Briefings in Bioinformatics, 2024, **25(6)**, bbae573
https://doi.org/10.1093/bib/bbae573
Problem Solving Protocol

A multi-task prediction method based on neighborhood structure embedding and signed graph representation learning to infer the relationship between circRNA, miRNA, and cancer

Lan Huang*, Xin-Fei Wang, Yan Wang*, Ren-Chu Guan, Nan Sheng, Xu-Ping Xie, Lei Wang, Zi-qi Zhao

Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, No. 2699, Qianjin Street, Changchun 130012, China

*Corresponding authors. Lan Huang, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: huanglan@jlu.edu.cn: @lanhuang; Yan Wang, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: wy6868@jlu.edu.cn: @yanwang

Abstract

Motivation: Research shows that competing endogenous RNA is widely involved in gene regulation in cells, and identifying the association between circular RNA (circRNA), microRNA (miRNA), and cancer can provide new hope for disease diagnosis, treatment, and prognosis. However, affected by reductionism, previous studies regarded the prediction of circRNA-miRNA interaction, circRNA-cancer association, and miRNA-cancer association as separate studies. Currently, few models are capable of simultaneously predicting these three associations.
Results: Inspired by holism, we propose a multi-task prediction method based on neighborhood structure embedding and signed graph representation learning, CMCSG, to infer the relationship between circRNA, miRNA, and cancer. Our method aims to extract feature descriptors of all molecules from the circRNA-miRNA-cancer regulatory network using known types of association information to predict unknown types of molecular associations. Specifically, we first constructed the circRNA-miRNA-cancer association network (CMCN), which is constructed based on the experimentally verified biomedical entity regulatory network; next, we combine topological structure embedding methods to extract feature representations in CMCN from local and global perspectives, and use denoising autoencoder for enhancement; then, combined with balance theory and state theory, molecular features are extracted from the point of social relations through the propagation and aggregation of signed graph attention network; finally, the GBDT classifier is used to predict the association of molecules. The results show that CMCSG can effectively predict the relationship between circRNA, miRNA, and cancer. Additionally, the case studies also demonstrate that CMCSG is capable of accurately identifying biomarkers across various types of cancer. The data and source code can be found at https://github.com/1axin/CMCSG.

Keywords: competing endogenous RNA; miRNA-disease association; circRNA-disease association; miRNA-circRNA interaction; molecular associations network

Introduction

A variety of complex human diseases, especially cancer, are related to abnormal transcriptional and post-transcriptional gene expression. Large-scale transcriptome analysis of human cells shows that not all genomes are transcribed into proteins, and some of them are involved in gene regulation as non-coding RNA (ncRNA) [1]. In the past decade, a large number of studies have shown that ncRNA can control the expression of mRNA through target genes, leading to tumor production, proliferation, and metastasis [2, 3]. This discovery provides a new opportunity for the interpretation of tumor biology and the treatment of complex diseases.

As a small ncRNA (<200 nucleotides), microRNA (miRNA) is a typical representative of ncRNA that regulates gene expression [4]. Through miRNA response elements (MREs), miRNA can target mRNA; and a variety of endogenous RNA can also target miRNA through MRE [5]. This competitive binding enables endogenous RNA to regulate each other's expression, which is called the competing endogenous RNAs (ceRNAs) hypothesis [6]. The ceRNA hypothesis was first confirmed in *Arabidopsis thaliana* [7]. With the development of research, ceRNA has been proven to play an important role in diseases such as cancer [8, 9]. Since the dysregulation of miRNA in cancer was first reported in 2002 [10], miRNA has been studied rapidly as a marker for the diagnosis and treatment of cancer. In 2013, the miRNA mimic for cancer treatment entered the clinical stage for the first time [11]. In addition, the detection of miRNA in biological fluids has become an important means of monitoring cancer [12]. According to ceRNA

Received: July 2, 2024. **Revised:** September 9, 2024. **Accepted:** October 30, 2024

theory, endogenous RNA with MRE can be used as ceRNA, including protein-coding transcripts, circular RNA (circRNA), and long non-coding RNA (lncRNA). At present, more than 7000 human circRNA have been identified by high-throughput technology [13]. Because of its unique ring structure and high exonuclease resistance, circRNA has great potential as ceRNA. Although the mechanism of circRNA is not perfect, it may still have high functionality in cancer.

To economically and efficiently identify potential circRNA biomarkers for diseases and to gain a deeper understanding of the disease mechanisms, computational prediction models related to circRNA have gained significant attention and development. Initially, models for predicting miRNA-disease associations were prioritized. Researchers have proposed numerous advanced prediction models to advance this field [14–16], such as Long et al. introduced a tri-channel neural network to predict potential miRNA-disease associations [17]. Wang et al. proposed SAEMDA, based on a stacked autoencoder, to predict potential miRNA biomarkers for diseases [18]. Li et al. developed the HHOMR model, based on hybrid high-order moment residuals, for effective prediction of unknown miRNA-disease associations [19]. Zhao et al. fused high-order and low-order structural information and proposed the MotifMDA model based on a hierarchical attention network, which effectively predicted new miRNA-disease associations [20]. In the realm of circRNA-disease association prediction, Wang et al. combined dataset integration strategies and attention mechanisms to propose the AMDECDA model for predicting potential circRNA biomarkers for diseases [21]. Lan et al. conducted a benchmark evaluation of computational methods for circRNA-disease association prediction [22]. Wang et al. constructed the GSLCDA model using an unsupervised deep graph structure learning method to predict potential circRNA biomarkers for diseases [23]. In predicting pathogenic circRNA biomarkers that competitively bind with miRNAs, Wang et al. used a multi-structural feature extraction framework to predict unknown circRNA-miRNA interactions [24]. Furthermore, Wang et al. introduced KS-CMI, which applies CMI predictions to real-world cases [25]. CMI prediction has also been extended to multi-source data fusion [26, 27] and knowledge graph domains [28]. Additionally, the construction and feature extraction methods of heterogeneous networks for various types of molecules have garnered widespread attention. For instance, Zhao et al. built a heterogeneous information network based on nine types of associations among five types of molecules and employed the HINLMI model, which incorporates neighborhood-level structural representations, to extract molecular features and predict associations between lncRNA and miRNA [29]. Similarly, Sheng et al. constructed a three-layer heterogeneous network for lncRNA-miRNA-disease and proposed the GCLMTP model, based on graph contrastive learning, to predict associations between lncRNA, miRNA, and disease [30].

The circRNA-miRNA-mediated model is a classic framework for gene regulatory expression, providing new research perspectives on cancer onset and treatment. Large-scale graph-based feature extraction methods can effectively construct graph-level models of biomolecules to support advanced downstream tasks and identify potential disease biomarkers. However, due to limitations in graph representation learning methods, existing approaches often modularize the complete biological network into isolated research units for computational modeling, focusing on single-type molecular association behaviors. Nonetheless, a cell, as a complete entity, is undoubtedly influenced by all components of the biological network system. Consequently, current studies are typically constrained to learning node-specific local and global topological structures and neighbor behavior association in the graph. Additionally, single-association modeling methods may face the risk of label leakage because using the same type of training data might lead the model to learn node features with label assumptions. Moreover, existing research often concentrates solely on training and predicting positive samples, which may result in the propagation of false-negative samples. Given the biological integrity of the ceRNA hypothesis, modeling complete biological regulatory behavior as a network aid in learning more comprehensive regulatory representations of molecules and discovering unknown associations within the network.

