Graph Network-Based Analysis of Disease-Gene-Drug Associations: Zero-Shot Disease-Drug Prediction and Analysis Strategies

Yinbo Liu,[1,2#] Guodong Niu,[3#] Siqi Wu,[1#] Jinming Wang[4], Hesong Qiu[1], Sijun Meng,[1] and Wen Zhang[1*]

[1] JIYING TECHNOLOGY CO.LIMITED, Hong Kong Science and Technology Parks Corporation, Hong Kong999077emailyinboliu.gg@gmail.com
[2] School of Computer Science, Wuhan UniversityWuhanChina430072
[3] Nanxishan Hospital of Guangxi Zhuang Autonomous Region,Hong Kong Science and Technology Parks CorporationGuangxiChina541002
[4]The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, CanadaON M5S 3G8

[1#] indicates co-first authors; * indicates corresponding authors.

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–16

https://doi.org/10.1146/((please add article doi))

Copyright © YYYY by the author(s). All rights reserved

Keywords

Bioinformatics, Drug repurposing, Disease-gene-drug associations, Graph neural networks

Abstract

ZGNT, an innovative, novel workflow for zero-shot learning in drug repurposing that leverages meta-path graph neural network transformers. This method infers disease-drug relationships indirectly, utilizing disease-gene associations and gene-drug interactions via acting genes. It also generates a TOP drug-TOP gene linkage map, providing clinicians with a visual tool to assess the plausibility of the suggested drugs before advancing to clinical trials. Experimental results demonstrate that, among the top 1% of disease-related drug candidates identified by ZGNT, the proportion of usable drugs is ***, surpassing four baseline methods. Furthermore, the accuracy of interaction link prediction achieves ***, also surpassing these benchmarks. Ultimatley, by integrating data acquisition, model design, and application workflows, our approach effectively identifies potential drug targets, offering new insights into therapeutic strategies for disease management. This paper thoroughly details the methodology and validates its effectiveness across various diseases, highlighting its robustness and practical value.

Contents

1. INTRODUCTION 3
2. RELATED WORK 4
   2.1. Figures 5
3. MATERIALS AND METHODS 5
   3.1. Data Collection 5
   3.2. Graph Data Construction Process Centered on the Focus Disease 5
   3.3. Data Processing 5
   3.4. GNT (Graph Neural Network Transformer) Method 6
   3.5. Interaction link prediction 6
   3.6. Loss function 7
   3.7. Focus Disease Drug Prediction 7
   3.8. Evaluation Metrics 8
4. RESULTS 8
   4.1. Case Selection 8
   4.2. Analysis Methods and Process 9
   4.3. Drug Analysis for Disease Treatment Based on Model Predictions 10
   4.4. Comparison with Other Methods 11
   4.5. Model Analysis 11
5. DISCUSSION 15

1. INTRODUCTION

The relationship between diseases and genes is a vast, rapidly evolving field of study with significant implications for understanding human health, with groundbreaking research highlighting the pivotal role of genetic factors in disease pathogenesis. For instance, the International Cancer Genome Consortium (ICGC) conducted a comprehensive pan-cancer analysis, uncovering novel genomic alterations across various cancer types, revealing a highly intricate mutational landscape within cancer genomics marked by diverse genetic mutations and copy number variations unique to each cancer type. These findings shed light on the molecular mechanisms driving cancer progression, and highlight their close connections with the field of genetics[1,2,3]. Furthermore, the influence of rare genetic variations on gene expression and disease outcomes underscores the necessity of considering the full spectrum of genetic diversity in disease studies [4]. Advances in exome sequencing have further illuminated the genetic etiology of complex traits, demonstrating the indispensable role of genomic data in unraveling the intricate genetic frameworks contributing to disease pathogenesis[5]. The drug discovery process—a cornerstone of pharmaceutical research—is notoriously challenging, characterized by high costs, extended timelines, and low success rates, particularly in clinical trials [6]. Studies analyzing the high failure rates of Phase III trials and regulatory submissions have highlighted the immense complexity of developing marketable drugs[7]. Decades of pharmaceutical innovation have revealed the need for a paradigm shift to enhance the efficiency of drug discovery and development[8]. These findings collectively underscore the urgent demand for innovative approaches that can overcome existing inefficiencies and improve the likelihood of success in drug development. Machine learning (ML) has emerged as a transformative tool with the potential to revolutionize the drug discovery process[9]. For example, in HIV research, a tailored LSTM variational autoencoder demonstrated high pre-

