xxxx
doi.xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Identifying disease-associated circRNAs based on edge-weighted graph attention and heterogeneous graph neural network

Chengqian Lu [1,2], Lishen Zhang [1,2], Min Zeng [1,2], Wei Lan [3], and Jianxin Wang [1,2,*]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, 410083, China
[3]School of Computer, Electronic and Information, Guangxi University, Nanning, 530004, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Circular RNAs (circRNAs) with varied biological activities are implicated in pathogenic processes, according to new findings. They are regarded as promising biomarkers for the diagnosis and prognosis due to their structural features. Computational approaches, as opposed to traditional experiments, can identify the circRNA-disease connections at a lower cost. Multi-source pathogenesis data can help to reduce data sparsity and infer probable connections at the system level. The majority of available approaches create a homologous network using multi-source data, but they lose the data's heterogeneity. Effective solutions that make use of the peculiarities of multi-source data are urgently needed.
Results: In this paper, we propose a model (CDHGNN) based on edge-weighted graph attention and heterogeneous graph neural networks for discovering probable circRNA-disease correlations prediction. The circRNA network, miRNA network, disease network and heterogeneous network are constructed based on the introduced multi-source data on circRNAs, miRNAs, and diseases. The features for each type of node in the network are then extracted using a designed edge-weighted graph attention network model. Using the revised node features, we learn meta-path contextual information and use heterogeneous neural networks to assign attention weights to different types of edges. CDHGNN outperforms state-of-the-art algorithms with comparable accuracy, according to the findings of the trial. Edge-weighted graph attention networks and heterogeneous graph networks have both improved performance significantly. Furthermore, case studies suggest that CDHGNN is capable of identifying particular molecular connections and can be used to investigate pathogenic pathways.
Contact: jxwang@mail.csu.edu.cn

1 Introduction

Back-spliced from precursor mRNAs, circular RNA (circRNA) is a single-stranded endogenous non-coding RNA with covalently closed loop structures (Jeck *et al*., 2014). Emerging evidences show that circRNAs play a variety of functions, such as transcriptional regulators, microRNA sponges and protein templates (Huang *et al*., 2020). For example, exon 15 of SMARCA5 is stopped because that circSMARCA5 becomes an R-loop after binding to its parent gene locus (Xu *et al*., 2020). As a consequence, abnormal expression or dysfunction of circRNA will rise a variety of diseases. Compared with linear transcripts, circRNA is more stable without free ends that are susceptible to exonuclease digestion. Therefore, circRNA is expected to be a promising diagnostic biomarker due to its unique structure and biological roles. Nonetheless, identifying experimentally verified disease-related circRNAs takes a lot of manpower and material resources.

Computational methods provide effective means for large-scale discovery of circRNA-disease associations. Existing approaches are categorized into three types. The first type of methods are based on network propagation algorithms. The KATZ measure is used to calculate the potential association probabilities based on a heterogeneous network formed from various biological data. (Fan *et al*., 2018; Zhao *et al*., 2019; Deng *et al*., 2019). Lei *et al*. (2020) proposed a model to classify associations based on a heterogeneous network with the random walk with start. Only topological information is taken into account, limiting the model's learning ability.

The second type of methods are based on machine learning models. Yan *et al*. (2018) presented a method based on the regularized least-squares of Kronecker product kernel with disease semantic similarity and Gaussian interaction profiles. Wei *et al*. (2020) applied matrix factorization based on disease semantic information, circRNA-gene, gene-disease and circRNA-disease. Xiao *et al*. (2019) employed a weighted low-rank approximation model with dual-manifold regularization on Gaussian interaction profile kernel, disease network and circRNA network. Wei *et al*. (2021) proposed a ranking model to obtain global ranking relationships between query circRNAs and diseases. Based on low-dimensional node representation, Xiao *et al*. (2021) proposed a network embedding-based adaptive subspace learning method to infer potential associations. The appropriate features necessitate professional knowledge and have an impact on classifier performance. The third type of methods are based on neural networks. Lu *et al*. (2020) applied neural networks to replace linear approximation based on matrix factorization for possible associations. Wang *et al*. (2020) extracted features of circRNAs and disease from multi-source and classified associations with ELM classifier. Lu *et al*. (2020) employed an unsupervised model to learn *k*-mer low-dimensional vectors and disease ontology representation, as well as BiLSTM to obtain circRNA-disease connections. Wang *et al*. (2020) gained features from biological data with stacked autoencoder algorithm and predicted associations with rotation forest classifier. After fusing features of circRNA sequence and disease semantics, Wang *et al*. (2020) pre-trained the generative adversarial networks with circRNA-disease pairs and inferred prediction with the extreme learning machine classifier. Ignoring the close connection between features of circRNAs (diseases) and their associations prevents the models from learning intrinsic representation. Mudiyanselage *et al*. (2021) applied graph convolutional networks on the constructed heterogeneous network. Lan *et al*. (2021) integrated multiple relationship among circRNA, miRNA, lncRNA and disease based on graph attention network. The methods mentioned above have improved accuracy of prediction. However, the heterogeneity between different data is disregarded when applying graph neural networks on multi-source data.