To address the aforementioned issues, we propose a method capable of modeling the circRNA-miRNA-cancer regulatory network (CMCN) and predicting associations between any two entities, termed CMCSG. Specifically, CMCSG constructs a ternary heterogeneous network based on the circRNA-miRNA-disease regulatory process. It then utilizes local and global neighborhood structure embedding modules to extract molecular topological features from the network. Subsequently, a signed graph representation learning approach, based on balance theory and status theory, is employed to construct chain-like social relationships for positive and negative samples, thereby propagating node structural features. Finally, these features are fed into a high-level predictor to infer relationships between circRNA, miRNA, and cancer.

Materials and methods

Dataset

The data used in this experiment are downloaded and preprocessed from the circR2Cancer database [31] (http://www.biobdlab.cn:8000/). All ceRNA data have experimental support, including the complete circRNA-miRNA-cancer regulatory association of 72 cancers, 731 pairs of miRNA and cancer associations, 648 circRNA-cancer associations, and 753 circRNA-miRNA associations. The data details are shown in Table 1. We uploaded the data to GitHub (https://github.com/1axin/CMCSG) under the name CMCSG. In addition, we introduce an independent dataset CMI-9905 (collated by Wang et al. [26]) in the functional verification to verify the effectiveness of the embedding method based on neighborhood structure.

CMCSG method

Figure 1 outlines our proposed multi-task prediction method. The method consists of three main parts. Firstly, the ternary heterogeneous association network CMCN between circRNA-miRNA-cancer is constructed according to the known data to embed the structural information between nodes. Secondly, for different prediction tasks, combined with the local and global neighborhood structure embedding module to extract the topology embedded representation of nodes from the CMCN, and denoising autoencoder (DAE) is used to enhance the features. Then, the signed graph attention network is used to propagate and aggregate the node information in the network, and the association representation of the node in the heterogeneous graph is obtained. Finally, the advanced predictor is used to predict the target association type (sub-network).

Construction of the ternary heterogeneous network of circRNA-miRNA-cancer

In this study, we used the regulatory associations of 72 cancers to construct a heterogeneous network CMCN. CMCN contains three kinds of associations among three kinds of molecules,

Table 1. The information of the CMCN.

dataset	pairs	circRNA	miRNA	disease	Average degree
CMA	753	515	477	non	1.5273
CCA	648	515	non	72	2.2003
MCA	731	non	477	72	2.6437

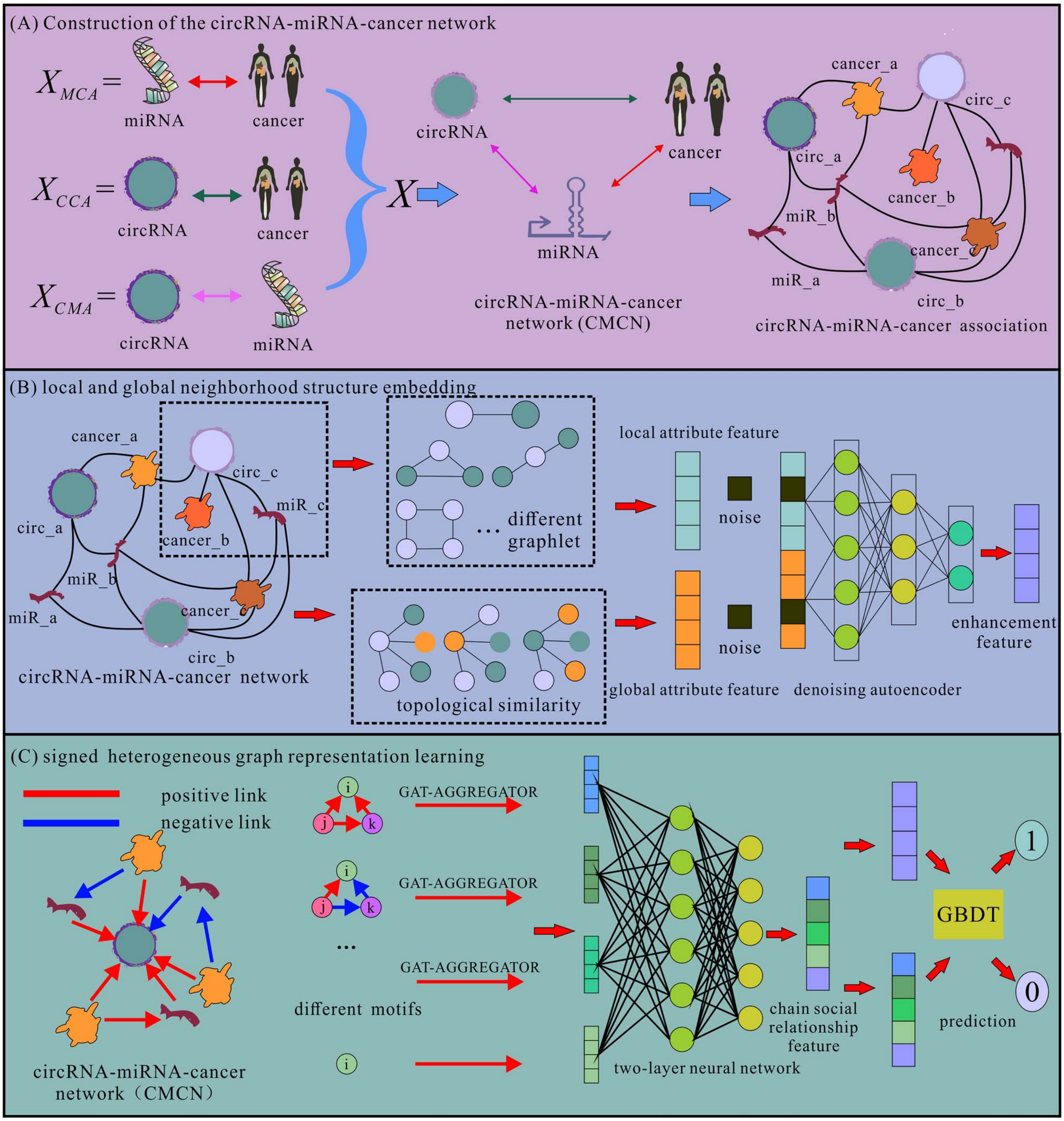


Figure 1. The flow chart of CMCSG.

including miRNA-cancer association, circRNA-miRNA interaction, and circRNA-cancer association. Combined with the relationship between biological entities, the molecules form a complete closed-loop triangular relationship structure. By embedding the network composed of two known associations, we can predict any unknown third association. Specifically, CMCN is defined as an undirected graph, which is expressed as G = <V, E>, where $V = \{V_{miRNA} \cup V_{circRNA} \cup V_{cancer}\}$ represents the set of nodes and E_{ij} represents the association between node V_i and node V_j. We use the adjacency matrix X to represent the association information

