCA) failed to capture semantic signals
- LLM embeddings trained on biomedical text provided richer node representations

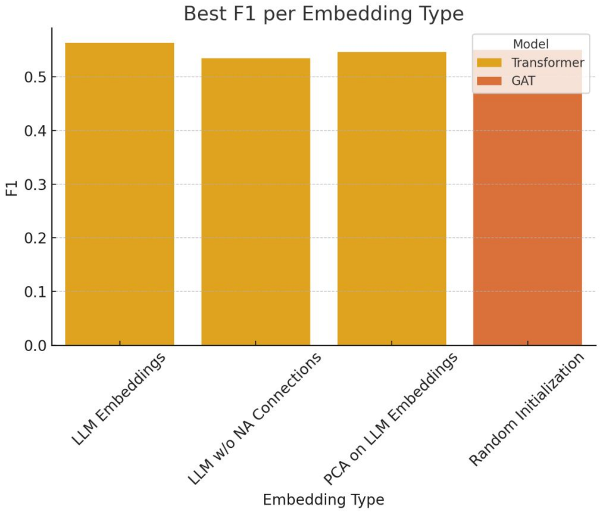


Embedding Type	Model	Therapeutic	Biomarker	Genomic Alterations	Accuracy	F1
Random Initialization	GCN	0.428	0.444	0.546	0.489	0.476
	GAT	0.388	0.563	0.582	0.551	0.550
	Transformer	0.371	0.517	0.527	0.504	0.504
PCA on LLM Embeddings	GCN	0.334	0.458	0.552	0.491	0.475
	GAT	0.340	0.542	0.601	0.546	0.539
	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 Boost Relation Classification

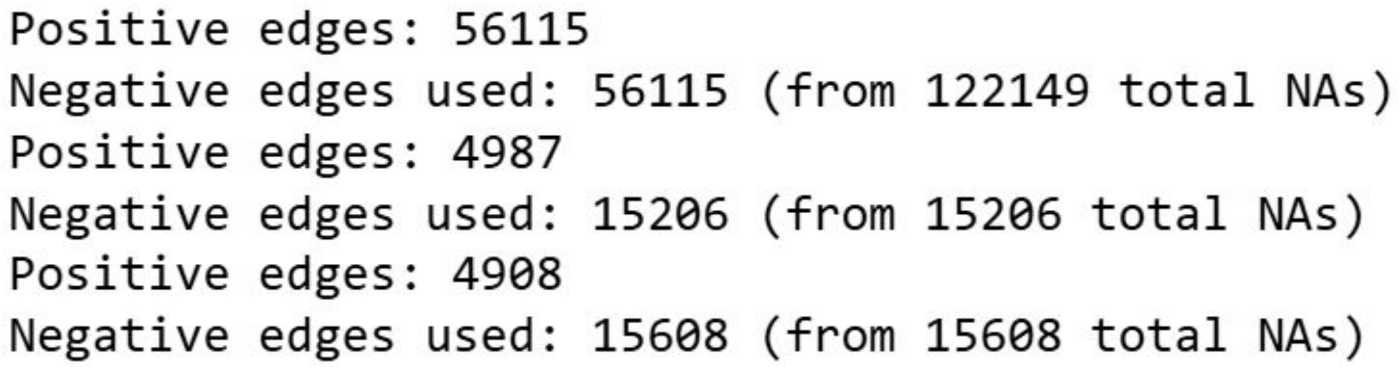
- We tested 4 embedding strategies across 3 GNN models
- Best performing setup: Transformer + LLM embeddings, with an F1 score of 0.563
- Simpler embeddings (random, PCA) failed to capture semantic signals
- LLM embeddings trained on biomedical text provided richer node representations


