Community Detection and Graph Neural Network-Based Prediction of Disease Genes in Human Gene Regulatory Networks

Mariia Zhokhova

Network Science Project
Program: Data Science
Department of Computer Science, Faculty of Science
University of Porto
202408799@up.pt

July 6, 2025

Abstract

This project investigates the human transcriptional gene regulatory network using the TRRUST database, aiming to uncover communities enriched in disease-associated genes and to characterize the regulatory roles of genes within these communities. We identified multiple gene communities and analyzed the distribution of activators, repressors, dual regulators, and genes with unknown regulatory roles across these groups. The results reveal heterogeneous regulatory role compositions, with several communities showing distinct enrichments in activators or repressors. Additionally, we developed several Graph Neural Network (GNN) models to predict genes likely involved in diseases based on their network context. The best-performing model demonstrated promising predictive capability, as reflected by evaluation metrics and confusion matrix analyses. Our integrative approach combining community detection, regulatory role characterization, and GNN-based prediction provides novel insights into the architecture of disease-related gene modules and offers a valuable tool for prioritizing candidate disease genes in gene regulatory networks. This methodology complements existing protein-protein and co-expression network studies by focusing on curated transcriptional regulatory interactions.

1 Introduction

Gene regulatory networks (GRNs) are critical for controlling gene expression and coordinating cellular processes. Understanding the structure and function of these networks is essential for deciphering the molecular basis of complex diseases. While much of the biological network research has historically focused on protein-protein interaction (PPI) networks [3, 13], transcriptional gene regulatory networks, which map interactions between transcription factors and their target genes, represent a relatively newer and less explored layer of gene regulation [5, 17].

The human transcriptional gene regulatory network provides valuable insights into the regulatory mechanisms underlying disease development. Advances in network science have enabled the identification of communities—clusters of genes that are densely connected and potentially functionally related—within large biological networks. Detecting such communities can highlight gene modules involved in specific biological processes or disease states. Moreover, characterizing the regulatory roles of genes within these communities, such as activators, repressors, or dual regulators, can deepen our understanding of how gene regulation influences disease.

Graph Neural Networks (GNNs) have emerged as powerful tools for learning from graph-structured data. By leveraging the network topology and node features, GNNs can predict gene functions and disease associations, enhancing the prioritization of candidate disease genes beyond traditional methods [12, 6].

In this study, we use the TRRUST database of curated human transcriptional regulatory interactions [9] to explore the organization of the gene regulatory network. We apply community detection techniques to identify gene modules enriched in disease-associated genes and analyze the composition of regulatory roles within these communities. Additionally, we develop and evaluate GNN models to predict disease involvement of genes based on their network context. This integrative approach advances the understanding of disease-related gene regulation and offers novel predictive capabilities for gene prioritization.

2 Data Acquisition

2.1 Human Transcriptional Gene Regulatory Network (GRN)

We utilized the TRRUST v2 database [10] as our primary source for curated human transcriptional regulatory interactions. TRRUST provides directed regulatory relationships between transcription factors (TFs) and target genes, including annotations of regulatory roles such as activators, repressors, dual regulators, or unknown. The dataset contains 9,396 regulatory interactions involving 2,862 unique genes out of approximately 19,000 protein-coding genes in the human genome [8], indicating partial but substantial coverage of the transcriptional regulatory landscape.

For disease associations, we integrated gene-disease knowledge from the Jensen Lab Disease Database [15]. This resource offers comprehensive information on gene-disease associations, which we used to label genes with known disease associations and study their distribution within gene regulatory network communities.

2.2 Gene Regulatory Roles

The TRRUST database includes annotations for the regulatory roles of TFs, classifying them as activators, repressors, dual regulators, or unknown. These annotations were used to characterize the regulatory roles of genes within identified communities in the gene regulatory network.

2.3 Gene Functional Annotation

To enrich gene features with functional information, we queried the UniProt REST API [1] to retrieve UniProt accession IDs and Gene Ontology (GO) terms for each gene symbol. We filtered for human genes (organism ID 9606) and extracted relevant GO biological process annotations.

3 Network Construction and Community Detection

3.1 Overview

Using the TRRUST regulatory interactions, we constructed a directed gene regulatory network (GRN) where nodes represent genes and edges represent transcriptional regulatory relationships. Each edge is annotated with the regulatory role (activation, repression, dual, or unknown) of the transcription factor with respect to its target.

We applied community detection algorithms to identify densely connected groups of genes within the GRN. We applied the Leiden algorithm [18] for community detection to identify densely connected groups of genes within the gene regulatory network. The Leiden method improves upon the Louvain algorithm by guaranteeing well-connected communities and faster convergence, making it well-suited for biological networks of this size and complexity.

The Leiden algorithm [18] improves on Louvain by guaranteeing well-connected communities and faster convergence, making it more reliable for biological networks where functional modules are expected to be tightly knit [18, 7].

The constructed human transcriptional gene regulatory network (GRN) comprises 2,862 nodes (genes) and 8,427 directed edges (regulatory interactions). The network is sparse, with an average degree of approximately 5.89 and a density of 0.0010. It consists of 26 connected components, with the largest component containing 2,804 nodes, indicating most genes form a single large regulatory module (Figure 2c).