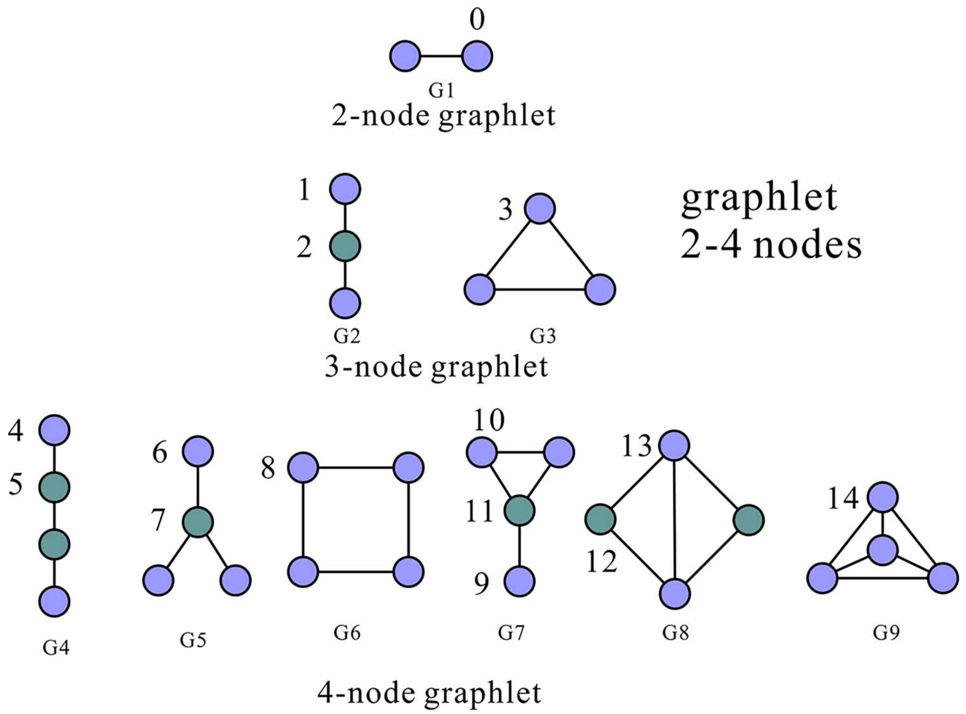


Figure 2. Different graphlets of K molecules (minimum 2 and maximum 4).

stored in graph G. The matrix X is defined as:

$$X = \begin{cases} X_{MCA} \in R^{N_{\mathrm{m}} \times N_{ca}} \\ X_{CCA} \in R^{N_{ci} \times N_{ca}} \\ X_{CMI} \in R^{N_{ci} \times N_m} \end{cases} \quad (1)$$

Where N_m, N_{ca}, and N_{ci} represent the number of miRNA, cancer, and circRNA respectively. If there is an association between node V_i and node V_j, X_{ij} is set to 1, otherwise 0.

Neighborhood structure embedding module based on local and global

The extraction of molecular attributes based on local topology

In the CMCSG method, we combine the definition of neighbor structure in the network to embed the node topology, which is implemented using the Role2vec [32] algorithm.

For an undirected graph G = <V, E > consisting of N nodes and E edges, the goal of role2vec is to divide different nodes into different 'graphlets', and map the topology information to a representation of a new feature space.

Graphlets are collections of subgraph structures within a graph that capture the complex relationships between nodes and their neighborhoods. CMCSG employs graphlet structures to represent the local structural features of nodes in the CMCN. Compared to focusing solely on individual edges or node neighbors, CMCSG can capture richer local patterns and topological structures, thereby providing a more accurate depiction of the roles of nodes and their relative positions within the network.

To obtain an effective target representation, we divide the associations of K nodes (minimum 2, maximum 4) into different graphlets, and define the role attributes of the nodes according to the graphlets, as shown in Fig. 2.

According to the structure definition in Fig. 2, we first extract all the structural information in graph G. Then we use the function Y to map the structure of node n_i to node type.

$$Y \approx F\left(UV^T\right) \quad (2)$$

Where F is a non-linear function.

Next, an attributed random walk is adapted to traverse graph G to learn the feature representation of the nodes. The attributed random walk summarizes the node structure types into different graphlets and uses them as node numbers, and then puts the obtained representation as a corpus into doc2vec in skip-gram for training to obtain the final embedded representation.

The extraction of molecular attributes based on the global topology

To compensate for the global topology of unrelated molecules, CMCSG introduces the struc2vec [33] algorithm to capture the structural similarity information between unrelated molecules through a biased random walk in the constructed multi-layer weighted graph.

In the undirected graph G < V, E > constructed from the data set, the direct nearest neighbor set of vertex U and the vertex set with distance d are represented by R_d(U), R_1(U) respectively, and S(s) represents the order degree of the vertex set. The structural distance $f_d(u,v)$ of the node set whose distance from the vertex U is less than d can be defined as:

$$f_{\mathrm{d}}\left(u, v\right) = f_{d-1}\left(u, v\right) + g\left(s\left(R_d(u)\right), s\left(R_d(v)\right)\right), d \geq 0, \left|R_d(u)\right|, \left|R_d(u)\right| > 0 \quad (3)$$

Where g () is used to measure the distance between ordered degree sequences.

Next, based on Dynamic Time Warping (DTW), the function used to measure the distance between sequences of different lengths containing duplicate elements is defined as W().

$$W\left(a,b\right)=\frac{\max\left(a,b\right)}{\min\left(a,b\right)}-1 \tag{4}$$

After the distance is defined, we construct a hierarchical weighted graph based on the ordered degree sequence distance between vertices for random walk node traversal. In weighted graphs at different levels, the edge weight is defined as T

$$T\left(u_d,u_{d+1}\right)=\log\left(\Gamma_d(u)+e\right),d=0,\ldots,d^*-1 \tag{5}$$

where Γ_d(u) is the number of edge weights greater than the average edge weight in the edge connected to vertex U at layer q.

$$\Gamma_q(u)=\sum\nolimits_{v\in V}1\left(w_q\left(u,v\right)>\overline{w}_q\right) \tag{6}$$

$\overline{w}_k$ is the average of all boundary rights in layer q.

In the constructed graph g, we use a biased random walk to sample vertex sequences. In layer q, the walking probability from vertex U to V is P.

$$p_q\left(u,v\right)=\frac{e^{-f_q(u,v)}}{Z_q(u)} \tag{7}$$

Z_q(u) is the normalization factor for Apex u in the q layer.

$$Z_q(u)=\sum\nolimits_{v\in V,v\neq u}e^{-f_q(u,v)} \tag{8}$$

With P, the vertices sampled each time are more likely to point similarly to the current vertex structure, independent of their positions in the graph.

Finally, by walking in Graph g, we get the global topological structure feature representation of each node.

The feature enhancement based on the noise reduction method

The information extraction in large-scale graphs has the characteristics of high independence, difficult fusion, and long dimensions. CMCSG introduces the Denoising Autoencoder [34] to learn the low-dimensional hidden features from the features obtained. Specifically, CMCSG obtains corrupted features by adding noise to the original feature and then forces the neural network to learn a low-dimensional representation of pure features from the corroded features. This strategy ensures that the learned features are not just low-dimensional replicas of the original features while removing uncertain effects in the original features. In detail, CMCSG first adds Gaussian noise to corrupt the original features t, then reconstructs the corroded features T and the original features t through the sigmoid function respectively. The sigmoid function can be calculated as follows:

$$J=f_{siomoid}(\mathrm{t})=\frac{1}{\left(1+e^{-\mathrm{t}}\right)} \tag{9}$$

$$j=f_{siomoid}(\mathrm{T})=\frac{1}{\left(1+e^{-\mathrm{T}}\right)} \tag{10}$$

Next, the neural network is forced to learn the original representation of the feature from the corroded feature (the deep neural network is trained to minimize the average reconstruction error) to obtain a low-dimensional replacement of the original feature. The average reconstruction error is defined as:

$$\lambda^*,\beta^*=\underset{\lambda\beta}{\operatorname{argmin}}\frac{1}{n}\sum_{m=1}^{n}L\left(J^{(m)},j^{(m)}\right) \tag{11}$$

Where n is the number of train data, λ^*, β^* is the optimal values of λ, β.