In this paper, we offer a heterogeneous graph neural network model (named CDHGNN) that can learn not only the hidden properties of each type of biological molecule, but also the heterogeneity between different source data. We introduce multi-source data of circRNA, miRNA and disease to alleviate data sparsity and explore molecular associations. After that, we construct circRNA network, miRNA network, disease network and heterogeneous network. We devise an edge-weighted graph attention network to get node representation, in contrast to prior graph neural network methods that disregard edge weights. Furthermore, we adopt heterogeneous transformer network to learn the contextual information of the meta-path and get attention weights for different types of edges. The experimental results reveal that CDHGNN outpaces state-of-the-art computational methods. Edge-weighted graph attention network and heterogeneous graph network improve accuracy. It is worth noting that CDHGNN finds molecular connections and the relevant pathways in pathogenesis.

The contributions of this work are summarised as follows:

- To view molecular associations in pathogenesis, we integrate multi-source data of circRNA, miRNA and disease and construct corresponding biological networks.
- We devise an edge-weighted graph attention neural network that considers the value of associations when obtaining node intrinsic properties.
- We first study the heterogeneity of multi-source data. CDHGNN learns the meta-contextual path's information and assigns attention weights to various types of edges based on a heterogeneous neural network.
- The experimental results and case studies show that CDHGNN outperforms state-of-the-art methods and is helpful to explore relevant pathways in pathogenesis.

2 Materials and Methods

In this part, we present our model CDHGCN, which is applied to identify the potential circRNA-disease associations. As shown in Fig. 1, we first construct circRNA network, miRNA network, disease network and heterogeneous network. Then, we extract node features with a devised edge-weighted graph attention network. Finally, we learn contextual information and assign attention weights on the meta-path based on a heterogeneous neural network.

2.1 Benchmark Dataset

To alleviate data sparsity and understand potential associations systematically, we collect circRNA sequences, miRNA functional relationships, disease ontology, gene-disease network, circRNA-miRNA, miRNA-disease and circRNA-disease associations from benchmark database. We retrieve circRNA sequences data from CircBase (Glazar *et al*., 2014) and get 140,732 circRNA sequences. We download miRNA functional relationships from database MISIM v2.0 (Li *et al*., 2019), which includes functional relationships among 664 miRNAs. We obtain disease ontology from Disease Ontology (Schriml *et al*., 2019), which contains 11,652 phenotype ontology. The UMLS Metathesaurus Browser, a vast biomedical thesaurus, is used to get disease definitions. We get gene-disease associations from (Pinero *et al*., 2020), which contains 262,989 associations between 13,705 genes and 1,977 diseases. We gather circRNA-miRNA associations from starBase (Li *et al*., 2014), which contains 18,320 circRNA-miRNA associations between 886 circRNAs and 638 miRNAs. We get miRNA-disease associations from HMDD 3.0 (Huang *et al*., 2019), which includes 27,872 miRNA-disease associations between 1,054 miRNAs and 226 diseases. We download circRNA-disease associations from MNDR 3.0 (Ning *et al*., 2021), which includes 3,206 circRNA-disease associations between 2,396 circRNAs and 165 diseases. We have unified and standardized the nomenclature of circRNA according to circBase. The unified and standardized operation is executed for the nomenclature of disease according to OMIM and UMLS. After that, we delete duplicate data of species other than human species. Finally, we get 2,013 associations between 1,313 circRNAs and 144 diseases, 10,570 associations between 638 miRNAs and 128 diseases, and 13,315 associations between 824 circRNAs and 612 miRNAs. The details of the data are depicted in Table 1.

Table 1. The details of the data

Dataset	Num
Disease ontolgy	11,652
Disease definition	152
CircRNA sequences	140,732
MiRNA-disease	10,570 (638 miRNAs, 128 diseases)
CircRNA-miRNA	13,315 (824 circRNAs, 612 miRNAs)
CircRNA-disease	2,013 (1313 circRNAs, 144 diseases)

2.2 Network construction

After importing multi-source data on the pathogenesis of circRNAs, heterogeneous biological network is an effective method to find the promising associations between circRNAs and diseases. There are three types of nodes like circRNA, miRNA and disease. To obtain features of each type of

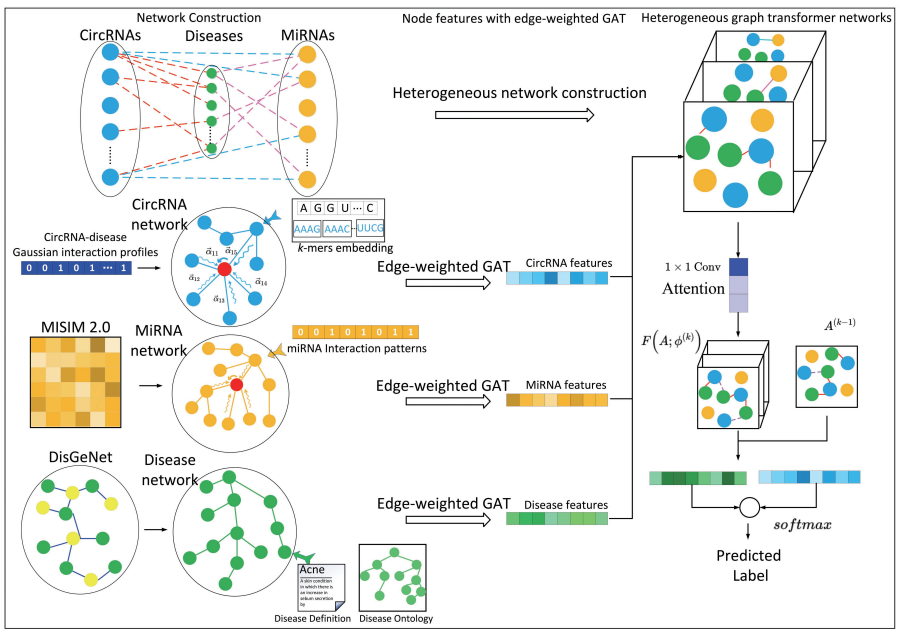

Fig. 1. The overflow of CDHGNN. Step 1: construct circRNA-circRNA network, miRNA-miRNA network, disease-disease network and heterogeneous network; Step2 : extract each type of node features with edge-weighted graph attention network; Step3: Learn contextual information and assign attention weights on the meta-path in the heterogeneous network.

node in the respective biological network, CDHGNN establishes circRNA network, miRNA network and disease network, respectively.