Centrality measures reveal a highly skewed degree distribution where most genes have few connections, while a small number act as hubs with very high degree (Figure 1). Degree centrality and betweenness centrality show heavy-tailed distributions, reflecting the presence of influential genes that may act as key regulatory bottlenecks (Figures 1 and 1). Closeness centrality displays a more symmetric distribution but includes nodes with zero closeness due to disconnected components (Figure 2a).

A scatter plot of degree versus betweenness centrality (Figure 2b) highlights the correlation between these two metrics, confirming that hub genes also tend to have high control over information flow within the network.

These network characteristics demonstrate the complex hierarchical structure of the GRN and justify the application of community detection algorithms to identify biologically meaningful gene modules.

3.2 Comparison of Network Centrality Measures Between Diseased and Non-Diseased Genes

We compared the distributions of centrality measures—degree centrality, betweenness centrality, and closeness centrality—between genes labeled as diseased (label = 1) and non-diseased (label = 0).

- **Degree Centrality:** Diseased genes showed a higher median (0.00105) and mean (0.00264) degree centrality compared to non-diseased genes (median = 0.00070, mean = 0.00171). The difference was statistically significant (Mann-Whitney U test, $p = 3.79 \times 10^{-13}$).

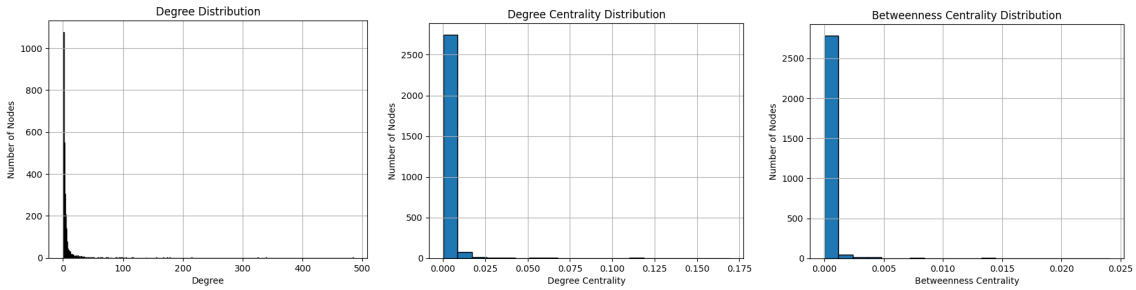
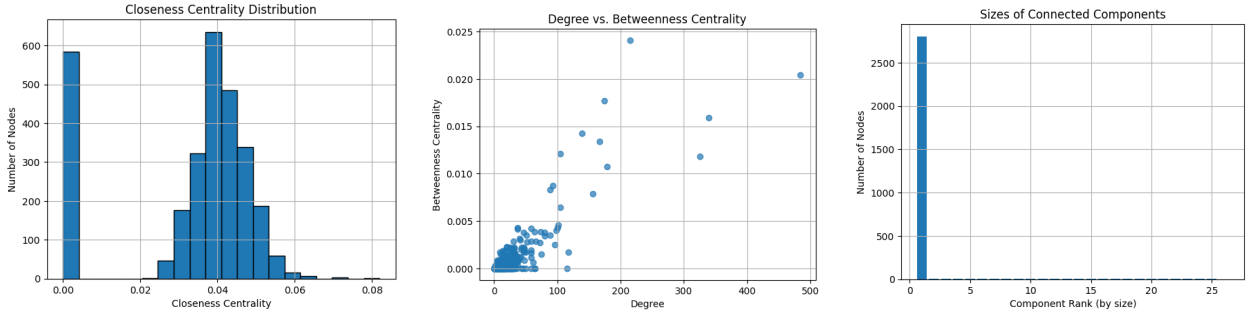


Figure 1: Distributions of (left) Degree, (center) Degree Centrality, and (right) Betweenness Centrality in the human transcriptional gene regulatory network.



(a) Closeness centrality distribution (b) Degree vs. Betweenness Centrality (c) Connected components sizes

Figure 2: Network metric visualizations: (a) closeness centrality distribution, (b) degree versus betweenness centrality scatter, and (c) connected components size distribution.

- **Betweenness Centrality:** Both groups had median betweenness centrality of zero; however, diseased genes had a higher mean betweenness (0.00022) than non-diseased genes (0.00009), with the difference also significant ($p = 3.96 \times 10^{-5}$).

- **Closeness Centrality:** Diseased genes had a slightly higher median (0.03933) and mean (0.03479) closeness centrality than non-diseased genes (median = 0.03804, mean = 0.03173), with a significant difference ($p = 2.49 \times 10^{-6}$).

These results indicate that genes associated with disease tend to occupy more central and potentially influential positions in the transcriptional gene regulatory network compared to non-diseased genes.

3.3 Top Genes by Centrality Measures

Figure 3a shows the top genes ranked by betweenness centrality, colored by disease association status. Hub genes such as *TP53* and *MYC* exhibit high betweenness, suggesting key regulatory roles.