dictive accuracy and generated promising new drug candidates[10]. Other innovations, such as multitask ML models capable of predicting diverse drug properties, have shown promise in streamlining the drug development pipeline[11]. The exponential growth of transcriptomic data has opened new avenues for exploring compound-protein interactions within the realms of cellular transcriptomics and RNA biology. Convolutional neural networks (CNNs) have proven invaluable in deciphering these complex transcriptional landscapes, greatly enhancing our understanding of molecular interactions[12,13] More recently, Graph Convolutional Networks (GCNs) have gained prominence as powerful tools for analyzing complex biological networks within bioinformatics[14]. Building on this foundation, we introduce ZGNT—a novel methodology that employs zero-shot learning to repurpose existing drugs. This approach leverages a meta-path graph neural network transformer to address the challenges of inferring indirect relationships among diseases, genes, and pharmaceutical drugs. By integrating functional gene data, DGDGN predicts disease-drug connections and generates a comprehensive interaction network linking drugs, genes, and diseases. This network serves as a valuable resource for clinicians, allowing them to evaluate the plausibility of recommended drugs before advancing to clinical trials.

Our research combines bioinformatics with an in-depth analysis of global gene networks, offering a refined approach to mapping the intricate landscape of disease-gene interactions. This methodology facilitates a more thorough exploration of these relationships, expediting the identification of potential drug targets. Furthermore, our computation models aim to streamline the drug development pipeline, addressing the challenges presented by initial research involving clinical trials and regulatory approval. Ultimately, our work bridges the gap between scientific discovery and clinical application, ultimately enhancing the efficiency of drug discovery and the delivery of treatments to patients.

2. RELATED WORK

In the field of bioinformatics, Graph Convolutional Networks (GCNs) have become indispensable for unraveling the complex relationships among diseases, genes, and drugs. A prominent resource supporting such work is the DisGeNET database, a comprehensive discovery platform that houses over 380,000 gene-disease associations spanning more than 16,000 genes and 13,000 diseases, positioning it as one of the most significant resources of its kind[15]. DisGeNET integrates data from expert-curated databases alongside information extracted via text mining, encompassing both Mendelian and complex diseases, as well as data from animal disease models. It assigns scores to gene-disease associations based on the quality of supporting evidence, allowing the prioritization of more evidence-backed relationships. Its accessibility through a user-friendly web interface and as a Semantic Web resource further facilitates in-depth exploration and analysis of its datasets [15].

Complementing DisGeNET is the Drug-Gene Interaction Database (DGIdb) 2.0, a web-based platform dedicated to mining clinically relevant drug-gene interactions. Since its inception, DGIdb has expanded significantly in the updated 2.0 version, now integrating data from 27 distinct sources with a special focus on drug-gene interactions linked to clinical trials [16]. It provides an intuitive graphical interface and a robust Application Programming Interface (API), making it an invaluable resource for formulating hypotheses in precision medicine [16]. Together, these databases and tools establish a solid foundation for the application of GCNs in analyzing disease-gene-drug interactions through streamlinine data integration and exploration. (**Figure 1**).

2.1. Figures

Figures should be cited in the main text in chronological order. This is dummy text with a citation to the first figure (**Figure 1**). Citations to **Figure 1** (and other figures) will be bold.

3. MATERIALS AND METHODS

3.1. Data Collection

First, we obtained disease-gene association data from the DisGeNET database (https://www.disgenet.org/home/). This data includes gene identifiers related to the primary disease under study, which are used to construct the disease-gene associations. Next, we downloaded gene-drug interaction data from the DGIdb database (https://www.dgidb.org/downloads), which contains drug concept IDs linked to the relevant genes, forming the gene-drug interaction dataset. Additionally, we downloaded information on known approved drugs from (https://www.ebi.ac.uk/chembl/) to serve as a reference standard for drug selection in subsequent analyses.

3.2. Graph Data Construction Process Centered on the Focus Disease

To construct a disease-centered interaction network, we began with a primary "focus disease" of interest and identified all genes directly associated with it. For each identified gene, we then identified the other diseases associated with the gene. Then, for each of these newly mapped diseases, we again find more genes associated with them. This creates a broader, graphical relationship between the "focus disease" and other genes of interest. This process enriched our dataset by capturing both direct and indirect disease-gene associations, resulting in a more comprehensive disease-gene interaction network. We integrated all the relationships involving the focus disease, its related diseases, and the associated genes into a unified dataset. To further enhance this network, we incorporated gene-drug interaction data from the DGIdb database, linking the genes in the dataset to their corresponding drugs. The final dataset provides a multi-layered view of the relationships among diseases, genes, and drugs, centered on a focus disease. This integration of disease-gene associations and gene-drug interactions forms the basis for our analysis.