The Gaussian interaction profiles (GIPs) embody association patterns of circRNA or disease (Lu *et al*., 2018). We get the association matrix between circRNAs and diseases $A_{cd} \in \mathbb{R}^{nc \times nd}$, where *nc* is the number of circRNAs while *nd* is the number of diseases, respectively. We calculate the Gaussian interaction profile kernel similarity as the similarity between two circRNAs:

$$Sim(c_i, c_j) = exp(-\varphi_c \|GIPs(c_i) - GIPs(c_j)\|^2),$$
$$\varphi_c = \frac{1}{nc} \sum_{i=1}^{nc} \|GIPs(c_i)\|^2, \tag{1}$$

where φ_c controls the kernel bandwidth.

To build a miRNA functional similarity network, we download functional relationships of miRNA from MISIM v2.0. To make the values have the same scale, we normalized the data with Z-score normalization and use it as the functional similarity among miRNAs.

To construct disease network, we download gene-disease data from DisGeNET. We get the association probability between genes and diseases. Then, we calculate disease similarity between disease *i* and disease *j* as follows:

$$Sim(d_i, d_j) = \frac{\sum_{v \in (G_{d_i} \cap G_{d_j})} (\boldsymbol{P}(v, d_i) + \boldsymbol{P}(v, d_j))}{\sum_{v \in G_{d_i}} \boldsymbol{P}(v, d_i) + \sum_{u \in G_{d_j}} \boldsymbol{P}(u, d_j)} \tag{2}$$

where G_{d_i} is the set of genes related with disease *i*, G_{d_j} is the set of genes related with disease *j*, where $\boldsymbol{P}(\cdot, \cdot)$ is the association probability matrix between genes and diseases.

There are three type of edges among circRNAs, miRNAs and disease, such as circRNA-miRNA, circRNA-disease and miRNA-disease. A heterogeneous network is defined as $G = (V, E)$, where V and E signify the sets of nodes and edges, respectively. The heterogenous network is accompanied by a node type mapping function $\psi_v : V \to \mathcal{T}^v$ and an edge type mapping function $\psi_e : E \to \mathcal{T}^e$. A node type mapping $\psi_v(v_i)$ uniquely corresponds to a node v_i, i.e., $\psi_v(v_i) \in \mathcal{T}^v, v_i \in V$. Analogously, an edge type mapping $\psi_e(e_i)$ uniquely corresponds to an edge e_i, i.e., $\psi_e(e_i) \in \mathcal{T}^e, e_i \in E$. The situation is $|\mathcal{T}_e| > 1$ in the heterogeneous network. We define the heterogeneous network with a set of adjacency matrices $\{A_i\}_{i=1}^{|\mathcal{T}_e|}$, where $A_i \in \mathbb{R}^{N \times N}$, $N = nc + nm + nd$. A node feature matrix $X \in \mathbb{R}^{N \times F}$ is the input vector to the heterogeneous network, where F denotes the learned features for N nodes from GAT. A meta-path $\mathcal{P}$ is a path through the heterogeneous network, i.e., $v_1 \xrightarrow{t_1} v_2 \xrightarrow{t_2} \ldots \xrightarrow{t_i} v_{i+1}$, where $t_i \in \mathcal{T}^e$. Then the the adjacency matrix $A_{\mathcal{P}}$ of the meta-path $\mathcal{P}$ is defined as $A_{\mathcal{P}} = A_{t_i} \cdots A_{t_2} A_{t_1}$, where A_{t_i} is an adjacency matrix for the *i*-th type of edges on the meta-path.

2.3 Node embeddings with edge-weighted graph attention networks

We extract features of various types of nodes from the three constructed similarity networks. As each molecule functions differently in biological systems, we apply graph attention network (GATs) to specify different weights to different nodes in a neighborhood (Velickovic *et al*., 2018). To account for the effect of edge weights on aggregation, we devise a novel edge-weighted GATs. Input a set of initial node features, $\mathbf{x} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}, \vec{x}_i \in \mathbb{R}^f$, where *n* means the number of nodes, and *f* is the feature dimension of each node. Transform the initial features into high-level features with a linear operation $h_i^{(l)} = W^{(l)} x_i^{(l)}$, where W is a shared weight matrix and $W \in \mathbb{R}^{f' \times f}$. The generated features of *l*-layer is $h_i^{(l)} = \{\vec{x}'_1, \vec{x}'_2, \ldots, \vec{x}'_n\}, \vec{x}'_i \in \mathbb{R}^{f'}$. The edge weight between node *i* and node *j* reflects the importance on each connection. Concatenate $h_i^{(l)}$, $h_j^{(l)}$ and edge weight ω_{ij} between node *i* and node *j* in *l* l-th layer and calculate a un-normalized pair-wise attention score as:

$$e_{ij}^{(l)} = \text{LeakyReLU}\left(\vec{a}^{(l)^T}\left(h_i^{(l)} \| h_j^{(l)} \| W_e \omega_{ij}\right)\right), \tag{3}$$

where $\vec{a}^{(l)}$ is a learnable weight vector, $a : \mathbb{R}^{f'} \times \mathbb{R}^{f'} \to \mathbb{R}$. Masked attention is applied to insert graph structure into the model architecture.