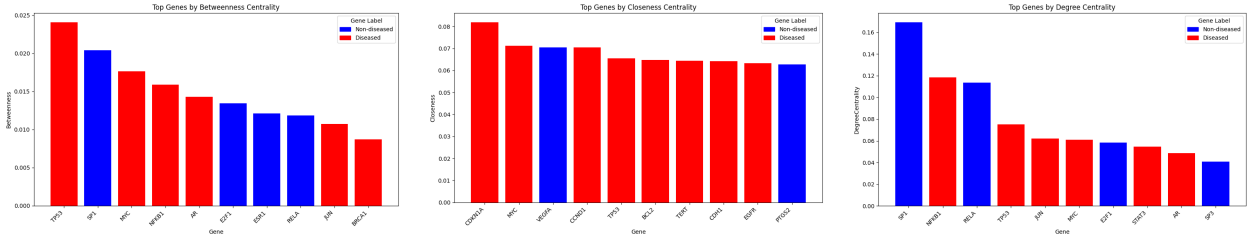
Figure 3b depicts the top genes by closeness centrality. Genes like *CDKN1A* and *VEGFA* appear central in the network, reflecting their potential influence over gene regulation pathways.

Figure 3c illustrates the top genes by degree centrality. Both diseased and non-diseased genes are well represented among the hubs, indicating complex regulatory interactions.

3.4 Statistical Comparison of Network Centrality Between Disease and Non-Disease Genes

To assess differences in network centrality metrics between disease-associated and non-disease genes, we performed Mann-Whitney U tests, a non-parametric alternative to t-tests suitable for non-normally distributed data. Specifically, we compared degree centrality, betweenness centrality, and closeness centrality distributions for the two gene groups.

The results (Table 1) show statistically significant differences ($p < 0.001$) in degree and betweenness centrality, with disease genes exhibiting higher median centrality values, suggesting they occupy more influential positions in the gene regulatory network. Closeness centrality differences were also significant but less pronounced. These findings support the hypothesis that disease-associated



(a) Top genes by betweenness centrality (b) Top genes by closeness centrality (c) Top genes by degree centrality

Figure 3: Top genes ranked by different centrality measures, colored by disease association status (red = diseased, blue = non-diseased).

genes tend to be network hubs or bottlenecks, consistent with their critical biological roles.

3.5 Community Detection and Functional Enrichment

3.5.1 Leiden Algorithm

Applying the Leiden algorithm to the gene regulatory network revealed 43 distinct communities with a modularity score of 0.4593, indicating a strong modular structure. Each community corresponds to a cluster of genes densely interconnected by transcriptional regulatory interactions.

Functional enrichment analysis of these communities revealed recurring biological themes. Most communities are significantly enriched in processes related to RNA polymerase II regulation and chromatin remodeling, reflecting the central role of transcriptional regulation and epigenetic modification in gene expression control. Metal ion binding is another common functional category across many communities, highlighting its importance in transcription factor activity and gene regulation.

Several communities also show enrichment in apoptosis, immune response, and membrane transport processes, suggesting specialized regulatory modules involved in cell death, defense mechanisms, and molecular transport.

This modular organization supports the hypothesis that gene regulatory networks are structured into functionally coherent modules, which may correspond to biological pathways or disease-associated gene clusters. Such insights can guide targeted analyses of disease gene regulation and prioritization in further studies.

3.5.2 Biological Interpretation of Community Structure

Our community detection analysis revealed 43 distinct gene modules within the human transcriptional gene regulatory network. These communities exhibit strong functional coherence, with many modules enriched for key biological processes such as RNA polymerase II regulation,

Figure 4: Visualization of the human transcriptional gene regulatory network partitioned into 43 communities using the Leiden algorithm.

chromatin remodeling, apoptosis, immune response, and metal ion binding. Such clustering reflects the modular organization of gene regulation, where genes involved in related pathways or cellular functions form tightly connected regulatory units.

RNA polymerase II regulation appeared as a dominant theme across numerous communities, underscoring its central role in transcriptional control. Chromatin remodeling, critical for modulating DNA accessibility and gene expression, was another frequent enrichment, often co-occurring with apoptotic processes in several modules. This suggests that these communities may coordinate transcriptional responses essential for cell cycle control and programmed cell death, pathways frequently implicated in disease.

Notably, specific communities also showed enrichment for immune response-related genes, highlighting the regulatory networks underpinning host defense mechanisms. The presence of metal ion binding processes in multiple

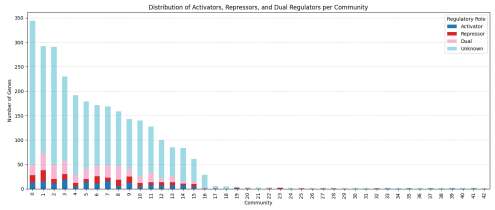
Table 1: Centrality statistics and Mann-Whitney U test p-values comparing disease-associated and non-disease genes.