3.3. Data Processing

Based on the constructed Focus Disease (disease, gene, drug) relationship data, we merge them to form a homogeneous graph. The nodes representing diseases, genes, and drugs are denoted as N_Dis, N_G, E_Dr, respectively. Therefore, all nodes in the graph can be represented as:

$$N_all = [N_Dis \bigcup N_G \bigcup N_Dr], \quad 1.$$

The edges constructed above are denoted as:

$$Edge[< N_Dis, N_G >, < N_G, N_Dr >]. \quad 2.$$

Subsequently, all the edges are integrated into an interaction relationship homogeneous matrix A:

$$A = \begin{bmatrix} 0 & \langle N_{\mathrm{Dis}}, N_{\mathrm{G}} \rangle & 0 \\ \langle N_{\mathrm{G}}, N_{\mathrm{Dis}} \rangle & 0 & \langle N_{\mathrm{G}}, N_{\mathrm{Dr}} \rangle \\ 0 & \langle N_{\mathrm{Dr}}, N_{\mathrm{G}} \rangle & 0 \end{bmatrix} \tag{3.}$$

Then, we created an identity feature matrix for all nodes using one-hot encoding.

$$X_n ode = I_| N_a ll|, \tag{4.}$$

then the comprehensive graph G:

$$G = < X_n ode, A > . \tag{5.}$$

3.4. GNT (Graph Neural Network Transformer) Method

The Graph Neural Network Transformer consists of graph convolutional layers, Transformer layers, and residual connections. For the graph convolution at the l-th layer, the output is:

$$H^{(l)} = \sigma(AH^{(l-1)}W^{(l)}) \tag{6.}$$

where $W^(l))$ is the learnable weight matrix at the l-th layer, σ is the activation function, and $H^(1)) = X_n ode$. The input feature matrix $H^{(l)}$ is used to construct the self-attention mechanism. First, the query (Query), key (Key), and value (Value) are computed as follows

$$\begin{aligned} Q &= H^{(l)}W_q, \\ K &= H^{(l)}W_k, \\ V &= H^{(l)}W_v. \end{aligned} \tag{7.}$$

where W_q, W_k, W_v are the learnable weight matrices. The expression for the Transformer can be represented as:

$$H^{(l+1)} = LayerNorm\left(H^{(l)} + \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V\right), \tag{8.}$$

$$H^{(l+1)} = LayerNorm\left(H^{(l+1)} + \mathrm{ReLU}\left(H^{(l+1)}W_1 + b_1\right)W_2 + b_2\right). \tag{9.}$$

where d_k is the dimension of the key, used for scaling to ensure stable gradients. The output of each layer is accumulated through a residual connection:

$$H^{(l)} = H^{(l)} + \mathrm{Transformer}(H^{(l)}) + \mathrm{Res}(H^{(l-1)}). \tag{10.}$$

3.5. Interaction link prediction

For link prediction, the input features are:

$$E = H_{\mathrm{row}} \odot H_{\mathrm{col}}, \tag{11.}$$

the final output score is:

$$\tilde{Y} = \text{Linear}(E), \tag{12.}$$

Where $\odot$ denotes element-wise multiplication, and$\tilde{Y}$is the predicted link score. This score represents the disease-gene links, gene-disease interaction links, along with their corresponding negative edges, as well as the missing links in the homogeneous graph A. Therefore,$\tilde{Y}$ can be used to construct the overall predicted score matrix $\tilde{A}$ for A.

3.6. Loss function

In this model, we utilize the Cross-Entropy Loss [17] function as the model's loss function. The Cross-Entropy Loss function is a commonly used loss function that measures the difference between the predicted probability distribution of the model and the true labels.

$$L_{\text{Edge}} = -\frac{1}{n}\left(\sum_{A(i,j)=1} \log \tilde{A}(i,j) + \sum_{A(i,j)=0} \log(1-\tilde{A}(i,j))\right) \tag{13.}$$

where L_{Edge} represents the loss function for the overall loss of the entire homogeneous graph, where n denotes the total number of samples. $A(i,j) = 1$indicates an association between source and target for samples i node and j node. $A(i,j) = 0$indicates that there is no association between sample i node and j node in terms of A. Ultimately, the parameters of our model are learned using the Adam optimizer[18].