Where L is the reconstruction error.

In this experiment, we finally use DAE to extract 20-dimensional molecular functional feature descriptors.

Feature extraction based on signed graph representation learning

In this study, the information of the CMCN is stored in the adjacency matrix of 1064 × 1064, which means that 1,132,096 pairs of associations can be generated. However, only 4264 pairs of associations are known, therefore, most of the existing data cannot meet the requirements of a large number of relationships between nodes, which makes it difficult to capture valuable information in the graph. Recent research shows that negative links have the same or even higher value than positive links in social networks [35], and the association between positive and negative links can be effectively connected through the balanced path based on the balance theory, thereby effectively improving the learning effect of sample features in the sparse data In this study, we combine signed graph attention network (SiGAT) [36] for heterogeneous graph representation learning. The balanced path based on balance theory can effectively utilize positive and negative samples to construct chained social relationships, which not only leverages the learning value of negative samples but also extends multi-order molecular associations. The definition of molecular association structures based on status theory enables the evaluation and learning of molecular contributions under different neighborhood structures, thereby obtaining molecular features with high association values.

First, CMCSG defines different social relationships (friends or foes) for molecules according to the positive and negative associations between them.

$$U=U^{+}\cup U^{-} \tag{12}$$

$$U^{+}\cap U^{-}=\varnothing \tag{13}$$

Furthermore, we introduce balance theory to extend the positive and negative associations between molecules. Specifically, if molecule A is associated with molecule B and molecule B is associated with molecule C, then molecule A is associated with molecule C (i.e. the friend of a friend is still a friend). Conversely, if molecule A is associated with molecule B and molecule B is not associated with molecule C, then molecule A is not associated with molecule C (i.e. the friend of an enemy is still an enemy). From the perspective of molecular associations, if RNA A can target RNA B and RNA B can target RNA C, then RNA A and RNA C may exhibit functional or structural similarities. Conversely, if RNA A can target RNA B but RNA B cannot target RNA C, RNA A and RNA C may have significant functional or structural differences.

$$\begin{cases}\mathrm{A}\overset{+}{\rightarrow}\mathrm{C} & if \quad \mathrm{A}\overset{+}{\rightarrow}BandB\overset{+}{\rightarrow}\mathrm{C}\\ \mathrm{A}\overset{-}{\rightarrow}\mathrm{C} & if \quad \mathrm{A}\overset{+}{\rightarrow}BandB\overset{-}{\rightarrow}\mathrm{C}\\ \mathrm{A}\overset{-}{\rightarrow}\mathrm{C} & if \quad \mathrm{A}\overset{-}{\rightarrow}BandB\overset{+}{\rightarrow}\mathrm{C}\end{cases} \tag{14}$$

Combined with the balanced path, CMCSG allows the recursive definition of molecular roles:

$$\begin{gathered}R>1\\ I_u\left(L+1\right)=\left\{\mathrm{x}_u|\mathrm{x}_k\in I_u(L)and\mathrm{x}_u\in U_k^{+}\right\}\cup\left\{\mathrm{x}_u|\mathrm{x}_k\in J_u(L)and\mathrm{x}_u\in U_k^{-}\right\}\\ J_u\left(L+1\right)=\left\{\mathrm{x}_u|\mathrm{x}_k\in J_u(L)and\mathrm{x}_u\in U_k^{+}\right\}\cup\left\{\mathrm{x}_u|\mathrm{x}_k\in I_u(L)and\mathrm{x}_u\in U_k^{-}\right\}\end{gathered} \tag{15}$$

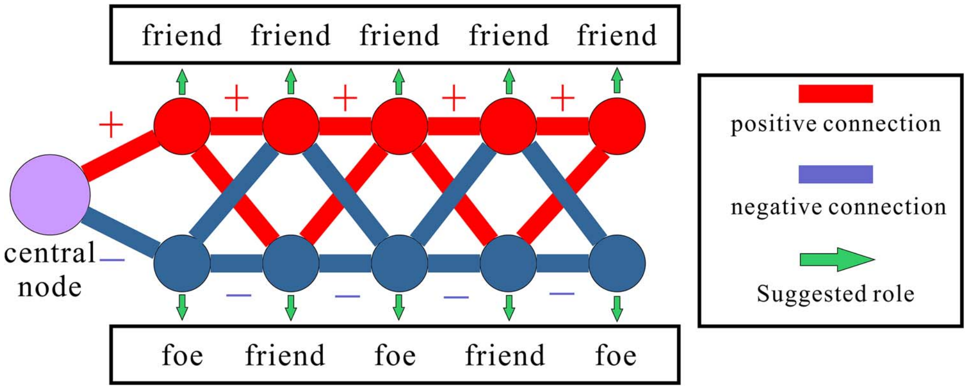


Figure 3. The definition of the balanced path to the molecular association.

Where I and J are the sets of positive and negative associations respectively. The definition of the balanced path to the molecular association is shown in Fig. 3.

Furthermore, molecules with both positive and negative links not only exhibit the transitivity of associations as described by balance theory [37] but also highlight different levels of importance of molecules within the network based on various association structures. Therefore, we introduce status theory [38] to provide a detailed description of node contributions within higher-order community structures. This characteristic can be expressed as:

$$\begin{cases} A > C & if \quad A \overset{+}{\rightarrow} B, B \overset{+}{\rightarrow} C \textbf{ and } A \overset{+}{\rightarrow} C \\ A > C & if \quad A \overset{+}{\rightarrow} B, B \overline{\rightarrow} C \textbf{ and } A \overset{+}{\rightarrow} C \\ \ldots & \ldots \qquad \ldots \\ C > A & if \quad A \overset{+}{\rightarrow} B, B \overline{\rightarrow} C \textbf{ and } A \overline{\rightarrow} C \\ C > A & if \quad A \overline{\rightarrow} B, B \overline{\rightarrow} C \textbf{ and } A \overline{\rightarrow} C \end{cases} \tag{16}$$

Specifically, CMCSG integrates status theory to describe a node's status, or contribution, within community structures. When molecule A positively associates with target molecule B, molecule A demonstrates a higher contribution value compared to molecule B. Conversely, when molecule A negatively associates with target molecule B, molecule A exhibits a lower contribution value relative to molecule B. With the relationship transitivity mechanism of balance theory, status theory is also applicable in ternary relationships; e.g. if molecule A positively associates with target molecule B and molecule B positively associates with target molecule C, then the contribution of molecule A is significantly higher than that of molecule C. By summarizing ternary association relationships under both positive and negative samples, we have constructed various association structure motifs [39], as illustrated in Fig. 4**.**

As shown in Fig. 4, after combining undirected relationships, four directed relationships, and two social theories, 38 motifs representing molecular biological processes are finally constructed to represent the social relations of molecules. To maximize the position of each molecule in the motifs, we adopt a targeted cycle strategy, i.e. the molecules in each motif cycle all the roles.