Centrality	Group	Median	Mean	p-value
Degree	Non-diseased (n=1792)	0.00070	0.00171	2*3.79e-13
	Diseased (n=1070)	0.00105	0.00264	
Betweenness	Non-diseased (n=1792)	0.00000	0.00009	2*3.96e-05
	Diseased (n=1070)	0.00000	0.00022	
Closeness	Non-diseased (n=1792)	0.03804	0.03173	2*2.49e-06
	Diseased (n=1070)	0.03933	0.03479	

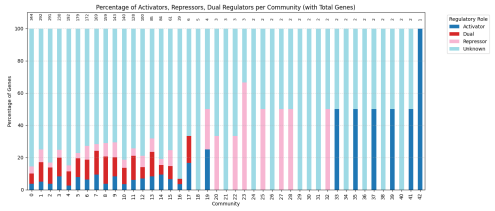
modules points to the importance of metal cofactors in enzymatic activities within these regulatory circuits.

For example, Community 1 is characterized by enrichment in chromatin remodeling and apoptosis, suggesting a role in maintaining genomic stability and regulating cell death pathways. Similarly, Community 2's association with immune response and cytokine signaling pathways reflects its potential involvement in inflammatory and immune diseases.

The diverse functional annotations across communities, coupled with varying distributions of regulatory roles (activators, repressors, dual regulators), highlight the complexity of transcriptional regulation in human cells and its relevance to disease pathogenesis.



(a) Absolute counts of regulatory roles.



(b) Percentage composition of regulatory roles.

Figure 5: Distribution of transcription factor regulatory roles across the 43 communities.

3.5.3 Distribution of Regulatory Roles Across Communities

We analyzed the distribution of transcription factor regulatory roles—activators, repressors, dual regulators, and unknown—across the 43 communities identified by the Leiden algorithm.

Figure 5a shows the absolute counts of genes with each regulatory role per community. Larger communities naturally contain more genes, but a substantial fraction of genes remain unannotated with unknown roles in many communities.

Figure 5b presents the normalized percentages of regulatory roles within each community. This highlights distinct enrichments in regulatory function, with some communities predominantly composed of activators, others enriched in repressors or dual regulators, illustrating functional specialization in gene regulation modules.

3.5.4 Network Metrics of Top Communities

We analyzed several network metrics within the top 16 largest communities identified by the Leiden algorithm. The average degree across these communities ranged from approximately 2.6 to 4.6, indicating varying levels of connectivity within gene modules. Community 2 exhibited the highest average degree (4.65), suggesting a densely connected cluster.

Clustering coefficients were generally low to moderate (0.01–0.47), with Community 2 again showing the highest average clustering coefficient (0.47), implying tightly knit local neighborhoods in that community. Other communities displayed lower clustering, reflecting more sparse local connectivity.

Average degree centrality and betweenness centrality also varied between communities, with degree centrality ranging up to 0.066 and betweenness centrality remaining low overall (below 0.0012). These results suggest that some communities have genes that act as key hubs or bottlenecks in the regulatory network, while others are more diffusely connected.

Overall, the heterogeneity in network metrics highlights functional specialization and distinct regulatory architectures among gene communities in the human transcriptional gene regulatory network.

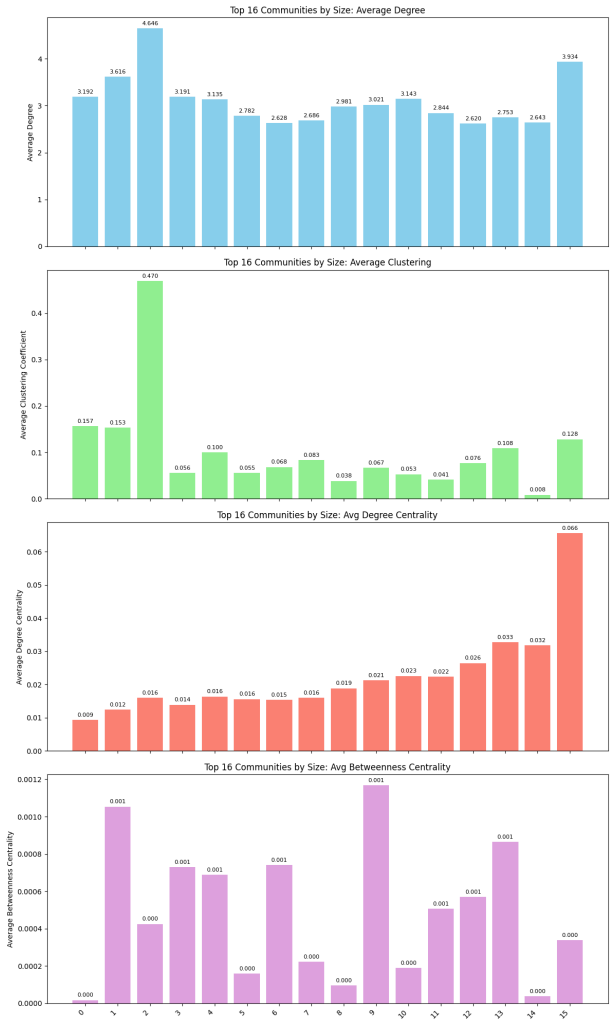


Figure 6: Network metrics for the top 16 largest communities detected by the Leiden algorithm. The plots show average degree, clustering coefficient, degree centrality, and betweenness centrality, with value labels displayed on each bar.