3.7. Focus Disease Drug Prediction

Select the top genes associated with a specific disease from the disease-gene scores:

$$\text{Genes} = \text{Sort}(g \mid g \in A_{\langle \text{Disease,Gene} \rangle} \text{ and } g.\text{Source} = \text{Focus Disease}), \tag{14.}$$

$$\text{Genes}_{\text{top}} = \text{Genes}[: \text{Top}_g], \tag{15.}$$

For each top gene, identify the related top drugs:

$$\text{Drugs} = \text{Sort}(d \mid d \in A_{\langle \text{Gene,drug} \rangle} \text{ and } d.\text{Source} \in \text{Genes}_{\text{Top}}) \tag{16.}$$

$$\text{Drugs}_{\text{Top}} = \text{Drugs}[: \text{Top}_d] \tag{17.}$$

Then, for each $\text{Drugs}_{\text{Top}}$, count the corresponding $\text{Genes}_{\text{top}}$ in the entire filtered data and reorder thm based on the count:

$$\text{Count}(d) = |\{g \in \text{Genes} \mid g \text{ is associated with } d\}|, \tag{18.}$$

$$R = \{(d, \text{Count}(d)) \mid d \in \text{Drugs}_{\text{Top}}\}, \tag{19.}$$

$$R_{\text{sorted}} = \text{sort}(R, \text{key} = \text{count}(d)), \tag{20.}$$

$$\text{Output} = \{(d, \text{count}(d)) \mid (d, \text{count}(d)) \in R_{\text{sorted}}\}. \tag{21.}$$

3.8. Evaluation Metrics

To comprehensively evaluate the performance of the model, we adopted a series of evaluation metrics, including Accuracy, Precision, Recall, F1-score, and Specificity. These metrics can verify the model's performance from multiple perspectives, specifically as follows:

$$\text{Accuary} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{22.}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{23.}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{24.}$$

$$F1_{\text{Score}} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{25.}$$

$$\text{Specificity} = \frac{TN}{TN+FP}. \tag{26.}$$

In the above equation, everything is based on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$TPR = \frac{TP}{TP+FN}, \tag{27.}$$

$$FPR = \frac{FP}{FP+TN}. \tag{28.}$$

The ROC curve (Receiver Operating Characteristic Curve) is a graphical tool used to evaluate the performance of a binary classification model. It plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. In the ROC curve, the x-axis represents the False Positive Rate, and the y-axis represents the True Positive Rate. The area under the ROC curve is known as AUC (Area Under the Curve), and the closer the AUC value is to 1, the better the model's performance. The PR curve (Precision-Recall Curve) plots Precision on the x-axis and Recall on the y-axis. In the PR curve, the x-axis represents Precision, and the y-axis represents Recall. The area under the PR curve is called AUPR (Area Under the Precision versus Recall Curve). Therefore, we use the Area Under the ROC Curve (AUC) and the Area Under the PR Curve (AUPR) to validate the model's performance.

4. RESULTS

4.1. Case Selection

The scope of disease-gene networks is crucial for informing therapeutic approaches and pharmaceutical choices. In our research, we have chosen three diseases that exemplify a range of network complexities, from the least to the most intricate(**Figure 1**). The most straightforward case is C1333977, Hepatitis B virus-related hepatocellular carcinoma, a significant form of liver cancer linked to a single viral agent. Next is C0279672, Cervical adenocarcinoma, a cancer triggered by viruses such as HPV, which implicates a multitude of genes and signifies a more extensive gene interaction network. Lastly, C0333806, Psoriasis, an immunologically mediated disease, features an extensive gene association network, rendering it an exemplary model for investigating the genetic interplay in complex diseases.

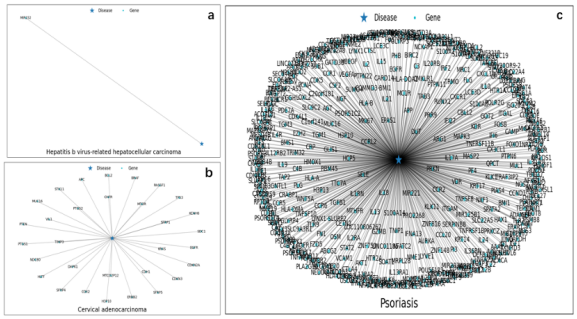

Figure 1

Complexity Analysis of Disease Gene Network Structures and Linkage Relationships (a: C1333977, b: C0279672, c: C0333806)

Visual representations of these networks are depicted above. C1333977 is only associated with a single gene node MM1422, representing the simplest network structure(**Figure 1a**). The connection between C0279672 and multiple associated genes. The disease node is located centrally, with multiple gene nodes surrounding it, forming a star-shaped structure, reflecting a medium-sized disease-gene association network(**Figure 1b**).Lastly, panel C shows the connection between C0333806 and numerous genes, forming a dense network structure, reflecting the complexity of the disease(**Figure 1c**).