Best Accuracy per Embedding Type


Best F1 per Embedding Type

Balancing Improves Recall, But Precision Wins

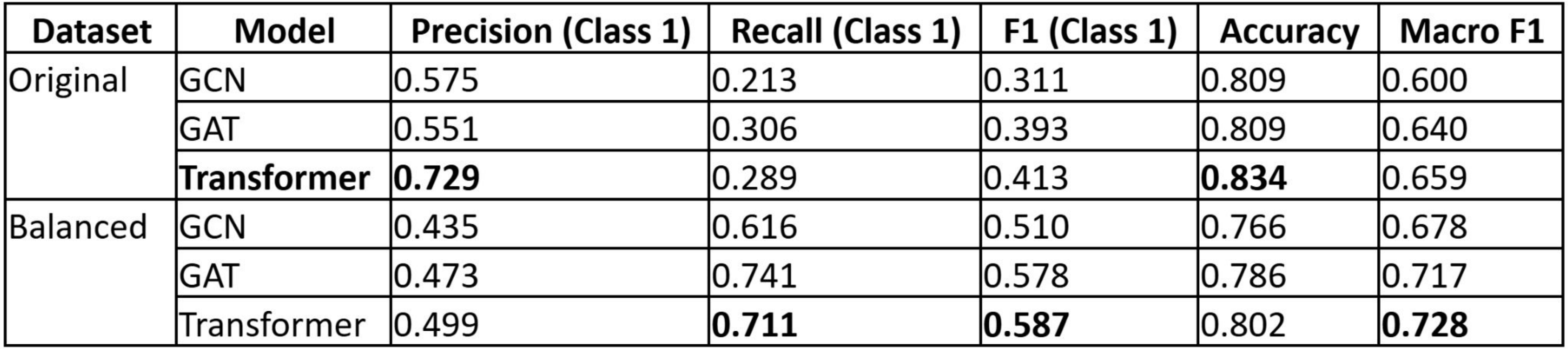
- We trained each model on two dataset versions: the original skewed set and a balanced set



```
Positive edges: 56115
Negative edges used: 56115 (from 122149 total NAs)
Positive edges: 4987
Negative edges used: 15206 (from 15206 total NAs)
Positive edges: 4908
Negative edges used: 15608 (from 15608 total NAs)
```

Balancing Improves Recall, But Precision Wins

- We trained each model on two dataset versions: the original skewed set and a balanced set
- Balancing helped with recall, but reduced precision significantly
- On the original data, the Transformer model achieved 0.729 precision
- For discovery tasks, fewer but reliable predictions are preferred
- Binary Classification results:



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**

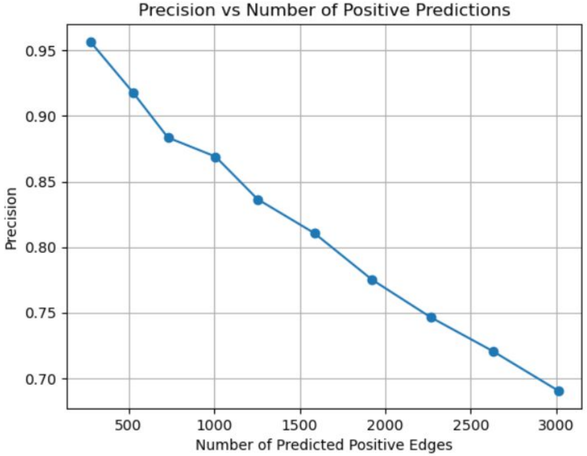
In Biomedical Discovery, Confidence › Coverage

- High precision minimizes false discoveries, which is critical in biology
- Each wrong prediction can mislead further lab work or research
- It’s better to predict fewer associations, but be more confident in each
- This guided our choice of evaluation metrics and threshold tuning

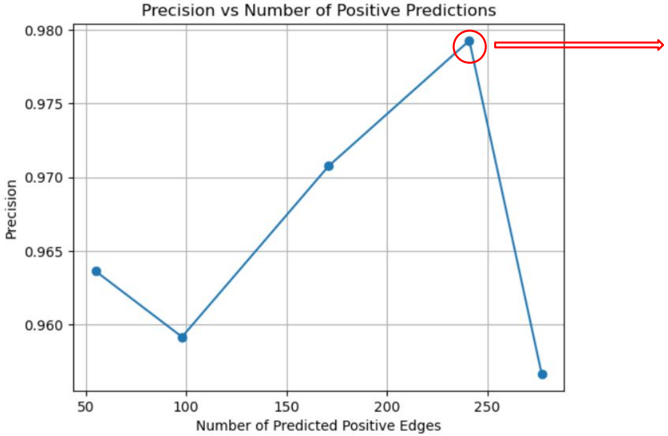
Optimizing the Confidence Threshold

- We tuned the probability threshold of the binary classifier (Val 0.96, Test 0.98)
- Higher thresholds reduce the number of predicted positives (recall), but improve precision
- At threshold 0.96, the model made 240 predictions with a 98% precision
- This tuning is essential for controlling model behavior

Threshold: 0.5-0.95 (0.05 jumps)