4 Modelling and Prediction

4.1 Feature Selection

Our feature set integrated diverse sources of biological and network information for each gene, including:

- **Gene Regulatory Role:** Categorical variable indicating whether the gene functions as an activator, repressor, dual regulator, or has an unknown regulatory role based on TRRUST annotations.
- **Community Membership:** The Leiden community ID assigned to the gene, representing its membership in a functionally coherent module.
- **Network Centrality Metrics:** Quantitative features including betweenness centrality, degree centrality, and closeness centrality calculated within the gene regulatory network. These metrics reflect the gene's influence and connectivity in the regulatory landscape.
- **Gene Ontology (GO) Features:** Binary indicators for selected GO biological process terms associated with each gene. Selection of GO terms was performed based on their *odds ratio*, measuring the strength of association between the presence of a GO term and disease gene status. GO terms with higher odds ratios indicate stronger enrichment in disease-associated genes and were prioritized for inclusion.

The combined feature matrix used for classification comprised 2,862 genes (nodes) with 754 features each. These features included categorical indicators of gene regulatory roles and community membership, continuous network centrality measures, and binary flags for selected Gene Ontology terms.

The relatively high-dimensional feature space reflects the integration of diverse biological and topological information, enabling the model to capture complex patterns associated with disease involvement.

4.2 Evaluation Metrics

To evaluate model performance on disease gene classification, we employed several complementary metrics. Accuracy provides the overall fraction of correctly predicted genes, but can be misleading in imbalanced datasets. Precision measures the proportion of predicted disease genes that are truly disease-associated, while recall (sensitivity) quantifies the fraction of actual disease genes correctly identified by the model. The F1-score harmonizes precision and recall into a single metric. Finally, the area under the Receiver Operating Characteristic curve (AUC) evaluates model discrimination capability across all classification thresholds.

In our biological context, recall is particularly critical because failing to identify disease-associated genes (false negatives) could miss key targets for further study or therapeutic intervention. Hence, models with higher recall, even at the cost of some false positives, may be preferred for prioritizing candidate disease genes.

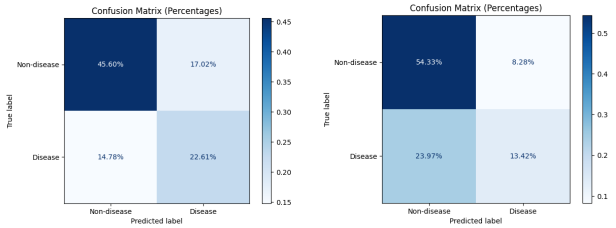
4.3 Baseline Model: Random Forest Classification

To establish a baseline for predicting disease-associated genes within the human transcriptional gene regulatory network, we implemented a Random Forest (RF) classifier. This classical machine learning method is widely

used for classification tasks due to its robustness, interpretability, and ability to handle heterogeneous feature sets.

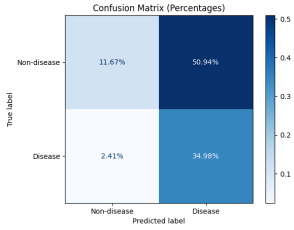
The Random Forest baseline model achieved an overall accuracy of 64% on the test set, demonstrating moderate predictive performance. The classification report (Table 2) reveals higher precision (0.67) and recall (0.84) for the non-diseased class (label 0), indicating the model reliably identifies non-disease genes.

However, for the disease-associated class (label 1), precision dropped to 0.53 and recall to 0.31, resulting in a lower F1-score of 0.39. Given the primary goal of accurately detecting disease genes, the relatively low recall indicates that many true disease genes are missed by this baseline. These results highlight the challenge of predicting disease involvement solely from gene roles, community membership, network metrics, and selected GO features. They motivate the development of more sophisticated models, such as Graph Neural Networks, which can leverage the full network structure and node features jointly for improved classification.

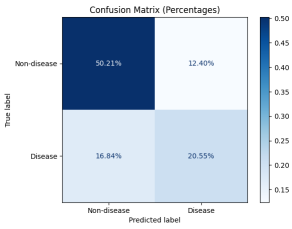


(a) DiseaseGeneGCN

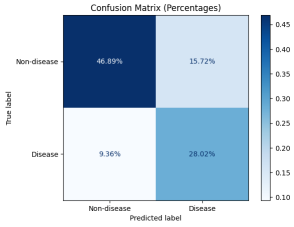
(b) ResidualGCN



(c) DiseaseGeneGAT



(d) ResidualAttentionGCN



(e) DiseaseGeneGraphSAGE

Figure 7: Normalized confusion matrices for the GNN models, showing percentages of predicted versus true labels on the test set.

Table 2: Classification report for Random Forest model predicting disease gene status.