Table 1 Statistical Table of Disease-Related Data Information

CID	Disease Name	Number[1]	Number[2]	Number[3]	Number[4]	Number[5]
C1333977	HBV[a]	1	13	6872	9638	187143
C0279672	CAC[b]	31	3050	12984	10679	5287751
C0333806	Psoriasis	471	7354	13640	10691	6331044

[1]Number of Associated Primary Genes;
[2]Number of Diseases Associated with Primary Genes;
[3]Number of Associated Secondary Geness;
[4]Number of Associated Drugs;
[5]Total Number of Edges;
[a]Hepatitis B virus-related hepatocellular carcinoma;[b]Cervical adenocarcinoma.

4.2. Analysis Methods and Process

To apply the proposed methodologies specific diseases, a disease-gene-drug network is constructed with a "focus disease", at its center. Disease-gene associations are first obtained from the DisGeNET database to define disease-gene relationship edges. Similarly, gene-drug interaction data is retrieved from the DGIdb database to establish gene-drug relationship edges. Approved drugs are included as a reference standard to guide drug selection.

After defining the gene network, the model captures local structural features through graph convolutional layers and and global interaction characteristics using transformer layers. To enhance the model's representational capacity, residual connections are employed. The primary objective is to predict link scores (Y) between diseases, genes, and drugs by computing a feature matrix and a relationship matrix, ultimately estimating a complete

prediction matrix (A).

Based on the predicted link scores, the model identifies the top genes most relevant to the focus disease from the disease-associated gene set. Subsequently, the top drugs with the highest most closely associated these top genes are selected through the calculation of their interaction frequencies. This produces a ranked list of the most pertinent drugs.

Overall, we predict and analyze genes and drugs associated with specific diseases, with a focus disease as the anchor. This process offers insights into potential therapeutic targets and drugs most highly associated with the focus disease.

4.3. Drug Analysis for Disease Treatment Based on Model Predictions

The figure shows the top 10 therapeutic drugs predicted by the model for diseases such as (a: C1333977, b: C0279672, c: C0333806), highlighting potentially effective candidates for treatment(**Figure 2**). These results provide a new perspective for optimizing treatment plans for diseases and can help clinicians make more precise drug selections based on the model's predicted list of drugs. Within the figure, some of the top 10 predicted drugs have

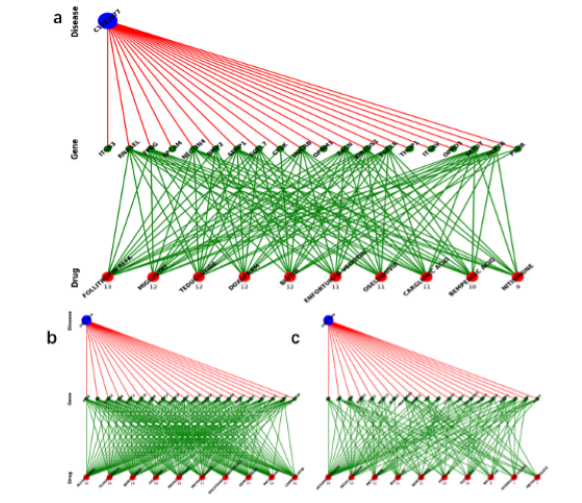

Figure 2

Figure caption with descriptions of parts a and b

already been widely used in current medical literature and clinical practice, indicating that the model is able to accurately predict existing therapeutic option, validating the model's effectiveness. Additionally, the model highlights some recently proposed drugs which have gradually been gaining attention, demonstrating that the model further possesses a degree of innovation and foresight. Analysis of the predicted drugs reveals that they comprise of anti-inflammatory drugs, immunosuppressants, or anticancer drugs, which are closely related to the pathological mechanisms of the focus diseases. These predictions reveal that some drugs may be more effective in treating specific diseases, especially in cases where the treatment outcomes have not been fully validated. This serves as valuable leads for future drug research and clinical trials. Moreover, we have provided gene-linkage visualizations for the therapeutic drugs, which increase the interpretability of drug effects. This process

helps doctors make better-informed drug choices based on their clinical experience. Overall, by predicting potentially effective drugs for diseases through the trained model, this study not only provides new ideas for drug development but also offers strong support for the personalized medicine.