Normalizing coefficients across all choices of j with the softmax function:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(e_{ik}^{(l)}\right)}. \tag{4}$$

Multi-head attention aggregation is adopted to aggregate the embeddings of neighbors together simultaneously

$$z_i^{(l+1)} = \|_{m=1}^{M}\sigma\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{m}W^{m}h_j^{(l)}\right) \tag{5}$$

For the initial node features in circRNA network, CDHGNN utilizes Doc2Vec algorithm (Le *et al*., 2014) to get sequence motif embedding on the seuqnce. For the initial node features in miRNA network, miRNAs' GIPs, $f_i^m = \{0, 1, \ldots, 1\}$, $f_i^m \in \mathbb{R}^{nd}$, are employed as association vectors. For disease, we concatenate disease semantic similarity based on disease ontology and disease semantic embedding from disease definitions with Doc2Vec.

2.4 Heterogeneous graph transformer networks

Heterogeneous networks composed of multiple types of data are models successfully applied in various applications. However, most GNN models treat this type of complex network including multiple types of data as a homogeneous network, or need to manually specify meta-paths in advance. These methods may cause loss of information. In this part, we applied graph transformer networks (GTN) to learn soft selections of edge types and get hidden relationships among circRNAs, miRNAs and diseases (Yun *et al*., 2019). The input to GTN is multiple networks with diverse types of nodes and edges. The 1×1 convolution layers assigns a weight to each type of edge in $\{A_i\}_{i=1}^{|\mathcal{T}_e|}$, and then $\mathrm{softmax}(W_\phi)$ utilizes attention mechanism to determine the influence on the final meta-path. A soft adjacency matrix is generated from the weighted sum defined as:

$$Q = F(A; W_\phi) = \phi(A; \mathrm{softmax}(W_\phi)), \tag{6}$$

where ϕ is the convolution layer and $W_\phi \in \mathbb{R}^{1\times1\times|\mathcal{T}_e|}$. Then, a composition meta-path is established by multiplying the adjacency matrices. To consider the characteristics of the original edges, we introduce the identity matrix I. Applying GCN for each type of edge on the meta-path, node representations are shaped as:

$$Z = \mathop{\Big\|}_{i=1}^{C}\sigma\left(\tilde{D}_i^{-1}\tilde{A}_i^{(l)}XW_{gt}\right) \tag{7}$$

where $\|$ denotes the concatenate operator, C is the number of channels, $\tilde{A}_i^{(l)} = A_i^{(l)} + I$, $\tilde{D}_i$ is the degree matrix of $\tilde{A}_i^l$, and $W_{gt} \in \mathbb{R}^{F\times F}$ is a shared trainable weight matrix. We utilize the cross-entropy loss function to measure the performance and the Adam algorithm to optimize the model.

3 Results and Discussion

3.1 Evaluation metrics

We implement 5-fold cross-validation to verify the effectiveness of the model and compare it with state-of-the-art methods. All the verified associations among circRNA, miRNA, and disease are treated as positive samples, while candidate samples are potential relationships that have not been validated. To alleviate the imbalance, we stochastically generate negative samples with the same number of positive samples. Then all the samples are randomly divided into 5 parts. One of them is treated as a separate test set that does not participate in the model's training. The remaining 4 parts take part in the training. We repeat the whole process five times until each part is tested once. THe general evaluation criteria are utilized, including Accuracy (ACC), Precision (Pre), Recall and F1-score defined as follows:

$$\begin{aligned}
\text{Pre.} &= \frac{TP}{TP+FP}\\
\text{Recall} &= \frac{TP}{TP+FN}\\
\text{Acc.} &= \frac{TP+TN}{TP+FP+FN+TN}\\
\text{F1}-\text{score} &= \frac{2*\text{ Precision }*\text{ Recall}}{\text{Precision }+\text{ Recall}}
\end{aligned} \tag{8}$$

where TP and TN are the numbers of correctly identified positive and negative samples, FP and FN are the number of incorrectly identified positive and negative samples, respectively.

Table 2. The impact of learning rate on model performance

Learning rate	AUC	AUPR	Acc	Pre	Recall	F1-Score
0.001	0.851	0.792	0.817	0.799	0.786	0.792
0.003	0.878	0.801	0.822	0.782	0.792	0.803
0. 005	**0.886**	**0.817**	**0.824**	**0.808**	**0.817**	**0.804**
0.007	0.865	0.795	0.811	0.801	0.807	0.796
0.01	0.821	0.792	0.798	0.800	0.803	0.768
0.03	0.791	0.724	0.771	0.657	0.694	0.674

Table 3. The impact of node dimensions on model performance

Node dim	AUC	AUPR	Acc	Pre	Recall	F1-Score
32	0.873	0.775	0.787	0.761	0.784	0.759
64	0.879	0.798	0.813	0.795	0.802	0.792
128	**0.886**	**0.817**	**0.824**	**0.808**	**0.817**	**0.804**
160	0.882	0.796	0.814	0.797	0.808	0.788
192	0.865	0.761	0.776	0.758	0.765	0.762

3.2 Effects of Parameters

In this part, we evaluate the impact of several key parameters on the model, including learning rate and node dimension. The learning rate is a critical parameter for the minimum of the loss function. We execute the grid search to obtain the optimal value from 0.001 to 0.3. From Table 2, it shows that CDHGNN performs the best with a value of 0.005. As the parameter increases from 0.001 to 0.005, the accuracy of the model is improving. But as the parameter continues to increase from 0.005, the performance of the model has declined. A large learning rate will cause the model to converge to a sub-optimal solution. The default value of the learning rate is set to 0.005.

