The Application of Graph Neural Networks in the Mining of Gene-phenotypic-Drug Networks for Cardiovascular Diseases

Sichen Yang
Jinqiu International High School, Wuhan, China

Abstract: Cardiovascular diseases, as the leading global health threat, involve a complex interactive network of genetic variations, phenotypic characteristics and drug effects in their pathogenesis. Traditional research methods are limited by the singularity of data dimensions and the concealment of relationships, making it difficult to systematically analyze the multi-level associations of genes, phenotypes, and drugs (G-P-D). Graph Neural Network (GNN) provides a new paradigm for mining G-P-D networks of cardiovascular diseases by integrating non-Euclidean structured data. This article systematically expounds the value of GNN in cardiovascular disease research from three dimensions: theoretical framework, application scenarios, and challenges. It focuses on analyzing its application logic in gene function analysis, phenotypic association discovery, and drug repositioning, providing theoretical support for the construction of a precision medical decision-making system.

Keywords: Graph Neural Network; Cardiovascular Diseases; Gene-Phenotypic - Drug Network; Precision Medicine; Heterogeneous Data Integration

1. Introduction

Cardiovascular diseases are the chronic disease group with the highest mortality rate worldwide, covering multiple subtypes such as coronary heart disease, heart failure, and hypertension [1]. Its pathological mechanism is highly complex, involving multiple interactions such as genetic variation, environmental exposure, lifestyle and drug intervention [2]. For instance, the CYP2C19 gene polymorphism significantly alters the efficacy of antiplatelet therapy by influencing the metabolic efficiency of clopidogrel [3]; APOE gene variations are closely related to the stability of atherosclerotic plaques and the reactivity to statins [4]. Such associations can only be revealed by integrating genotype, clinical phenotype and drug response data. However, traditional research methods are limited by the singularity of data dimensions and the concealment of relationships, making it difficult to systematically analyze the multi-level network characteristics of gene - phenotype - drug.

The traditional research paradigms mainly include genomic association analysis, phenomics analysis and drug clinical trials. Although GWAS has identified a large number of gene loci related to cardiovascular diseases, it can only explain part of the genetic risk. The remaining Missing Heritability may stem from gene-environment interaction or epigenetic modifications [5]. Phenomics research focuses on the relationship between clinical indicators and prognosis, but it is difficult to capture the dynamic associations among phenotypes and gene regulatory pathways [6]. Due to the limited sample size and individual heterogeneity in drug clinical trials, it is difficult to comprehensively evaluate the efficacy and safety of drugs. This kind of method is essentially a "single-dimensional slice" analysis, lacking the ability to model the overall structure of the G-P-D network, resulting in the key correlations being easily overlooked.

Graph neural networks, as a deep learning framework for processing non-Euclidean structured data, can effectively model heterogeneous relationships in G-P-D networks through node embedding and message passing mechanisms [7]. Its core advantage lies in supporting the unified representation of multiple types of nodes (genes, proteins, clinical indicators, drug molecules), and revealing the implicit biological associations through edge weight learning [8]. For instance, in hypertension research, GNN can integrate gene expression data, blood pressure monitoring records, and antihypertensive drug information to construct a dynamic network of "genes - blood pressure fluctuations - drug dosage",

providing a basis for individualized medication [9]. In addition, GNN can adapt to the time-varying characteristics of network structure during the progression of cardiovascular diseases, such as the systematic changes in gene expression profiles and drug reactivity during the pathological evolution of myocardial infarction patients from the acute phase to the chronic phase [10].

This paper systematically explores the application value of GNN in the mining of G-P-D networks for cardiovascular diseases from three aspects: theoretical framework innovation, application scenario expansion, and breakthrough of technical challenges. Theoretically, analyze how GNN ADAPTS to the characteristics of cardiovascular data through heterogeneous graph modeling, attention mechanisms, and dynamic graph learning; At the application level, the focus is on elaborating the technical logic in the localization of pathogenic genes, the discovery of phenotypic associations, and drug relocalization. At the challenge level, core issues such as data heterogeneity, model interpretability, and difficulties in dynamic modeling are discussed. This study aims to provide methodological support for precision medical decision-making and promote the paradigm transformation of cardiovascular disease research from "single-dimensional association" to "networked reasoning".