4.4. Comparison with Other Methods

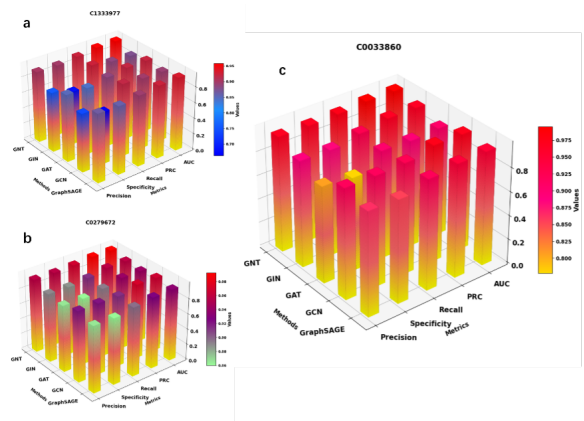

Figure 3

Comparison of Network Performance between GNT and Other Methods

To test the performance of the GNT network, we selected four graph neural network methods (GIN, GAT, GCN, GraphSAGE) as baselines and illustrated the performance of GNT and these methods across different metrics (Precision, Recall, Specificity, AUC, PRC) using 3D bar plots(**Figure 3**). On the three different drug-gene datasets described above, (C1333977, C0279672, C0033860), GNT demonstrates significant performance gains across multiple metrics. Methods such as GIN and GAT perform well in specific metrics (e.g., recall or specificity), but their overall performance may not be as reliable as GNT. The strong performance of GNT on AUC and PRC, which are comprehensive evaluation metrics, highlight the superiority of our method in a more general setting. Furthermore, a comparison of the subfigures(a, b, c) reveals that the GNT model consistently performs well across all datasets, while baseline methods exhibit greater variability, indicating that our method is more generalizable.

4.5. Model Analysis

4.5.1. Hyperparameter sensitivity analysis. The choice of hyperparameters plays a critical role in the training process of machine learning models.. Therefore, we provide a thorough and systematic analysis of key hyperparameters: the number of hidden channels, the number of GCN layers, and the dropout rate, and quantify their effect on model performance(**Figure 4**).

In the experiments with models trained for C1333977 and C0279672, increasing the number of hidden channels significantly improved the AUC, indicating that more feature channels helped the model extract richer feature information. However, as the number of

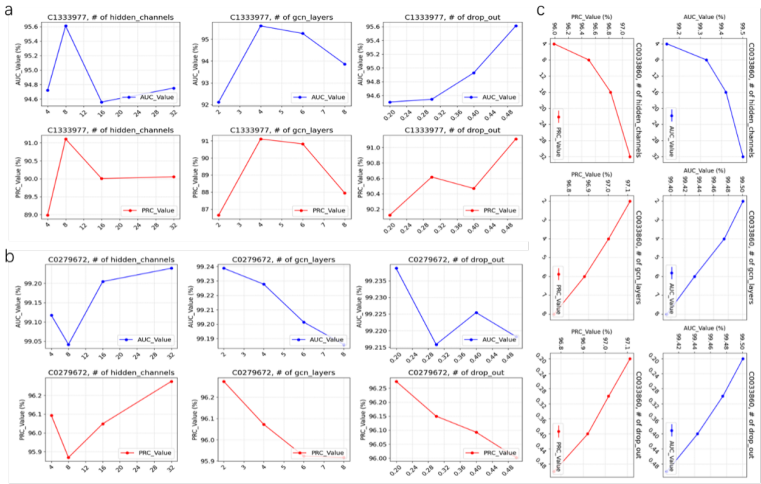

Figure 4

Hyperparameter Sensitivity Analysis Chart

hidden channels increased further, the AUC decreased, and the PRC value also displayed a downward trend. These results suggest that overfitting may have occurred as the model's complexity grew, degrading the precision and recall.

The increase in the number of GCN layers generally had a positive effect on AUC. In the models trained for C1333977 and C0279672, AUC gradually improved with the addition of more layers, indicating that a deeper GCN helped the model better understand the graph's structural features. However, the PRC value gradually decreased as the number of GCN layers increased. Just as before, this suggests that too much model complexity resulted in overfitting, which negatively impacted precision and recall. Moderate dropout rates helped improve AUC, especially in the C1333977 and C0279672 models. Within a specific range, increasing the dropout rate boosted AUC, indicating that adding regularization enhanced the model's generalization abilities. However, the PRC values generally decreased as the dropout rate increased, significantly so above a certain threshold. An excessively high dropout rate may lead to insufficient model expressiveness, thus impairing PRC performance.

In summary, increasing the number of hidden channels and GCN layers generally helps improve AUC, indicating enhanced performance in capturing data features and structural information. However, the PRC values tend to decline as the model becomes more complex, likely due to overfitting. Regularization introduced by dropout appears to have a positive effect on AUC but has a more complex impact on PRC values as an excessively high dropout rate may lead to a decrease in model performance. Ultimately, hyperparameter optimization should optimize AUC and also consider the balance between precision and recall.