The dimension of concatenated node features is a key factor for model performance. We conduct the cross-validation to assess the impact of node dimensions (Table 3). It indicates that the performance of the model becomes better when the dimensionality of the node increases from 32 to 128. The model achieves best performance when the node dimension is 128. However, as the node dimension continues to increase, the performance of the model decreases. The default value of node dimension is set to 128.

3.3 Changes of attention scores

We use attention mechanism to determine the influence of each type of edge. During the training, changes in the weights of each type of edge reflect the importance of the pathogenic process (Fig. 2). CM stands for

circRNA-miRNA matrix. CD stands for circRNA-disease Matrix. MD stands for miRNA-disease Matrix. I is an abbreviation for identity matrix. The CD matrix has the highest allocated weight, implying that it has the greatest influence on the final prediction accuracy. CM and MD are also important for accuracy. I has minimal impact on accuracy. The conclusion is consistent with the actual biological situation.

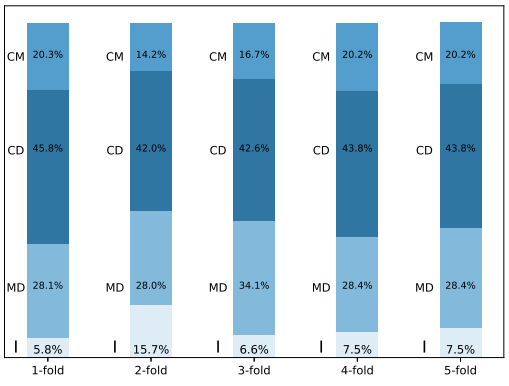

Fig. 2. Changes of attention values during 5-fold cross-validation

3.4 Comparison with other methods

We contrast CDHGNN with five state-of-the-art methods, including KGANCDA (Lan *et al*., 2021), MGRCDA (Wang *et al*., 2021), CDASOR (Lu *et al*., 2020), GCNCDA (Wang *et al*., 2020) and NSL2CD (Xiao *et al*., 2021). We also evaluate the effects of edge-weighted graph attention network and heterogeneous graph transformer network separately. The first model composed of un-weighted GAT and heterogeneous GTN is marked as CDHGNN_{u}. The second model composed of edge-weighted GAT and homogeneous neural network is marked as CDHGNN_{h}. The detailed comparison results are shown in Table 4. We can see that CDHGNN achieves the best performance(AUC:0.886, AUPR:0.817, Accuracy:0.824, Precision:0.808, Recall:0.814, F1-score:0.804). The edge-weighted GAT improves the performance (AUC:1%, AUPR:2%, Acc:2%, Pre:1%, Recall:1%, F1-score:0.7%). The heterogeneous neural networks improves even more (AUC:1.9%, AUPR:4%, Acc:4%, Pre:3.3%, Recall:1.6%, F1-score:0.9%). The detailed AUC results are shown in Fig. 3. The auc value of CDHGNN is much higher than that of other methods, indicating that the accuracy of CDHGNN is higher. In addition, we compare the percentage of correctly retrieved associations from top 10 to top 40 predictions (Fig. 4). All the results show that not only edge-weighted GAT effectively improve the model, but also heterogeneous graph neural networks make the model more accurate. Compared with edge-weighted GAT, the utilization of a heterogeneous graph neural network improves the accuracy more significantly. It indicates that the introduction of multi-source biological information can enhance prediction.

3.5 Case studies

Case studies are conducted to assess the prediction capacity with existing literature and public databases. Sort predicted relationships in descending order after training the model with all the empirically verified circRNA-disease associations. Table 5 shows 14 of the top 20 predicted connections were verified. Originated from vacuolar ATPse assembly factor, hsa_circ_0091702 mitigates sepsis-correlated acute kidney injury by regulating miR-9-3p/SMG1/inflammation and oxidative stress (Shi *et al*., 2020). Circ-AKT3 (hsa_circ_0000199) is related to acute kidney injury via miR-144-5p/Wnt/β-catenin pathway (Xu *et al*., 2020).

Table 4. The performance comparison with state-of-the-art methods.