2. The Theoretical Basis of Graph Neural Networks and Their Compatibility with Cardiovascular Diseases

2.1 Core Mechanisms and Variants of GNN

Graph neural networks update the target node embedding by iteratively aggregating the information of neighboring nodes. The core idea lies in encoding the network structure information into the node representation. The basic GNN model integrates neighbor features through mean aggregation or Max pooling operations, but such methods assume that all neighbor nodes are equally important and have difficulty handling the complexity of heterogeneous nodes and edges. In view of the particularity of cardiovascular disease data, GNN has derived multiple variants to meet the needs of different scenarios.

Graph attention networks can analyze the asymmetric relationship between transcription factors and target genes in gene regulatory networks by dynamically allocating the weights of neighboring nodes through the introduction of an attention mechanism. For instance, in the research on the regulation of genes related to myocardial hypertrophy, the GAT model can identify the differentiated effects of key transcription factors (such as GATA4) on downstream target genes (such as MYH7), and its attention weight distribution is highly consistent with the experimentally verified regulatory intensity. Heterogeneous graph neural networks support the joint modeling of multiple types of nodes and edges, and reveal the multi-level regulatory pathways of "gene-miRNA-phenotype" in cardiovascular diseases by defining meta-paths. For example, in coronary heart disease research, HGNN can integrate gene, miRNA and clinical phenotype data, and discover the mechanism by which miR-155 affects the inflammatory response by regulating the TLR4 gene through the "gene-miRNA-phenotype" meta-pathway.

Temporal graph neural networks combined with recurrent neural networks (RNN) or Transformer architectures can simulate the dynamic evolution of network structures during the progression of cardiovascular diseases. For instance, during the pathological evolution of hypertensive patients from the compensatory stage to the decompensated stage, TGNN can capture the temporal correlations of gene expression profiles (such as ACE genes), blood pressure fluctuations, and drug dosages (such as ACE inhibitors), providing a basis for individualized medication adjustments. In addition, graph contrastive learning optimizes the initial embedding through self-supervised tasks (such as node comparison and subgraph comparison), alleviating the scarcity of cardiovascular data annotation. In the study of heart failure classification, the GCL model was pre-trained and embedded using unlabeled electronic health record data, and its performance in subsequent classification tasks was significantly better than that of traditional supervised learning.

2.2 Structural Characteristics of G-P-D Network in Cardiovascular Diseases

The G-P-D network of cardiovascular diseases presents three major characteristics: multimodality, sparsity and dynamics. Multimodal nature is reflected in the fact that node types cover DNA sequences, metabolite concentrations and chemical structures, while

edge types include gene co-expression, protein interactions and drug-target binding. For example, in the coronary heart disease-related network, there are three types of edges simultaneously: gene-gene interaction, gene-phenotypic association, and drug-gene regulation. Sparsity is manifested in the fact that only about 2% of gene pairs in the real biological network have functional associations, which makes traditional association analysis vulnerable to false positive interference. The dynamics are reflected in the systematic changes of the network structure during the disease progression, such as the reconstruction of the gene expression profile in patients with heart failure from the compensated stage to the decompensated stage.

GNN ADAPTS the above features through multi-channel embedding, sparse connection optimization and dynamic graph learning mechanisms. Multi-channel embedding allocates independent embedding spaces for different types of nodes and achieves cross-modal information fusion through shared weight layers. Sparse connection optimization employs DropEdge or attention masking techniques to suppress the interference of meaningless edges on model training. Dynamic graph learning combined with recurrent neural networks or Transformer architectures captures the time-varying patterns of network structures, such as simulating the dynamic interaction between inflammatory factors and lipid metabolism during the progression of atherosclerotic plaques.

3. Application of GNN in Gene Function Analysis of Cardiovascular Diseases

3.1 Localization and Functional Annotation of Pathogenic Genes

Although traditional genomic association analysis has identified a large number of gene loci related to cardiovascular diseases, it can only explain part of the genetic risk. The remaining "deletion heritability" may result from gene-environment interaction or epigenetic modifications. GNN can enhance the efficiency of pathogenic gene discovery by integrating multi-omics data. Its application logic includes three steps: constructing the gene-phenotype isomerism map, embedding propagation and clustering, and functional enrichment verification.