Based on traditional GAT, SiGAT obtains different neighborhoods from different motifs. Specifically, for each node y, we construct different local neighborhoods according to the different motifs and then use the parameters W_{mi} and a_{mi} to construct a GAT-AGGREGATOR to obtain the message X_{mi}. For each motif, the formula of GAT-AGGREGATOR with parameters W_{mi} and a to extracting feature X_{mi} can be calculated as:

$$\alpha_{yn}^{m_i} = \frac{\mathbf{exp}\left(Leaky\,\mathbf{Re}\,LU\left(a_{m_i}^T\left[W_{m_i}X(y)\middle\|W_{m_i}X(n)\right]\right)\right)}{\sum_{k\in N_{m_i}(u)}\mathbf{exp}\left(Leaky\,\mathbf{Re}\,LU\left(a_{m_i}^T\left[W_{m_i}X(y)\middle\|W_{m_i}X(k)\right]\right)\right)} \tag{17}$$

$$X_{m_i}(y) = \sum_{v\in N_{m_i}(y)} \alpha_{yv}^{m_i} W_{m_i} X(n) \tag{18}$$

Finally, the features of molecules in each neighborhood are connected to a two-layer neural network to obtain the final social relation embedding representation Z. The loss function L can be calculated as:

$$L(Z) = -\sum_{U^+\in N(y)^+} \mathbf{log}\left(S\left(Z_y^T Z_{U^+}\right)\right) - C\sum_{U^-\in N(y)^-} \mathbf{log}\left(S\left(-Z_y^T Z_{U^-}\right)\right) \tag{19}$$

Among L, S() is the sigmoid function, $N(n)^{+/-}$ is the set of positive and negative neighbors of node n, and C is the balance parameter.

Results

Evaluation criteria

In this study, we used five-fold cross-validation (five-fold CV) to evaluate the performance of CMCSG in unknown type association predictions. Specifically, we construct two types of associations in CMCN for feature extraction to predict the third unknown association. Five-fold CV divides the data into five equal subsets for five model training and prediction, each using a non-repetitive subset as the test set and the remaining subset as the training set until all subsets are verified as test sets, the average of the five experiments is calculated as the final result. In addition, we introduce accuracy (Acc.), precision (Prec.), F1-score, area under the ROC curve (AUC), and area under the P-R curve (AUPR) to comprehensively evaluate the performance of the model. These evaluation indicators are calculated as follows:

$$\text{Acc.} = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$\Pr ec. = \frac{TP}{TP + FP} \tag{21}$$

$$F1 - score = \frac{2prec \times recall}{prec + recall} \tag{22}$$

Among them, TP (true positive) and FP (false positive) represent the positive samples of correct and wrong prediction of the model, TN (true negative) and FN (false negative) represent the

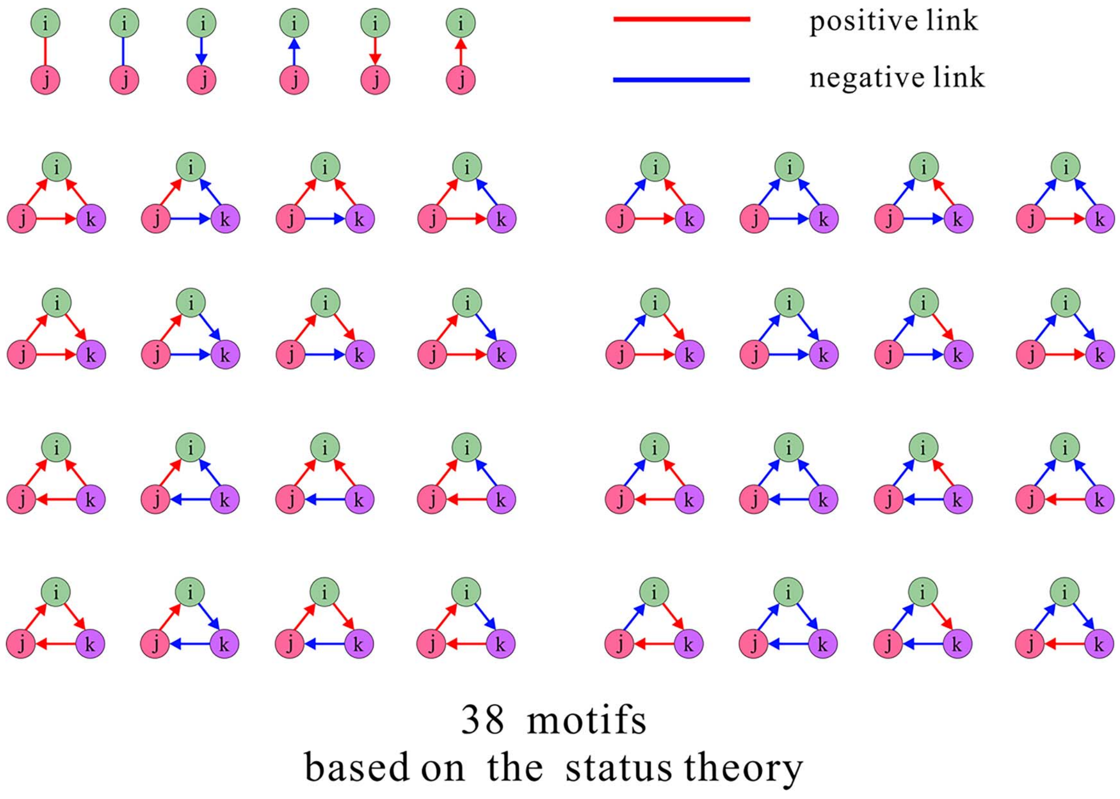


Figure 4. 38 motifs based on the status theory.

negative samples of correct and wrong prediction of the model, respectively.

Performance evaluation

To verify the ability of the CMCSG to predict the association between circRNA, miRNA, and cancer, we conducted independent prediction tasks for the miRNA-cancer association (MCA), circRNA-cancer association (CCA), and circRNA-miRNA interaction (CMI) in CMCN, and the result of the five-fold CV is recorded in Table 2. In addition, we draw the ROC and PR curve of three associated prediction tasks, as shown in Fig. 5.

As shown in Table 2, the average Acc, Prec, F1, AUC, and AUPR of the CMCSG in MCA prediction is 0.7387, 0.7391, 0.7386, 0.8200, and 0.8073, respectively; in CCA prediction, the average Acc, Prec, F1, AUC, and AUPR is 0.7415, 0.7417, 0.7414, 0.7956 and 0.7800, respectively; In CMI prediction, the average Acc, Prec, F1, AUC, and AUPR is 0.6407, 0.6411, 0.6405, 0.6921, and 0.6618, respectively. The results of the five-fold CV show that CMCSG can effectively predict the association of MCA, CCA, and CMI, and two of the three sets of associations have more than 75% AUC. The results show that CMCSG is a powerful method to predict the association between circRNA, miRNA, and cancer.