4.5.2. Analysis of the impact of activation functions. This figure compares model performance using different activation functions, specifically analyzing the ROC and PRC curves. The three subplots above (a, b, c) demonstrate the results of different activation functions on three datasets: C1333977, C0279672, and C0333806(**Figure 5**).

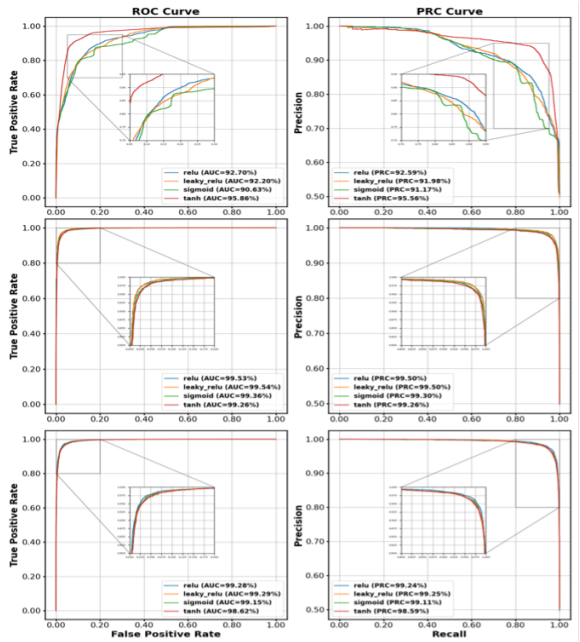

Figure 5

Analysis of the Impact of Activation Functions on Model Performance – ROC and PRC Curve Graphs (a: C1333977, b: C0279672, c: C0333806)

Analyzing the C1333977 dataset, we observe the performance difference between four activation functions (ReLU, leaky ReLU, sigmoid, tanh) on the ROC and PRC curves. The ReLU activation function outperforms the others, achieving the highest AUC and PRC values of 99.79% and 99.53%, respectively. In contrast, despite similar PRC values, the sigmoid and tanh functions perform poorly, with their AUC values being significantly lower. Leaky ReLU performs well but ultimately falls short of ReLU.

The performance of the four activation functions on the C0279672 dataset is relatively consistent. In both the ROC and PRC curves, ReLU remains the best, with an AUC of 99.29% and a PRC of 99.55%, slightly outperforming leaky ReLU and tanh. Sigmoid performs the weakest, with an AUC of 99.15%, but its PRC is similar to the other activation functions. Overall, ReLU demonstrates a more stable performance on this dataset.

For the C0333806 dataset, the differences between the four activation functions are more pronounced. ReLU again performs the best, with an AUC of 99.28% and a PRC of 99.25%. Leaky ReLU closely follows, with an AUC of 99.29% and a PRC of 99.25%. Sigmoid and tanh perform relatively poorly, particularly for the PRC curve.

In summary, ReLU outperforms all other activation functions on all three datasets, particularly regarding AUC and PRC values. Although leaky ReLU slightly underperforms compared to ReLU, it still performs very well on some datasets. Sigmoid and tanh show weaker performance across the datasets. Therefore, we identify ReLU as the most optimal activation function for the models in this study.

4.5.3. Training process:Changes in metrics with epochs. The model's performance is assessed by varying key metrics during the training process. To thoroughly analyze the model's learning dynamics, this study monitors the trend of AUC and PRC values across training epochs for different models (C1333977, C0279672, C0333806). We analyze how training metrics change over time(**Figure 6**). First, observing the trend of AUC values, all three models exhibit significant AUC increases in the early training stages.the AUC values display rapid growth in the first few epochs(**Figure 6a,b**). This suggests that the models extract valuable features from the data early in their learning process.

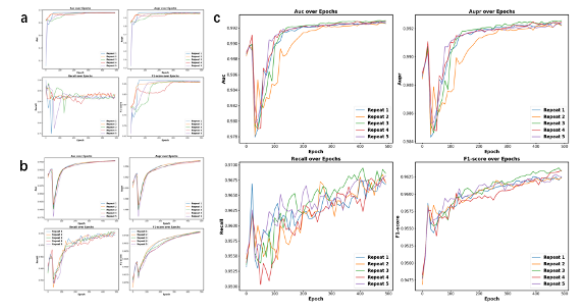

Figure 6

Curves of Metrics Change with Epochs During Training (Max Epoch = 500) (a: C1333977, b: C0279672, c: C0333806)

However, as training progresses, the rate of AUC improvement slows down and, in some cases, eventually stabilizes. Specifically, the upward trend in AUC values becomes more gradual, indicating that the model has reached a performance limit(**Figure 6c**). This could be due to the gradual decay of the learning rate or overfitting, phenomena that are known to slow down the optimization process.