Methods	AUC	AUPR	Acc	Pre	Recall	F1-Score
CDHGNN	**0.886**	**0.817**	**0.824**	**0.808**	**0.814**	**0.804**
CDHGNN_{u}	0.875	0.797	0.804	0.798	0.802	0.797
CDHGNN_{h}	0.856	0.757	0.764	0.765	0.786	0.788
KGANCDA	0.841	0.615	0.646	0.674	0.693	0.709
MGRCDA	0.845	0.714	0.751	0.767	0.785	0.789
CDASOR	0.814	0.725	0.756	0.742	0.703	0.728
GCNCDA	0.788	0.737	0.713	0.714	0.726	0.786
NSL2CD	0.803	0.704	0.735	0.657	0.696	0.713

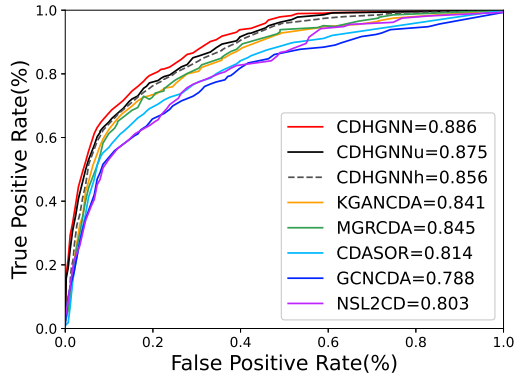

Fig. 3. The performance comparison with state-of-the-art methods in terms of AUC values.

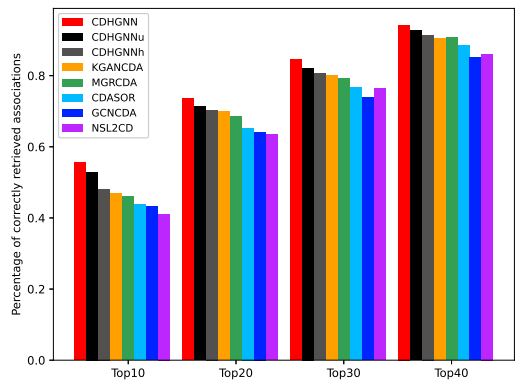

Fig. 4. Percentage of correctly retrieved associations

CircFUT8 (hsa_circ_0003028) sponges miR-570-3p and regulate the miR-570-3p/KLF10 axis as a tumor suppressor in bladder cancer (He *et al*., 2020). Binding with miR-296-5p, hsa_circ_0000515 activates the cell growth of bladder cancer (Cai *et al*., 2020). Fastening miR-200a-3p, exosomes-mediated transfer of circ_UBE2D2 (hsa_circ_0005728) enhances tamoxifen resistance in breast cancer (Hu *et al*., 2020). Sponging miR-532-3p, circRNA_103809 (hsa_circ_0072088) represses cell proliferation and metastasis of breast cancer (Liu *et al*., 2020). Participating in the miR-135a-5p/EMT axis, circRNA_0001946 (hsa_circ_0001946) functions as a tumor promoter (Zeng *et al*., 2020). Inhibiting colorectal cancer cell proliferation by its knockdown and regulating miR-296-5p/RUNX1 axis, circ_0000512 (hsa_circ_0000512) is a promising therapeutic target for colorectal cancer (Wang *et al*., 2020). Circ-RanGAP1 (hsa_circ_0063526)

promote gastric cell progression by mediating miR-877-3p/VEGFA axis (Lu *et al*., 2020). In gastric cancer, hsa_circ_0004872 related to a negative regulatory loop hsa_circ_0004872/miR-224/Smad44/ADAR1 functions as a tumor suppressor (Ma *et al*., 2020). Acting as endogenous RNA for miR-654-3p, circRHOBTB3 (hsa_circ_0006404) inhibits cell growth of gastric cancer by promoting p21 signaling pathway (Ma *et al*., 2020). Insulating miR-17, circ-ITCH (hsa_circ_0001141) functions as a tumor-suppressor factor by the Wnt/β-catenin signaling pathway in gastric cancer (Peng *et al*., 2020). Sponging miR-134-5p which activates BTG-2 expression, circZNF609 (hsa_circ_0000615) represses proliferation and migration of glioma cell (Tong *et al*., 2020). Targeting the miR-520a-5p/CDK4 regulatory axis, exosome-transmitted hsa_circ_0014235 activates malignant development of non-small cell lung cancer (Xu *et al*., 2020). The details of validated pathways for the top-20 predicted circRNA-disease associations is shown in Fig. 5. MiRNAs linked to circRNAs and involved in diseases have been discovered, which could aid in the discovery of new pathways.

Table 5. Validation of top-20 rank predicted associations.

Diseases	Top-ranked circRNAs			
	CircRNAs	miRNAs	Rank	Evidences
Acute kidney injury	hsa_circ_0091702	miR-9-3p	9	PMID: 32827242
	hsa_circ_0000199	miR-144-5p	17	PMID: 33200535
Bladder cancer	hsa_circ_0003028	miR-570-3p	19	PMID: 32072011
Breast cancer	hsa_circ_0000515	miR-296-5p	2	PMID: 32446265
	hsa_circ_0005728	miR-200a-3p	5	PMID: 32756532
	hsa_circ_0072088	miR-532-3p	11	PMID: 32499818
Colorectal cancer	hsa_circ_0001946	miR-135a-5p	4	PMID: 32508871
	hsa_circ_0000512	miR-296-5p	14	PMID: 32821119
Gastric cancer	hsa_circ_0063526	miR-877-3p	1	PMID: 31811909
	hsa_circ_0004872	miR-224	3	PMID: 33172486
	hsa_circ_0006404	miR-654-3p	7	PMID: 31928527
	hsa_circ_0001141	miR-17	16	PMID: 33060778
Glioma	hsa_circ_0000615	miR-134-5p	12	PMID: 31721211
Non-small lung cancer	hsa_circ_0014235	miR-520a-5p	15	PMID: 33292236

4 Conclusion

CircRNAs and illnesses are linked by their range of biological roles. Meanwhile, the special structure makes it a promising biomarker for the treatment of diseases. However, existing methods lack consideration of multi-source data heterogeneity. In this work, we devise a model to predict possible circRNA-disease associations based on a heterogeneous graph neural network. We created a unique edge-weighted graph attention network to grasp node features since edge weights convey the relevance of associations between nodes. We adopt an attention mechanism to learn contextual information and assign attention scores on the meta-path to infer potential associations. The experimental results reveal that CDHGNN outperforms state-of-the-art methods with comparable accuracy. It is worth noting that CDHGNN can find molecular connections and the relevant pathways in pathogenesis.