Threshold: 0.95-0.99 (0.01 jumps)

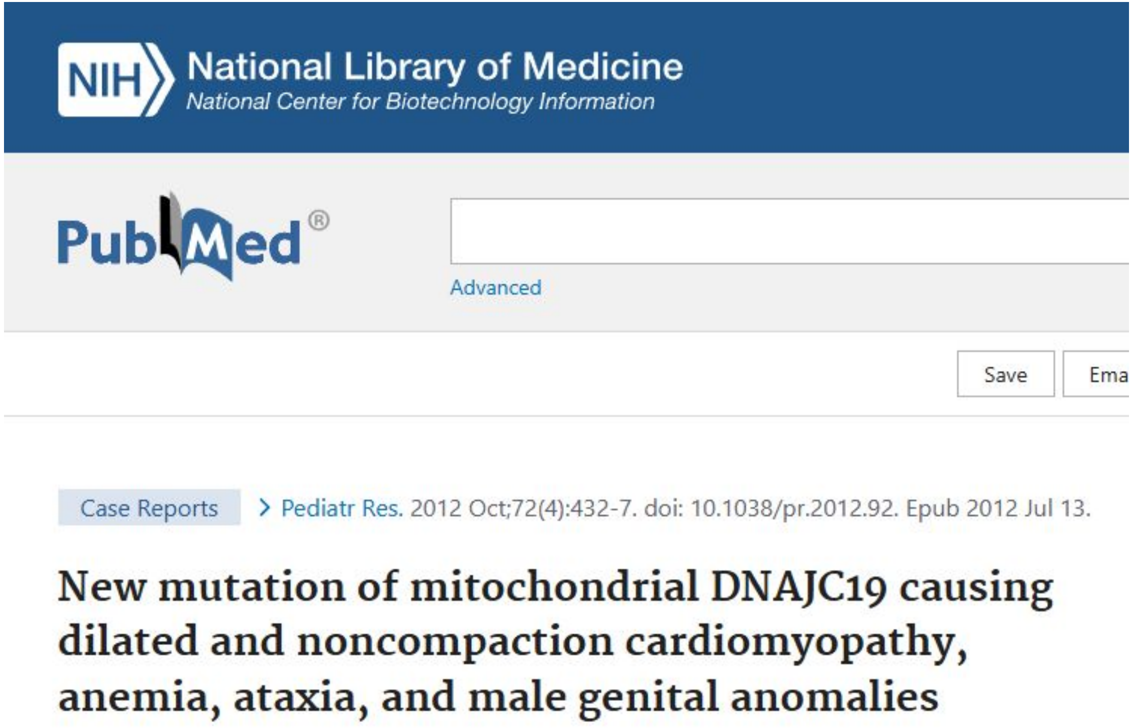


Sweet Spot

A False Positive?

A False False Positive!

A False False Positive!



NIH
National Library of Medicine
National Center for Biotechnology Information
PubMed®
Advanced
Save
Ema
Case Reports

Pediatr Res. 2012 Oct;72(4):432-7. doi: 10.1038/pr.2012.92. Epub 2012 Jul 13.

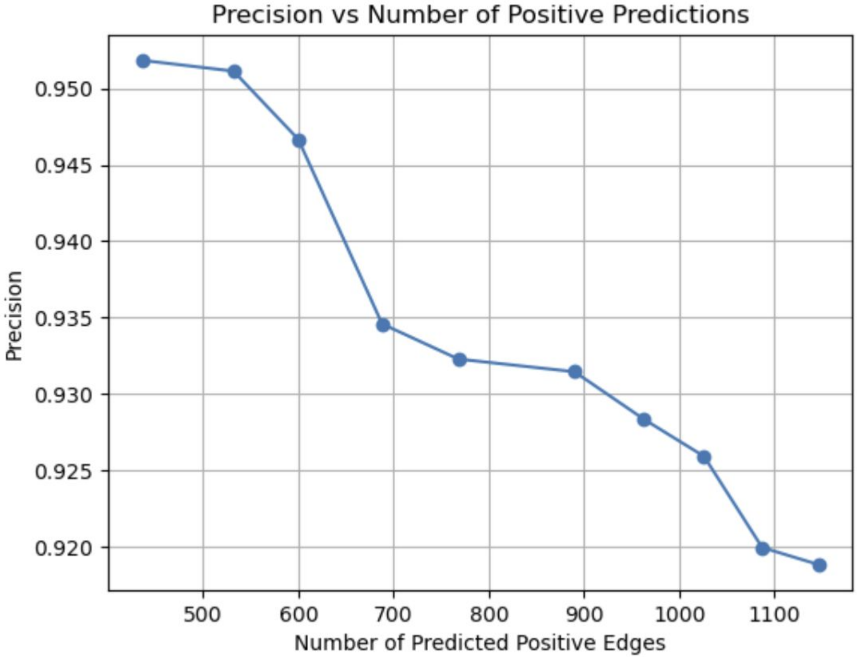
New mutation of mitochondrial DNAJC19 causing dilated and noncompaction cardiomyopathy, anemia, ataxia, and male genital anomalies