Class	Precision	Recall	F1-score	Support
Non-diseased	0.67	0.84	0.74	359
Diseased	0.53	0.31	0.39	214
Accuracy			0.64	
Macro avg	0.60	0.57	0.57	573
Weighted avg	0.62	0.64	0.61	573

4.4 Graph Neural Network Modeling

4.4.1 Model Architectures

We implemented and evaluated multiple Graph Neural Network (GNN) architectures to predict disease-associated genes based on the transcriptional gene regulatory network:

- **DiseaseGeneGCN** (Fig. 8): A two-layer Graph Convolutional Network (GCN) with ReLU activation and dropout.
- **ResidualGCN** (Fig. 9): A deeper GCN with four convolutional layers, batch normalization, and residual connections to improve gradient flow and representation power.
- **DiseaseGeneGAT** (Fig. 10): A two-layer Graph Attention Network (GAT) using multi-head attention to weigh neighbor contributions adaptively.
- **ResidualAttentionGCN** (Fig. 11): A deeper attention-based GCN with multiple GAT layers, batch normalization, and residual connections.
- **DiseaseGeneGraphSAGE** (Fig. 12): GraphSAGE model that samples and aggregates neighborhood features inductively, enabling generalization to unseen nodes.

To comprehensively explore the predictive power of graph neural networks on the gene regulatory network, we implemented multiple architectures with varying complexities and mechanisms of information aggregation. The **DiseaseGeneGCN** model serves as a baseline, employing straightforward graph convolution layers with ReLU activations and dropout for regularization, suitable for capturing local neighborhood structure. The **ResidualGCN** extends this by stacking multiple convolutional layers combined with batch normalization and residual connections, which help mitigate vanishing gradients and enable learning of deeper, more complex patterns within the network. The **DiseaseGeneGAT** leverages graph attention mechanisms that adaptively weight neighboring nodes, allowing the model to focus on the most relevant regulatory interactions and potentially improving interpretability. Building

on this, the **ResidualAttentionGCN** incorporates multiple attention layers with residual connections and normalization to combine the benefits of depth and adaptive neighbor weighting. Lastly, **DiseaseGeneGraphSAGE** uses an inductive aggregation strategy that generalizes well to unseen nodes by sampling and aggregating neighborhood features, making it suitable for larger or evolving biological networks. Together, these diverse models enable us to balance between simplicity, representational capacity, and generalizability in predicting disease-associated genes.

4.4.2 Training and Evaluation Procedure

We used PyTorch Geometric to implement these models. The training pipeline included:

- Data splits created via stratified random sampling to form training, validation, and test sets.
- Weighted cross-entropy loss to account for class imbalance.
- Early stopping based on validation Area Under the ROC Curve (AUC) to prevent overfitting.
- Hyperparameter tuning over hidden dimensions, dropout rates, and learning rates.
- Evaluation metrics including accuracy, precision, recall, F1-score, and AUC.

Cross-validation was employed to assess model robustness. Confusion matrices and ROC and Precision-Recall curves were generated to visualize performance.

4.4.3 Hyperparameter Tuning and Class Weighting

We performed hyperparameter tuning for the Graph Neural Network models by varying hidden layer sizes, dropout rates, and learning rates. The best validation AUC achieved during tuning was 0.5864, corresponding to a configuration with a hidden dimension of 32, dropout of 0.3, and learning rate of 0.001.

Earlier configurations, such as hidden dimension 128, dropout 0.5, and learning rate 0.01, also achieved respectable AUC scores around 0.5662, demonstrating the model's sensitivity to these parameters.

To address class imbalance in the dataset, we computed class weights inversely proportional to class frequencies. The weights applied were approximately 1.60 for the non-disease class and 2.67 for the disease-associated class, thereby penalizing misclassification of the minority class more heavily during training.

These steps contributed to improving the models' ability to detect disease genes in the transcriptional regulatory network.

4.4.4 Graph Neural Network Model Performance

Table 3 summarizes the performance of various graph neural network (GNN) architectures applied to the classification of disease-associated genes. Among the models tested, the *DiseaseGeneGraphSAGE* achieved the highest accuracy (74.91%) and AUC (0.8153), indicating superior overall classification ability. While the *DiseaseGeneGAT* model showed the highest recall (93.55%), it suffered from substantially lower precision and accuracy, suggesting a tendency to over-predict the disease class. The *ResidualGCN* exhibited high precision but a lower recall, pointing to conservative predictions. The *DiseaseGeneGCN* and *ResidualAttentionGCN* provided a more balanced trade-off between precision and recall. Overall, these results highlight the importance of considering multiple evaluation metrics to understand model behavior comprehensively, especially in scenarios where recall (sensitivity to disease genes) is critical.

Note: In this context, recall is particularly important because failing to identify true disease-associated genes (false negatives) may hinder downstream biological insights and therapeutic target discovery. The DiseaseGeneGAT model shows the highest recall (0.9355), indicating strong sensitivity, though at the cost of lower precision and accuracy. Conversely, DiseaseGeneGraphSAGE provides a balanced trade-off with high accuracy and recall, making it the best overall performer for disease gene prediction.

The confusion matrices illustrate the classification performance of each graph neural network model on the test dataset. The matrices are normalized to show percentages of predictions relative to the true labels.