The validation of the local and global neighborhood structure embedding method

CMCSG combines the local and global neighborhood structure embedding method to extract the topological feature representation of each node in the network. The purpose of this method is to retain the structure information of the nodes in the graph from the point of topology, to obtain the low-dimensional embedded representation of the nodes. Therefore, the nodes with similar structures in the network should have similar feature representations. This property shows that the nodes with similar structures have a closer distance in the feature space, and vice versa. In this part, we combine the independent data set CMI-9905 and subgraph G to verify the effectiveness of the neighborhood structure embedding method. Specifically, we first construct a first-order subgraph G based on the CMI-9905 data set, which contains 12 association information between four kinds of circRNA and 12 kinds of miRNA. Then, we use the neighborhood structure embedding method to extract the features of 16 molecules in the CMI-9905 data set and subgraph G, and use DAE to compress the features into three dimensions. Finally, the node features are visualized in 3D space to observe whether the nodes with similar structures have similar spatial distances.

The validation based on local topology

In this part, we construct a first-order subgraph G based on CMI-9905 data sets to verify the effectiveness of the neighborhood structure embedding method in molecular local topology feature extraction, the subgraph G is shown in Fig. 6. We use the neighborhood structure embedding method to extract the node features in the subgraph G, and then use DAE to compress the features into three dimensions and project them to the 3D space. The experimental results are shown in Fig. 7. (It is worth noting that 3-dimensional features are not the best feature enhancement strategy, which has been verified in comparative experiments, but it is still of reference value for effective visualization.)

In Fig. 7, nodes with different colors correspond to molecules with different degrees. Role2vec algorithm and struc2vec algorithm can effectively extract node features with degree 1, but for molecules with higher degrees, the effect is poor, and the neighborhood structure embedding method can effectively extract the topological features of all nodes.

Table 2. The prediction result of the five-fold CV for three kinds of association.

MCA	**Acc**	**Prec.**	**F1-score**	**AUC**	**AUPR**
1	0.7338	0.7338	0.7338	0.8065	0.7955
2	0.7338	0.7342	0.7337	0.8412	0.8383
3	0.7740	0.7740	0.7740	0.8507	0.8353
4	0.6952	0.6963	0.6948	0.7838	0.7728
5	0.7568	0.7572	0.7568	0.8178	0.7947
Mean	0.7387	0.7391	0.7386	0.8200	0.8073
std	±0.0265	±0.0261	±0.0266	±0.0240	±0.0254
CCA	Acc	Prec.	F1-score	AUC	AUPR
1	0.7192	0.7192	0.7192	0.8021	0.8174
2	0.7761	0.7763	0.7760	0.8230	0.7789
3	0.7452	0.7454	0.7451	0.7959	0.7921
4	0.7375	0.7380	0.7374	0.7781	0.7416
5	0.7297	0.7298	0.7297	0.7793	0.7701
Mean	0.7415	0.7417	0.7414	0.7956	0.7800
std	±0.0193	±0.0193	±0.0192	±0.0165	±0.0249
CMI	Acc	Prec.	F1-score	AUC	AUPR
1	0.6391	0.6391	0.639	0.6901	0.6611
2	0.6478	0.6488	0.6474	0.6896	0.6636
3	0.6179	0.6179	0.6179	0.6721	0.6557
4	0.6412	0.6419	0.6409	0.7004	0.6702
5	0.6578	0.6582	0.6577	0.7085	0.6586
Mean	0.6407	0.6411	0.6405	0.6921	0.6618
std	±0.0131	±0.0133	±0.0130	±0.0122	±0.0049

The validation based on the global topology

In this part, we verify the effectiveness of the neighborhood structure embedding method in global topological feature extraction based on the CMI-9905 data set. We use the neighborhood structure embedding method to extract node features in the CMI-9905 dataset, and then use DAE to compress the features into three dimensions and project them into 3D space. The experimental results are shown in Fig. 8.

In Fig. 8, the role2vec algorithm cannot effectively extract the structural features of molecules, while the struc2vec algorithm can effectively extract molecular structural features, but the effect is relatively discrete. The proposed method achieves the best feature extraction effect. In fact, while role2vec can effectively capture the local topological structure of nodes, it is limited when applied to larger graphs. On the other hand, although struc2vec can capture structural similarities between distant nodes in large-scale graphs, it fails to capture the local topology of nearby nodes. By combining both approaches, CMCSG balances local and global topological features of molecules, thus achieving optimal performance.

The validation of feature extraction based on signed graph representation learning

The CMCSG method can be divided into two modules, one is a feature extraction module based on local and global neighborhood structure embedding (CMC-T), and the other is a heterogeneous graph node feature representation module (CMC-S) that combines balance theory and state theory. In the previous section, we verified the effectiveness of CMC-T through the visualization of molecular features. In this part, we verify the effectiveness of CMC-S in the proposed method through ablation experiments based on miRNA-cancer association prediction. Specifically, we kept the data used unchanged, and then used CMC-S for feature extraction and prediction tasks, and determined the contribution of CMC-S to model prediction by comparing the prediction performance. The results of the ablation experiment are recorded in Table 3. To facilitate comparison, we project the mean values of the evaluation criteria in Fig. 9.

Figure 9 data indicate that the CMC-S module performs well in MCA prediction, demonstrating its effectiveness in extracting molecular features. However, CMC-S exhibits lower performance across all evaluation metrics compared to CMCSG, with AUC and AUPR values being 5% lower. This suggests that the CMC-T module provides a valuable complement to feature extraction. Graph representation learning demonstrates notable advantages in feature extraction, primarily due to the CMCSG module's integration of balance theory and status theory for node feature extraction. This approach effectively captures complex relationships within the graph structure, including node connectivity patterns and structural information. Such capability allows for a more comprehensive understanding of node roles within the network, leading to the extraction of more representative features. Additionally, graph representation learning facilitates the propagation of node attribute features within the graph, further enhancing the understanding of node-related information.

Optimal dimensional selection of features for DAE

In this work, CMCSG uses DAE to strengthen the molecular features. The dimension size of DAE not only determines the effectiveness of feature extraction but also affects its contribution to the final feature. Therefore, it is important to choose the appropriate extract dimensions of the DAE.

To compare the impact of the dimensions extracted by DAE on feature extraction, we respectively extracted 16, 32, 64, 128, and 256-dimensional features of nodes based on miRNA-cancer association prediction, and performed the prediction task. The experimental results are recorded in Table 4. To facilitate comparison, we project the experimental data into 3D space, as shown in Fig. 10.