A False Positive? Actually, a Hidden Truth

- One prediction was a link between the gene DNAJC19 and anemia, **which was not labeled in the dataset**
- However, literature confirms this gene is associated with MLASA, which includes sideroblastic anemia
- The model labeled it as “biomarker” slightly off

High Confidence Errors Reveal New Biology

- On the test dataset At threshold 0.98, we reviewed 451 predictions



High Confidence Errors Reveal New Biology

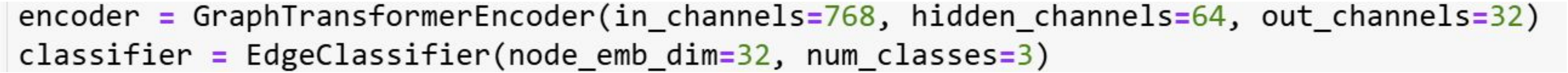
- Only 27 were labeled as false positives
- The few false positives, would have a higher chance to be novel connections
- Many of those were plausible: shared disease pathways, ontology links, or semantic similarities
- Errors often resulted from dataset gaps, not model flaws

Second Stage Helps Interpret Model Outputs

- We ran a multi class classifier on the false positives from stage 1
- Even when the type wasn’t always accurate, it provided useful context
- Examples include “biomarker” predictions where the gene is mentioned in literature as Genomic Alteration.

Final Setup: Optimized for Biomedical Use

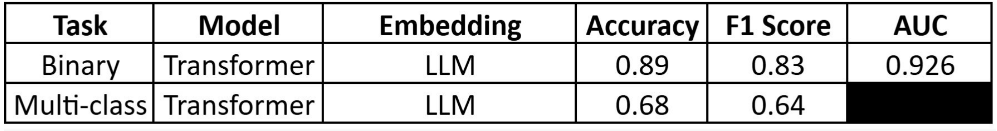
- GNN architecture: Transformer
- Node features: LLM based embeddings
- Training: 200 epochs on the original dataset
- Threshold: 0.96/0.98 for high confidence outputs
- Inference pipeline: binary prediction → relation classification



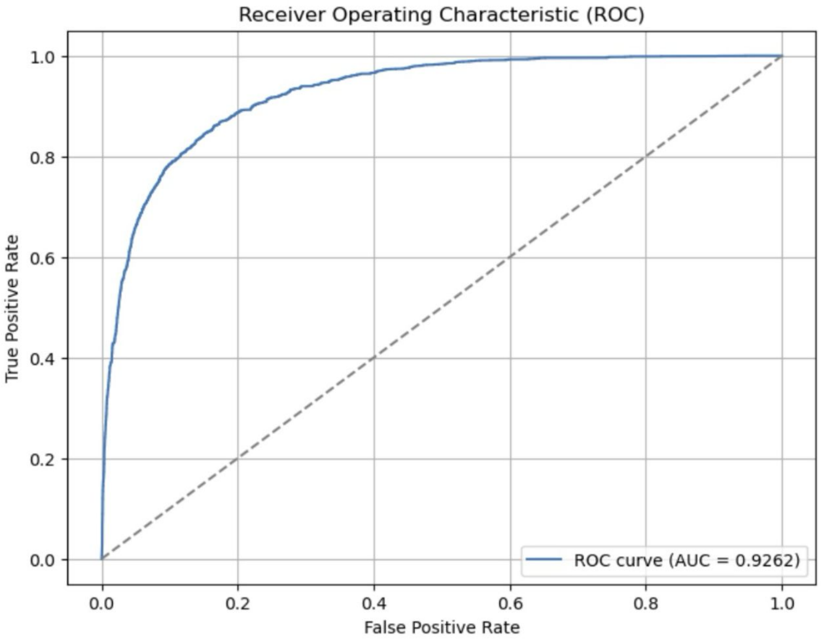
```
encoder = GraphTransformerEncoder(in_channels=768, hidden_channels=64, out_channels=32)
classifier = EdgeClassifier(node_emb_dim=32, num_classes=3)
```