4.4.5 Model Discussion

- **DiseaseGeneGCN** (Fig. 7a) shows a reasonable balance between true positive and true negative rates. However, the false positive rate is somewhat elevated, indicating the model tends to over-predict disease labels.
- **ResidualGCN** (Fig. 7b) improves true negative predictions but suffers from lower recall for the disease class, resulting in more false negatives.
- **DiseaseGeneGAT** (Fig. 7c) achieves very high recall for the disease class, correctly identifying most disease genes, but this comes at the cost of increased false positives, decreasing precision.
- **ResidualAttentionGCN** (Fig. 7d) strikes a better balance with fewer false positives than GAT and improved recall compared to ResidualGCN.

Table 3: Performance metrics for GNN models on disease gene classification.

Model	Accuracy	Precision	Recall	F1-score	AUC
DiseaseGeneGCN	0.6820	0.5705	0.6047	0.5871	0.7317
ResidualGCN	0.6775	0.6184	0.3589	0.4542	0.6959
DiseaseGeneGAT	0.4665	0.4071	0.9355	0.5673	0.6603
DiseaseGeneGraphSAGE	0.7491	0.6406	0.7495	0.6908	0.8153
ResidualAttentionGCN	0.7075	0.6235	0.5495	0.5842	0.7370

- **DiseaseGeneGraphSAGE** (Fig. 7e) exhibits the best overall performance, maintaining high true positive and true negative rates, suggesting superior discrimination capability between disease and non-disease genes.

5 Results

5.1 General Results

We identified 43 distinct communities within the human transcriptional gene regulatory network, revealing a modular structure with functional enrichment in processes such as RNA polymerase II regulation, chromatin remodeling, and apoptosis. Centrality metrics showed that disease-associated genes tend to have significantly higher degree, betweenness, and closeness centralities than non-disease genes, indicating their key roles in network control and information flow.

Performance evaluation of several Graph Neural Network (GNN) architectures (Table 3) showed that DiseaseGeneGraphSAGE achieved the highest overall accuracy (74.9

These results underscore the effectiveness of combining network topology, community structure, and advanced GNN methods for disease gene prediction.

5.2 Biological Interpretation

The modular gene communities we uncovered reflect coherent functional units, with many enriched in key regulatory processes vital to cellular health and disease. Hub genes like TP53, MYC, and CDKN1A occupy central positions, consistent with their critical roles in tumor suppression, cell cycle regulation, and apoptosis, as documented in literature [14, 4, 2].

Higher centrality scores for disease-associated genes imply their importance as regulatory bottlenecks whose dysfunction could disrupt gene expression programs and promote disease. Our models' ability to leverage these structural and functional features supports their biological relevance and utility in prioritizing candidate genes for further study.

6 Conclusion

This study highlights the power of integrating transcriptional regulatory network structure with graph-based machine learning to advance disease gene prioritization. By combining community detection, centrality analysis, and multiple GNN architectures, we gained insights into gene regulatory modules and identified predictive models that balance accuracy and sensitivity.

Future directions include expanding to multi-omics integration, improving model interpretability, and applying these approaches to other complex diseases for enhanced understanding and potential clinical translation.

7 Limitations and Future Work

Despite the insights gained from our integrative analysis, several limitations must be acknowledged. First, gene-disease association data remain incomplete and potentially biased, which may affect the labeling and interpretation of disease genes [15]. Second, the TRRUST database, while carefully curated, does not capture all transcriptional regulatory interactions, limiting the network's coverage and potentially missing key regulatory relationships [10]. Notably, although there are approximately 19,000 protein-coding genes in the human genome [8], our analysis covers only 2,862 genes present in the TRRUST network, which restricts the scope of gene regulatory interactions studied.

Additionally, gene regulatory networks are inherently noisy and dynamic, but our static network representation cannot capture temporal or context-specific regulatory changes [16].

Future work could address these limitations by integrating multi-omics data sources, such as epigenetic marks, proteomics, and transcriptomics, to build more comprehensive and context-aware regulatory networks [11]. Longitudinal data could further elucidate dynamic regulatory changes during disease progression. Moreover, incorporating additional network layers, including protein-protein interactions and co-expression networks, may improve gene prioritization and provide deeper biological insights. Finally, exploring explainability methods for GNNs could enhance the interpretability of disease gene predictions,

aiding biological validation.

References

[1] Uniprot rest api documentation. `https://www.uniprot.org/help/api`. Accessed: 2025-07-06.

[2] Tarek Abbas and Anindya Dutta. p21 in cancer: intricate networks and multiple activities. *Nature Reviews Cancer*, 9(6):400–414, 2009.

[3] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[4] Chi V Dang. Myc on the path to cancer. *Cell*, 149(1):22–35, 2012.

[5] E. M. Davidson and M. Levin. Gene regulatory networks. *Proceedings of the National Academy of Sciences*, 102(14):4935–4936, 2005.

[6] B. Shariat F. Fout, J. Byrd and A. Ben-Hur. Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 30:6530–6539, 2017.

[7] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[8] Adam Frankish, Mark Diekhans, Ana-Maria Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Jo Armstrong, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[9] H. Han, J. W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H. N. Jeon, H. Jung, S. Nam, M. Chung, J. H. Kim, and I. Lee. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018.