When constructing the gene-phenotype isomerism map, gene nodes are connected to clinical phenotype nodes, and the edge weights are determined by co-localization analysis or Mendelian randomization. For instance, in the study of hypertrophic cardiomyopathy, heterogeneous graphs can integrate genomic, transcriptomic and imaging data to form complex networks containing thousands of nodes and edges. In the embedding propagation stage, GNN learns the low-dimensional representation of gene nodes through multi-layer message passing, enabling function-related genes to aggregate in the embedding space. Cluster analysis can identify core gene modules, such as the MYH7 gene and its regulatory network. Functional enrichment verification compares the clustering results with the GO or KEGG pathway databases to confirm the biological significance of the module, such as finding that MYH7 causes myocardial cell contractile dysfunction by affecting the activity of calcium ion channels.

3.2 Causal Inference of Gene-phenotypic Associations

The SNP loci discovered by GWAS are mostly located in non-coding regions, and their pathogenic mechanisms need to be transmitted through intermediate phenotypes. GNN can be combined with causal discovery algorithms to construct a causal chain model of "gene - intermediate phenotype - disease". Its technical approaches include conditional independence testing, causal direction identification and counterfactual reasoning.

The conditional independence test utilizes the node representation embedded in GNN to calculate the conditional mutual information between genes and phenotypes and screen potential causal pairs. Causal direction identification determines the direct or indirect impact of genes on phenotypes through the distribution of attention weights, such as identifying the pathway by which the APOE gene affects plaque stability by up-regulating LDL receptor expression. Counterfactual reasoning simulates the phenotypic changes after gene editing to verify the robustness of the causal relationship. In the study of atherosclerosis, this model reveals the causal association between the APOE gene and plaque stability, providing a new target for targeted therapy.

4. Application of GNN in the Discovery of Phenotypic Associations in Cardiovascular Diseases

4.1 Integration and Dimensionality Reduction of Phenome Data

Cardiovascular disease phenotypes cover multi-dimensional indicators such as structure, function and biomarkers. Traditional methods such as PCA are difficult to handle nonlinear relationships, while GNN can achieve precise division of phenotypic space through nonlinear embedding. Its methodological innovations include multi-view graph convolution and self-supervised pre-training.

Multi-view convolution is used to construct independent subgraphs for different phenotypic types. For example, imaging indicators and biochemical indicators are modeled separately, and information is fused through a shared weight layer. This strategy can preserve the specificity of each modal while exploring cross-modal associations. Self-supervised pre-training utilizes contrastive learning tasks to optimize the initial embedding, such as predicting missing phenotypic values or reconstructing part of the observed data, thereby enhancing the performance of downstream classification tasks. In the phenotypic typing study of heart failure, the GNN model divided patients into two subgroups, namely "high output type" and "low output type", and the prognosis difference was significantly better than that of the traditional K-means clustering.

4.2 Association Modeling of Phenotypic and Drug Responses

There is significant heterogeneity in patients' responses to cardiovascular drugs. GNN can construct a "phenotypic - drug - gene" ternary relationship network to predict individualized medication regimens. Its modeling process includes constructing patient-drug heterogeneous graphs, meta-path analysis and response prediction.

When constructing patient-drug isomerism graphs, patient nodes connect medication records with clinical phenotypes, such as dosage, treatment course, and blood pressure changes. Meta-pathway analysis defines pathways such as "patient - medication - phenotypic change - medication adjustment", and extracts high-order association features. Response prediction predicts the efficacy or risk of adverse reactions of patients to specific drugs through graph classification tasks. In anticoagulant therapy, this model can identify patients carrying CYP2C9*3 gene variations in advance, avoiding the risk of bleeding caused by excessive warfarin.

5. Application of GNN in the Relocation and Development of Cardiovascular Drugs

5.1 Expansion of Drug-Target-Disease Networks

Traditional drug development relies on the linear paradigm of "one target - one disease", while GNN supports the systematic exploration of multi-target - multi-disease networks. Its technical implementation includes the construction of drug similarity networks, heterogeneous network alignment and path enrichment analysis.