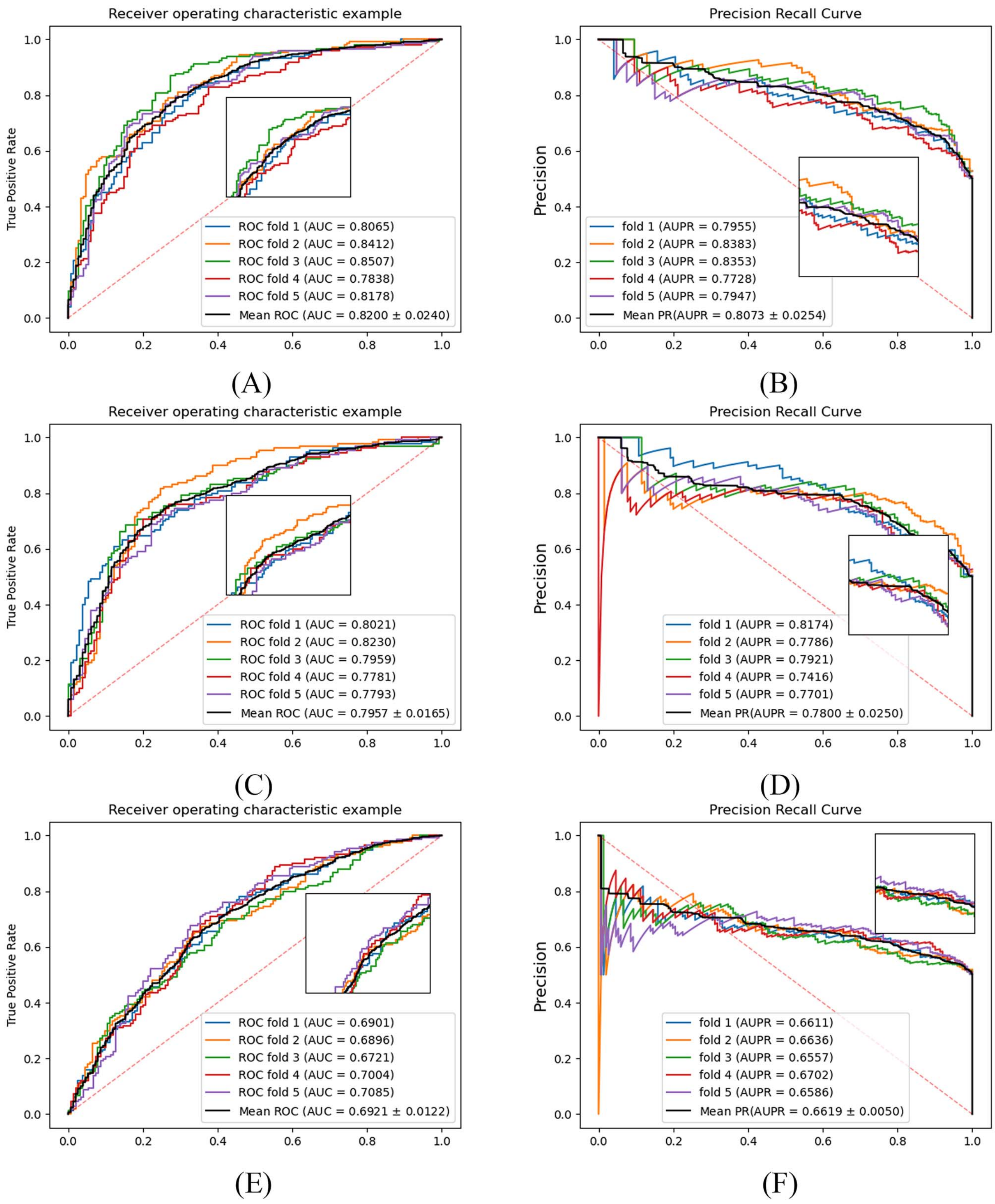


Figure 5. The ROC and PR curves of CMCSG ((A) and (B) are the ROC and PR curves of CMCSG in MCA prediction; (C) and (D) are the ROC and PR curves of CMCSG in CCA prediction; (E) and (F) are CMCSG ROC and PR curves in CMI prediction).

The DAE dimension comparison test results show that the 256-dimensional feature achieved the highest AUC value, but considering the five evaluation criteria and computational efficiency, the 128-dimensional feature achieved the best result. Therefore, the 128-dimensional feature was used as the best DAE extraction dimension in this study.

The selection of the best classification strategy

In this part, we select the best classification strategy for CMCSG through comparative experiments based on association prediction between circRNA, miRNA, and cancer. Specifically, we keep the data and node features used for training unchanged, and then use GBDT [40], RF (Random Forest) [41], LR (Logistic

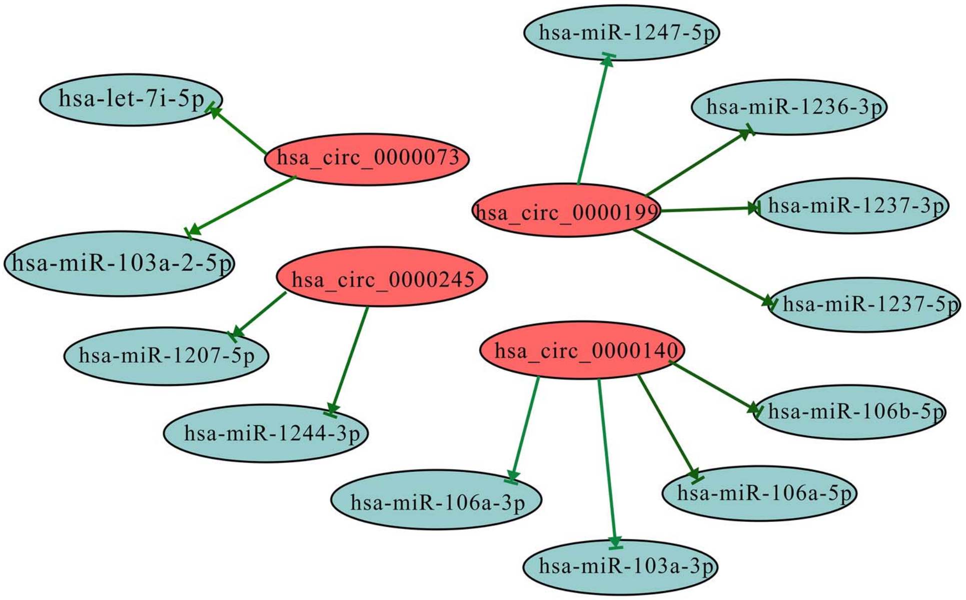


Figure 6. The subgraph G extracted from CMI-9905.

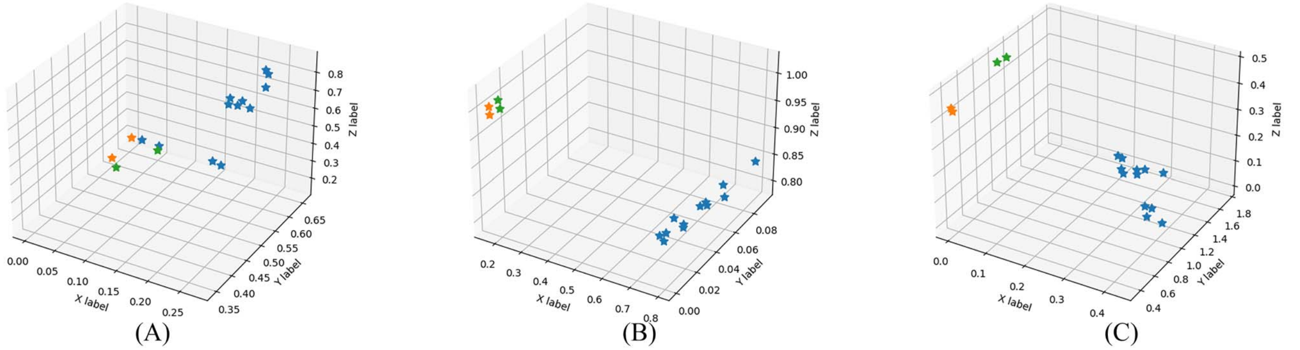


Figure 7. Three-dimensional visualization of neighborhood structure embedding method based on subgraph G ((A) is role2vec extraction, (B) is struc2vec extraction, (C) is CMCSG extraction)).