[10] Hao Han, Jae-wook Cho, Sangyoung Lee, Anurag Yun, Hyejin Kim, Dongwoo Bae, Sungbo Yang, Chanin Kim, Juwon Lee, Bin Kang, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

[11] Yana Hasin, Michael Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[12] M. Agrawal J. Zitnik and J. Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[13] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[14] David P Lane. p53, guardian of the genome. *Nature*, 358(6381):15–16, 1992.

[15] Jan Piñero, Àlex Bravo, Nerea Queralt-Rosinach, Alberto Gutiérrez-Sacristán, Joan Deu-Pons, Eva Centeno, Joaquín García-García, Ferran Sanz, and Lucia I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

[16] Jordan F Reiter and William C Skarnes. Dynamic gene regulatory networks and network medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 9(5):e1381, 2017.

[17] D. Thieffry. From gene regulatory networks to genetic networks: A new perspective. *Current Opinion in Genetics Development*, 17(2):213–220, 2007.

[18] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

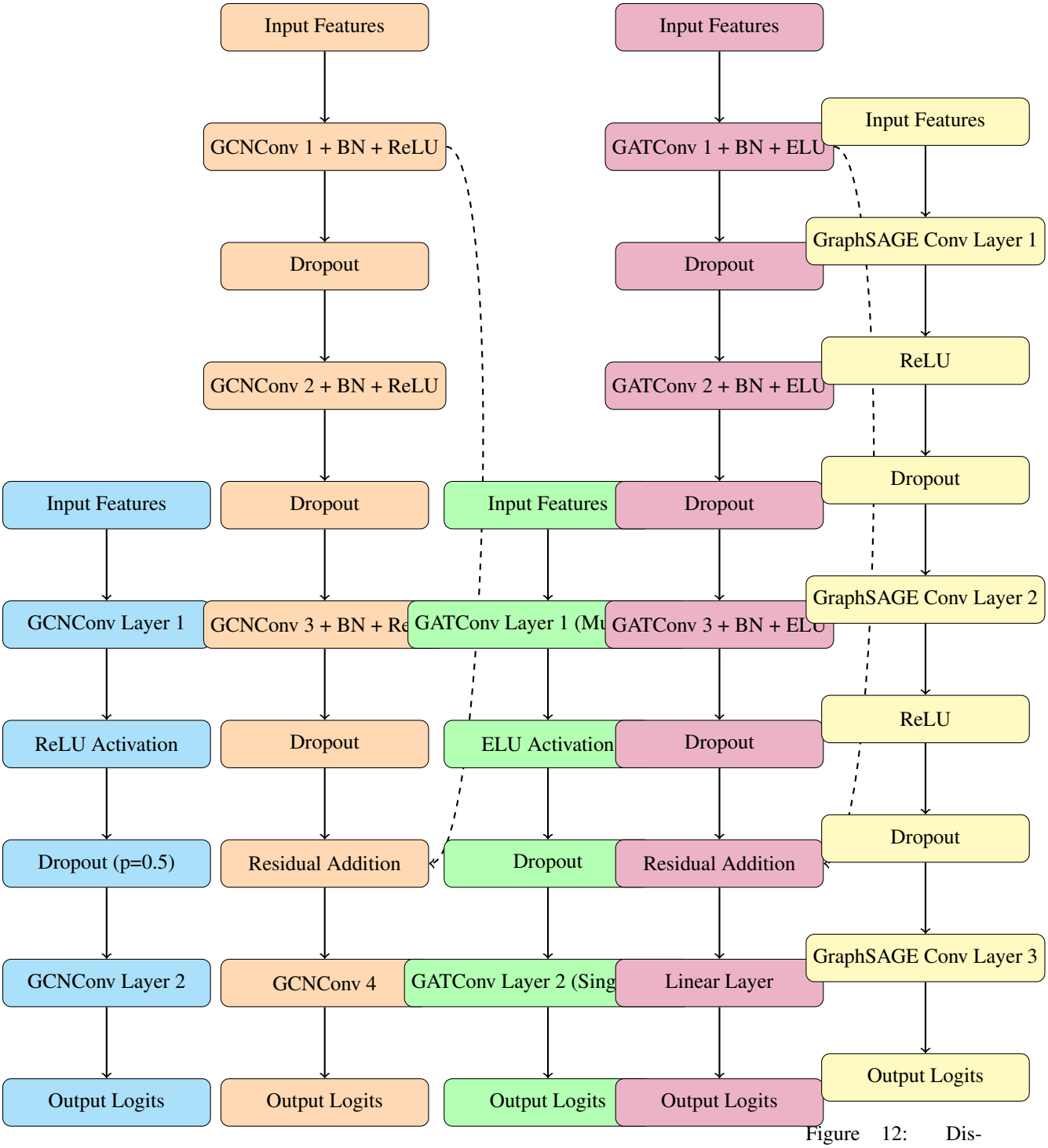


Figure 8: DiseaseGeneGCN

Figure 9: ResidualGCN

Figure 10: DiseaseGeneGAT

Figure 11: ResidualAttentionGCN

Figure 12: DiseaseGeneGraphSAGE