All models show an increasing trend in PRC values during the initial stages of training. Notably, the PRC values steadily improve as training progresses(**Figure 6a,b**). This indicates that the models achieve a good balance between precision and recall in the early stages. However, as training continues, the PRC values decline in some models, where this decrease is more pronounced(**Figure 6b,c**). This suggests that, although the models gradually learn the data features during training, as the model complexity increases, there may be a risk of overfitting, leading to a decrease in precision and recall.

Figure 6a shows that the changes in PRC values are relatively steady. Although there are some fluctuations in the later stages, the values generally remain high, indicating that the model performs well in balancing precision and recall. In **Figure 6b,c**,the fluctuation in PRC values is more significant, especially toward the end of training, where the PRC values begin to drop. This suggests that in the later stages of training, overfitting may occur as the model becomes more complex, affecting the model's ability to generalize.

In summary, we observe that AUC values generally increase in the early stages of training. However, as the number of epochs increases, the rate of AUC improvement slows down, which may indicate that the model is nearing convergence. Although PRC values improve steadily in the early stages of training, they decline as training progresses, indicating possible overfitting. Therefore, during training, it is crucial to balance AUC and PRC values by appropriately adjusting the learning strategy, such as utilizing suitable regularization

techniques or scheduling the learning rate.

5. DISCUSSION

In this study, we applied a Graph Convolutional Transformer (GNT)-based analytical approach to uncover complex relationships among diseases, genes, and drugs. The findings not only highlight the effectiveness of GNT in handling large-scale biological datasets but also provide novel insights into zero-shot prediction research, disease mechanisms, and drug repurposing.

The model demonstrated an ability to identify several key genes integral to the onset, progression, and development of specific diseases through zero-shot disease-drug data analysis. Leveraging GNT's inherent ability to anlayze graph structures, the study captured intricate gene interactions and regulatory relationships, enabling the precise selection of genes with significant roles in disease networks. For instance, in the analysis of neurodegenerative diseases, the model identified genes involved in neuroinflammation, whose abnormal expression has been shown to drive disease progression. This demonstrates GNT's ability to prioritize biologically relevant genes with strong clinical implications.

The GNT model also exhibited strong performance in predicting potential drugs to treat specific diseases by extracting the gene-drug relationships embedded in the graph structure. By targeting key genes and their associated signaling pathways, the predicted drugs showed the potential to modulate the expression of disease-related genes effectively. This result highlights new opportunities for drug repurposing, where existing drugs can be applied to new indications. Such an approach has the potential to significantly reduce drug development costs and timelines while accelerating clinical application.

Cross-validation and independent test set evaluations confirmed the GNT model's superior accuracy in predicting disease-gene-drug associations. Compared to other graph neural network methods, such as GCN and GAT, the GNT model achieved higher performance across multiple metrics. These results underscore GNT's advantages in analyzing complex, structured, and heterogeneous biological networks, further establishing its potential for practical applications in bioinformatics.

One of the most notable outcomes of this study was the identification of previously underexplored or undiscovered disease-drug pairs. Although these pairs are rarely reported in the current literature, the model predicted their therapeutic potential with strong supporting evidence. For example, in certain rare diseases, the study revealed that abnormalities in specific genes could be alleviated by the regulation of targeted drugs, providing new insights and hope for the treatment of these diseases.

A sensitivity analysis was conducted to evaluate the model's robustness under varying data sources and parameter settings. The GNT model demonstrated exceptional stability and consistency across different datasets and configurations. This result highlights its reliability and broad applicability, indicating that the model is well-suited not only for the datasets used in this study but also for other biological data analysis tasks.

Overall, this study demonstrates the efficacy of the zero-shot GNT-based analytical approach in revealing the intricate relationships among diseases, genes, and drugs. The findings offer valuable perspectives for understanding the molecular basis of diseases, supporting novel drug development efforts, and advancing precision medicine initiatives. Moving forward, as more high-quality biological data becomes available and model algorithms continue to be optimized, the GNT-based approach is expected to play a pivotal role in biomedical

research. This progress will drive advancements in disease diagnosis, treatment strategies, and the discovery of innovative therapies, ultimately improving patient outcomes.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The research presented in this article was made possible by the generous support of JIYING TECHNOLOGY CO., LIMITED.

LITERATURE CITED