Results

After training on train+validation sets and inferring on test dataset:

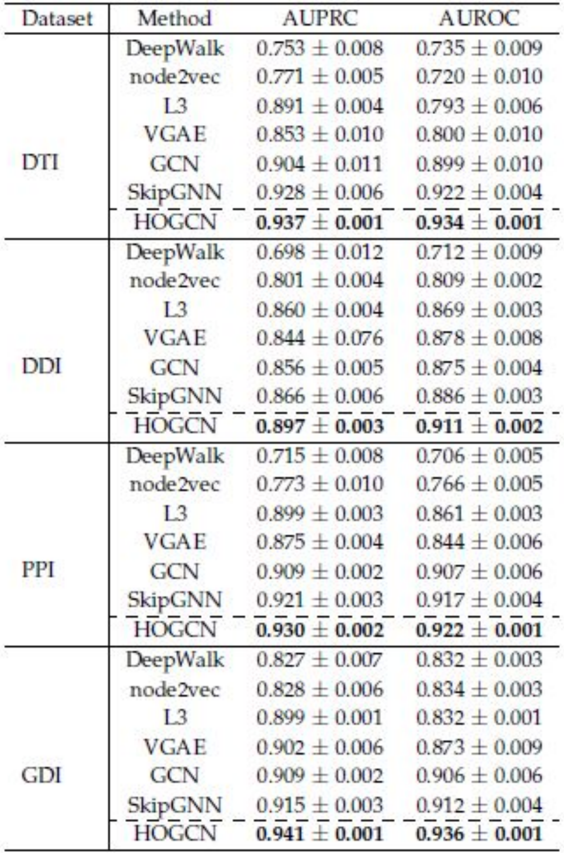


Task	Model	Embedding	Accuracy	F1 Score	AUC
Binary	Transformer	LLM	0.89	0.83	0.926
Multi-class	Transformer	LLM	0.68	0.64	[redacted]



Comparable to State of the Art

- Our model achieved AUC 0.926 on the binary-class task
- HOGCN (state of the art) reported AUC of 0.936 on similar task and dataset
- Crop-GPA performed similarly (or even worse) with heavier model complexity
- Our simpler architecture still matches these strong baselines



Dataset	Method	AUPRC	AUROC
DTI	DeepWalk	0.753 ± 0.008	0.735 ± 0.009
	node2vec	0.771 ± 0.005	0.720 ± 0.010
	L3	0.891 ± 0.004	0.793 ± 0.006
	VGAE	0.853 ± 0.010	0.800 ± 0.010
	GCN	0.904 ± 0.011	0.899 ± 0.010
	SkipGNN	0.928 ± 0.006	0.922 ± 0.004
	HOGCN	**0.937 ± 0.001**	**0.934 ± 0.001**
DDI	DeepWalk	0.698 ± 0.012	0.712 ± 0.009
	node2vec	0.801 ± 0.004	0.809 ± 0.002
	L3	0.860 ± 0.004	0.869 ± 0.003
	VGAE	0.844 ± 0.076	0.878 ± 0.008
	GCN	0.856 ± 0.005	0.875 ± 0.004
	SkipGNN	0.866 ± 0.006	0.886 ± 0.003
	HOGCN	**0.897 ± 0.003**	**0.911 ± 0.002**
PPI	DeepWalk	0.715 ± 0.008	0.706 ± 0.005
	node2vec	0.773 ± 0.010	0.766 ± 0.005
	L3	0.899 ± 0.003	0.861 ± 0.003
	VGAE	0.875 ± 0.004	0.844 ± 0.006
	GCN	0.909 ± 0.002	0.907 ± 0.006
	SkipGNN	0.921 ± 0.003	0.917 ± 0.004
	HOGCN	**0.930 ± 0.002**	**0.922 ± 0.001**
GDI	DeepWalk	0.827 ± 0.007	0.832 ± 0.003
	node2vec	0.828 ± 0.006	0.834 ± 0.003
	L3	0.899 ± 0.001	0.832 ± 0.001
	VGAE	0.902 ± 0.006	0.873 ± 0.009
	GCN	0.909 ± 0.002	0.906 ± 0.006
	SkipGNN	0.915 ± 0.003	0.912 ± 0.004
	HOGCN	**0.941 ± 0.001**	**0.936 ± 0.001**