Drug similarity networks calculate the similarity between drugs based on chemical structure or mechanism of action, for example, by constructing networks through ECFP fingerprints or target spectrum similarity. Heterogeneous network alignment aligns drug networks with disease gene networks to identify cross-indication medication opportunities. Path enrichment analysis was used to extract drug-target-gene-disease pathways to verify the biological rationality of the relocation hypothesis. In the study of beta-blockers, by analyzing their interaction with the heart failure gene network, it was found that carvedilol could improve myocardial remodeling by regulating the expression of the ADRB1 gene. This mechanism was subsequently confirmed by clinical trials.

5.2 Prediction of Synergistic Effects of Drug Combinations

Cardiovascular diseases often require combination therapy. GNN can predict the optimal combination plan by simulating the cascade effect of drugs, targets and phenotypes. Its prediction framework includes the construction of drug combination maps, collaborative score calculation and virtual screening.

When constructing a drug combination map, drug nodes are connected by sharing targets or phenotypic effects. The calculation of collaborative scoring is based on the quantification of the combination synergy of embedding similarity and path length, for

example, giving priority to drug pairs that act on the same pathway through different mechanisms. Virtual screening prioritizes the verification of the in vitro activity of high-scoring combinations, reducing experimental costs. In the treatment of hypertension, this model predicts that the combination of amlodipine and valsartan can achieve better blood pressure control by synergistically inhibiting the RAAS system and calcium channels. Subsequent clinical studies have verified that its antihypertensive effect is significantly superior to that of monotherapy.

6. Challenges and Future Directions

6.1 Current Limitations

GNN faces three major challenges in the application of cardiovascular diseases: data heterogeneity, insufficient interpretability and difficulty in dynamic modeling. Clinical phenotypic data have missing values and noise, and genomic data have batch effects. It is necessary to develop robust graph data augmentation methods. The "black box" feature of GNN limits its application in clinical decision-making, and it is necessary to combine SHAP values or attention heat maps to enhance model transparency. During the progression of cardiovascular diseases, the network structure evolves rapidly, and the existing TGNN models have limited ability to capture long-term dependencies.

6.2 Development Trends

Future research on GNN will focus on multimodal fusion, causal reinforcement learning and applications of federated learning. Multimodal fusion integrates emerging data such as single-cell sequencing and spatial transcriptomics to construct higher-resolution G-P-D networks. Causal reinforcement learning combines counterfactual reasoning and reinforcement learning to optimize dynamic treatment strategies. Federated learning enables the collaborative training of cross-institutional GNN models under the premise of protecting patient privacy, promoting large-scale clinical validation.

7. Conclusion

Graph neural networks, with their powerful capabilities in heterogeneous data integration and relationship reasoning, provide a revolutionary tool for the mining of gene-phenotypic-drug networks in cardiovascular diseases. From pathogenic gene localization to individualized medication prediction, GNNS are reshaping the research paradigm of cardiovascular precision medicine. In the future, with the accumulation of multimodal data and algorithm innovation, GNN is expected to play a core role in the entire chain of cardiovascular disease prevention, diagnosis and treatment, and ultimately achieve the clinical transformation of "gene-phenotypic-drug" linked decision-making.

References

[1] Gaidai, O., Cao, Y., & Loginov, S. (2023). Global cardiovascular diseases death rate prediction. Current problems in cardiology, 48(5), 101622.

[2] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., ... & GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American college of cardiology, 76(25), 2982-3021.

[3] Mega, J. L., Simon, T., Collet, J. P., Anderson, J. L., Antman, E. M., Bliden, K., ... & Sabatine, M. S. (2010). Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. Jama, 304(16), 1821-1830.

[4] Postmus, I., Trompet, S., Deshmukh, H. A., Barnes, M. R., Li, X., Warren, H. R., ... & Publications Committee Mathew Christopher G. 92 Blackwell Jenefer M. 80 81 Brown Matthew A. 83 Corvin Aiden 86 McCarthy Mark I. 98 Spencer Chris CA 77. (2014). Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. Nature communications, 5(1), 5068.

[5] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.

[6] Prasad, V., & Ioannidis, J. P. (2014). Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices. Implementation

Science, 9(1), 1.

[7] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI open, 1, 57-81.

[8] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

[9] Doshi, S., & Chepuri, S. P. (2022). A computational approach to drug repurposing using graph neural networks. Computers in Biology and Medicine, 150, 105992.

[10] Skarding, J., Gabrys, B., & Musial, K. (2021). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. iEEE Access, 9, 79143-79168.