Although CDHGNN performs admirably in terms of predicting probable associations, it still has several flaws that need be investigated further. For example, the pathogenic process of disease is a very complex molecular activity process. CDHGNN uses biological information among circRNA, miRNA and disease. It would be preferable if additional relevant biomolecular data were integrated to train the model. Furthermore, the molecular structure, which provides unique information about biomolecules, may be enhanced by adding more accurate features to the prediction model.

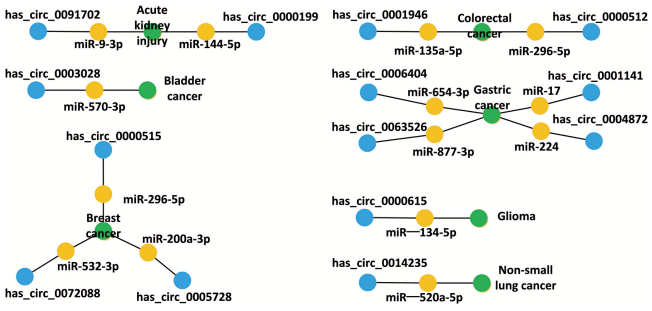

Fig. 5. The validated pathways for the top-20 predicted circRNA-disease associations.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61972423, No. 62002390), the 111 Project (No. B18059) and the Hunan Provincial Science and Technology Program (No. 2018wk4001).

Conflict of Interest: none declared.

References

Cai, F., Fu, W., Tang, L., et al. Hsa_circ_0000515 is a novel circular RNA implicated in the development of breast cancer through its regulation of the microRNA-296-5p/CXCL10 axis. *FEBS J*. 2021;288(3): 861-883.

Deng, G., Mou, T., He, J., et al. Circular RNA circRHOBTB3 acts as a sponge for miR-654-3p inhibiting gastric cancer growth. *J. Exp. Clin. Canc. Res*. 2020;39(1): 1-16.

Deng L., Zhang W., Shi Y., *et al*. Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci. Rep.*. 2019;9(1): 1-10.

Deng Z., Li X., Wang H., et al. Dysregulation of CircRNA_0001946 contributes to the proliferation and metastasis of colorectal cancer cells by targeting MicroRNA-135a-5p. *Front. Genet*. 2020;11: 357-369.

Fan, C., Lei, X., Wu, F. X., *et al*. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. *Int. J. Biol*. 2018;14(14): 1950.

Glazar P., Papavasileiou P., and Rajewsky N. CircBase: a database for circular RNAs. *RNA* 2014;**20**: 1666-1670.

He Q., Yan D., Dong W., et al. circRNA circFUT8 upregulates krüppple-like factor 10 to inhibit the metastasis of bladder cancer via sponging miR-570-3p. *Mol. Ther*. 2020;16: 172-187.

Hu, K., Liu, X., Li, Y., et al. Exosomes Mediated Transfer of Circ_UBE2D2 Enhances the Resistance of Breast Cancer to Tamoxifen by Binding to MiR-200a-3p. *Med. Sci. Monit*. 2020;26: e922253-1–e922253-12.

Huang, Z., Shi, J., Gao, Y., *et al*. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res*. 2019;47(D1): D1013-D1017.

Huang A., Zheng H., Wu Z., *et al*. Circular RNA-protein interactions: functions, mechanisms, and identification. *Theranostics* 2020;10(8): 3503.

Jeck, W.R. and Sharpless, N.E. Detecting and characterizing circular RNAs. *Nat. Biotechnol*. 2014;32(5): 453-461.

Lan, W., Dong, Y., Chen, Q., *et al*. KGANCDA: predicting circRNA-disease associations based on knowledge graph attention network. *Brief. Bioinform*. 2021; doi: doi.org/10.1093/bib/bbab494.

Le, Q., Mikolov, T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). Beijing, China: JMLR, 2014, 1188-1196.

Lei X., Bian C. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association. *Sci. Rep*. 2020;10: 1940.

Li, J. H., Liu, S., Zhou, H., *et al*. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014;42(D1): D92-D97.

Li, J., Zhang, S., Wan, Y., *et al*. MISIM v2. 0: a web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res*. 2019;47(W1): W536-W541.

Liu, M., Luo, C., Dong, J., et al. circRNA_103809 suppresses the proliferation and metastasis of breast cancer cells by sponging microRNA-532-3p (miR-532-3p). *Front. Genet*. 2020;11: 485.

Lu, C., Yang, M., Luo, F., *et al*. Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics*. 2018;34(19): 3357-3364.

Lu, C., Yang, M., Li, M., *et al*. Predicting human lncRNA-disease associations based on geometric matrix completion. *IEEE J. Biomed. Health Inform*. 2019;24(8): 2420-2429.