What We’ve Achieved

- Built a two-stage GNN pipeline: discovery + interpretation
- Achieved high precision and strong generalization
- Leveraged biomedical LLM embeddings to improve prediction
- Outperformed or matched benchmark models
- Identified plausible new gene–disease links

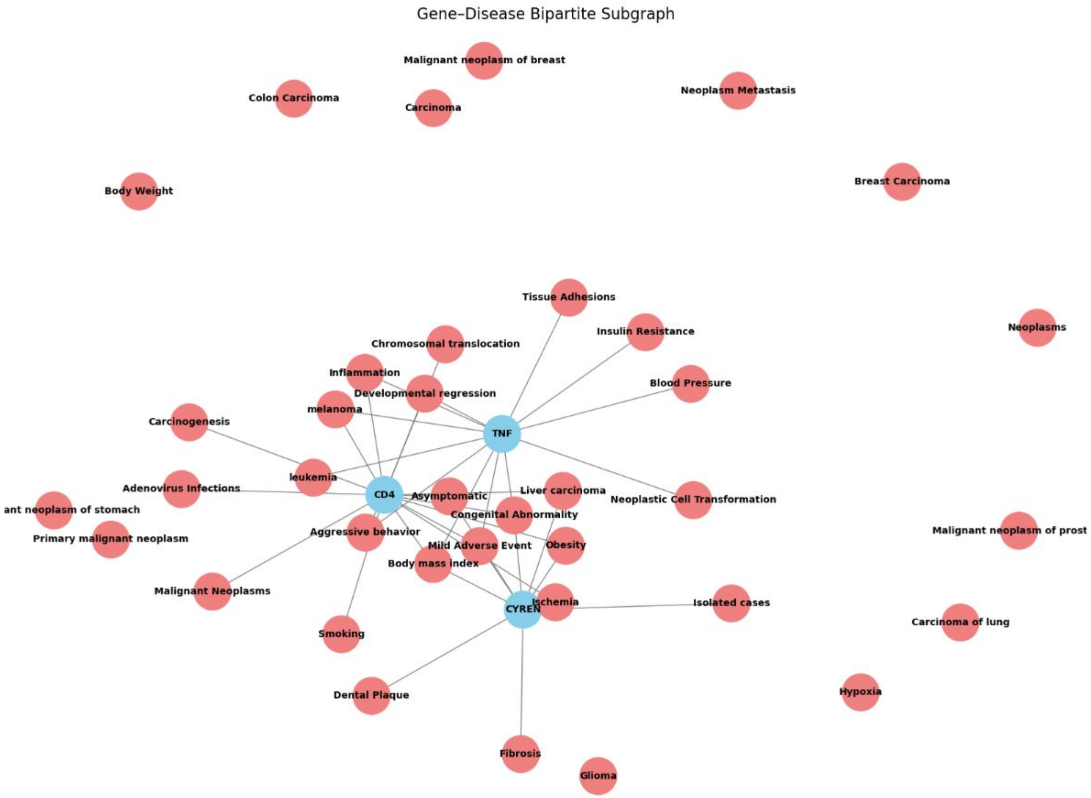
What Comes Next

- Expand the graph to include proteins, drugs, pathways
- Add retrieval-based models (GNN+LLM) to validate predictions
- Fine-tune BioBERT on domain-specific gene–disease corpora
- Get more features/data in order to create better embeddings
- Explore architecture tuning: attention heads, layer depth, and pooling strategies

Under the Hood: Technical Overview

- Node embeddings: **Pretrained BiomedNLP** language model, optionally reduced with PCA
- Training: Models trained for **200 epochs**, across both **balanced and unbalanced** datasets
- Classification tasks:
  - **Binary classification** to identify whether a gene disease association exists
  - **Multi-class classification** to assign a relation type (therapeutic, biomarker, genomic alteration)
- **Two stage pipeline:** Binary classifier → Multi-class classifier applied to **high confidence false positive predictions**
- Threshold analysis: Model outputs analyzed post training to identify optimal confidence cutoffs
- Evaluation: Used precision, recall, F1 score, accuracy, and AUC on validation and test sets

Questions ?


Gene–Disease Bipartite Subgraph

Thank You