Lu, C., Zeng, M., Zhang, F., *et al*. Deep matrix factorization improves prediction of human circRNA-disease associations. *IEEE J. Biomed. Health Inform*. 2020;25(3): 891-899.

Lu, C., Zeng, M., Wu, F. X., *et al*. Improving circRNA–disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. *Bioinformatics*. 2020;36(24): 5656-5664.

Lu J., Wang Y., Yoon C., et al. Circular RNA circ-RanGAP1 regulates VEGFA expression by targeting miR-877–3p to facilitate gastric cancer invasion and metastasis. *Cancer Lett*. 2020;471: 38-48.

Ma C., Wang X., Yang F., et al. Circular RNA hsa_circ_0004872 inhibits gastric cancer progression via the miR-224/Smad4/ADAR1 successive regulatory circuit. *Mol. Cancer*. 2020;19(1): 1-21.

Mudiyanselage, T. B., Lei, X., Senanayake, N., *et al*. Predicting CircRNA Disease Associations using Novel Node Classification and Link Prediction Models on Graph Convolutional Networks. *Methods*. 2021; doi: doi.org/10.1016/j.ymeth.2021.10.008.

Ning L., Zheng B., Wang N., *et al*. MNDR v3.0: mammal ncRNA–disease repository with increased coverage and annotation. *Nucleic Acids Res*. 2021;49(D1): D160-D164.

Peng Y., Wang H. H. Cir-ITCH inhibits gastric cancer migration, invasion and proliferation by regulating the Wnt/β-catenin pathway. *Sci. Rep*. 2020;10(1): 1-13.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., *et al*. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1): D845-D855.

Schriml L., Mitraka E., Munro J., *et al*. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019;47(D1): D955-D962.

Shi Y., Sun C. F., Ge W. H., et al. Circular RNA VMA21 ameliorates sepsis-associated acute kidney injury by regulating miR-9-3p/SMG1/inflammation axis and oxidative stress. *J. Cell Mol. Med*. 2020;24(19): 11397-11408.

Tong H., Zhao K., Wang J., et al. CircZNF609/miR-134-5p/BTG-2 axis regulates proliferation and migration of glioma cell. *J. Pharm. Pharmacol* . 2020;72(1): 68-75.

Veličković, P., Cucurull, G., Casanova, A., *et al*. Graph attention networks. In: *International Conference on Learning Representations*. Vancouver, Canada, 2018.

Wang, L., You, Z. H., Huang, Y.A., *et al*. An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network. *Bioinformatics*. 2020;36(13): 4038-4046.

Wang L., You Z. H., Li J. Q., et al. IMS-CDA: prediction of CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model. *IEEE Trans. Cybern*. 2020;51(11): 5522-5531.

Wang L., Yan X., You Z. H., et al. SGANRDA: semi-supervised generative adversarial networks for predicting circRNA–disease associations[J]. *Brief. Bioinform*. 2021;22(5): 1-11.

Wang L., You Z. H., Huang D. S., *et al*. MGRCDA: Metagraph Recommendation Method for Predicting CircRNA-Disease Association. *IEEE Trans. Cybern*. 2021, 10.1109/TCYB.2021.3090756.

Wang L., You Z., Li Y., et al. GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm. *PLoS Comput. Biol.* 2020;16(5): e1007568.

Wang L., Wu H., Chu F., et al. Knockdown of circ_0000512 inhibits cell proliferation and promotes apoptosis in colorectal cancer by regulating miR-296-5p/RUNX1 axis. *Onco. Targets Ther*. 2020;13: 7357-7368.

Wei H. and Liu B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform*. 2020;21(4): 1356-1367.

Wei, H., Xu, Y., Liu, B. iCircDA-LTR: identification of circRNA-disease associations based on Learning to Rank. *Bioinformatics*. 2021;37(19): 3302-3310.

Xiao, Q., Jiawei L., and Jianhua D. Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework. *IEEE J. Biomed. Health Inform*. 2019;23(6): 2661-2669.

Xiao, Q., Fu, Y., Yang, Y., *et al*. (2021). NSL2CD: identifying potential circRNA–disease associations based on network embedding and subspace learning. *Brief. Bioinform*. 2021;22(6): 1-11.

Xu, X., Zhang, J., Tian, Y., *et al*. CircRNA inhibits DNA damage repair by interacting with host gene. *Mol. Cancer*. 2020;19(1): 1-19.

Xu, X., Tao, R., Sun, L., et al. Exosome-transferred hsa_circ_0014235 promotes DDP chemoresistance and deteriorates the development of non-small cell lung cancer by mediating the miR-520a-5p/CDK4 pathway. *Cancer Cell Int*. 2020;20(1), 1-15.

Xu Y., Jiang W., Zhong L., et al. circ-AKT3 aggravates renal ischaemia-reperfusion injury via regulating miR-144-5p/Wnt/β-catenin pathway and oxidative stress. *J. Cell Mol. Med*. 2020; DOI: 10.1111/jcmm.16072.

Yan C., Wang J., and Wu F.-X. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinform.* 2018;19: 73-81.

Yun, S., Jeong, M., Kim, R., *et al*. Graph transformer networks. In: Proceedings of the 33rd Advances in Neural Information Processing Systems (Neurips 2019), Vancouver, Canada. 2019;32: 11983-11993.

Zhao Q., Yang Y., Ren G., *et al*. Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans. Nanobiosci*. 2019;18(4): 578-584.