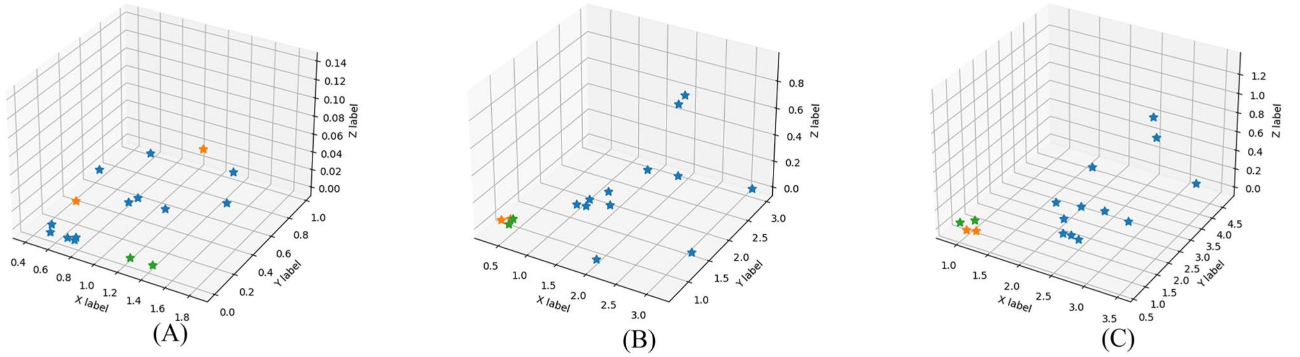


Figure 8. Three-dimensional visualization of neighborhood structure embedding method based on CMI-9905 ((A) is role2vec extraction, (B) is struc2vec extraction, (C) is CMCSG extraction)).

Regression) [42], SVM (Support Vector Machine) [43], KNN (K-Nearest Neighbor) [44] and LINR (Linear Regression) [45] classifiers to perform model MCA, CCA, and CMI prediction tasks respectively. The experimental ROC and P-R curves as shown in Fig. 11.

Figure 11 shows that the features extracted by the CMCSG method have achieved good results on all six classifiers, which proves the excellent performance of this method in feature extraction. Due to the special ensemble learning mechanism, the performance of the GBDT classifier is better than that of other single classifiers in three types of associated prediction tasks. Therefore, in this work, the GBDT classifier is used as the classification strategy of the model.

Comparison with existing models

To highlight the advantages of our proposed method in circRNA, miRNA, and cancer association prediction, we compared CMCSG with state-of-the-art (SOTA) models. To our knowledge, no existing

Table 3. Ablation experiment results.

CMC-S	Acc	Prec.	F1-score	AUC	AUPR
1	0.7065	0.7066	0.7065	0.7879	0.7655
2	0.7338	0.7349	0.7335	0.7763	0.7409
3	0.7397	0.7399	0.7397	0.8163	0.8065
4	0.7363	0.7364	0.7363	0.7848	0.7524
5	0.6644	0.6644	0.6644	0.7069	0.6867
Mean	0.7161	0.7164	0.7160	0.7744	0.7504
std	±0.0284	±0.0286	±0.0284	±0.0363	±0.0388

Table 4. The result of the different dimensions extracted.

CMA	Acc	Prec.	F1-score	AUC	AUPR
16	0.7202	0.7218	0.7198	0.7855	0.7635
32	0.7229	0.7236	0.7227	0.7866	0.7558
64	0.7442	0.7452	0.7439	0.8084	0.7875
128	0.7387	0.7391	0.7386	0.8200	0.8073
256	0.7585	0.7589	0.7584	0.8218	0.7844

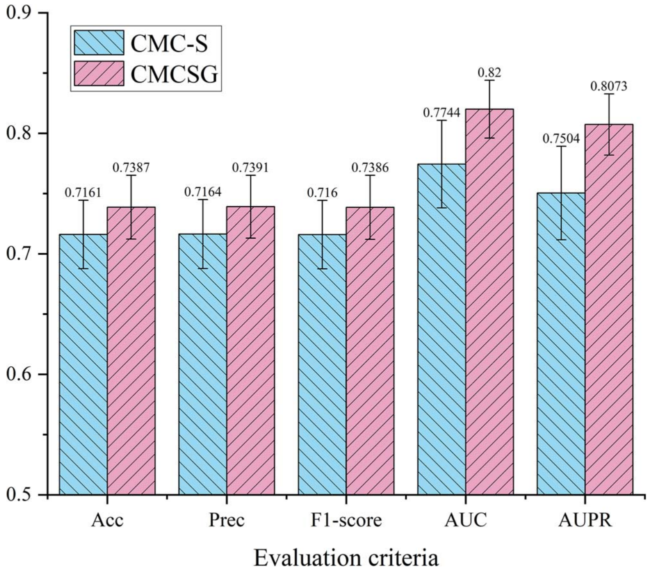


Figure 9. The average of the evaluation criteria of CMCSG and CMC-S.

models specifically address circRNA, miRNA, and cancer association prediction. In this section, we select SOTA models from the field of circRNA-miRNA interaction prediction and conduct comparisons across two commonly used datasets to validate the performance of our model.

Specifically, we compare our CMCSG model with NECMA [46], GCNCMI [47], CMIVGSD [48], and IIMCCMA [49] models using the CMI-753 dataset. Notably, due to the advantages of the CMCN construction, CMCSG does not incorporate any CMI associations in feature engineering, whereas the other models utilize similar CMI data for feature extraction and five-fold cross-validation. The CMI-753 dataset serves as a benchmark for prediction tasks in the CMI domain, presenting significant challenges due to its high sparsity and lack of RNA sequences. The SOTA models we compared include NECMA, a matrix factorization-based CMI prediction model; GCNCMI, which employs graph convolutional networks to extract molecular features for potential CMI prediction; CMIVGSD, which combines linear and nonlinear features for predicting unknown CMIs; and IIMCCMA, which uses matrix factorization and inductive matrix completion strategies to predict CMIs. The predictive performance of the models based on the CMI-753 dataset is derived from the work of Yao et al., and the comparison results are summarized in Table 5.

Table 5. Comparison results between CMCSG and SOTA based on the CMI-753 dataset.

CMI-753	AUC	AUPR
NECMA	0.4989	0.0003
GCNCMI	0.5679	0.0004
CMIVGSD	0.5755	0.0007
IIMCCMA	0.6702	0.0009
CMCSG	0.6921	0.6618

Table 5 indicates that in the prediction task on the CMI-753 dataset, the CMCSG model achieved the highest predictive performance, with the AUC exceeding that of the second-best model by more than 2%. This is due to the small data size and lack of attribute features in the CMI-753 dataset, which makes it challenging to extract effective molecular association information when relying solely on CMI links. The construction of the multi-source heterogeneous network CMCN effectively mitigated the sparsity of the CMI network, while the integration of graph representation learning based on social theories enabled the incorporation of both positive and negative samples into the modeling process. This approach also introduced chain features, further enhancing the understanding of node behavior within the network.

CMI-9905 is one of the most widely used benchmark datasets for CMI prediction tasks. Using this dataset, we compare our model with JSNDCMI [24], BCMCMI [50], BioDGW-CMI [51], DeepCMI [52], KS-CMI [25], BEROLECMI [53], and RBNE-CMI [54] models. JSNDCMI employs a multi-feature fusion framework to extract diverse structural features of nodes in the CMI network and predict unknown CMIs. BCMCMI extracts node features based on meta-paths for CMI prediction. DeepCMI combines rich text

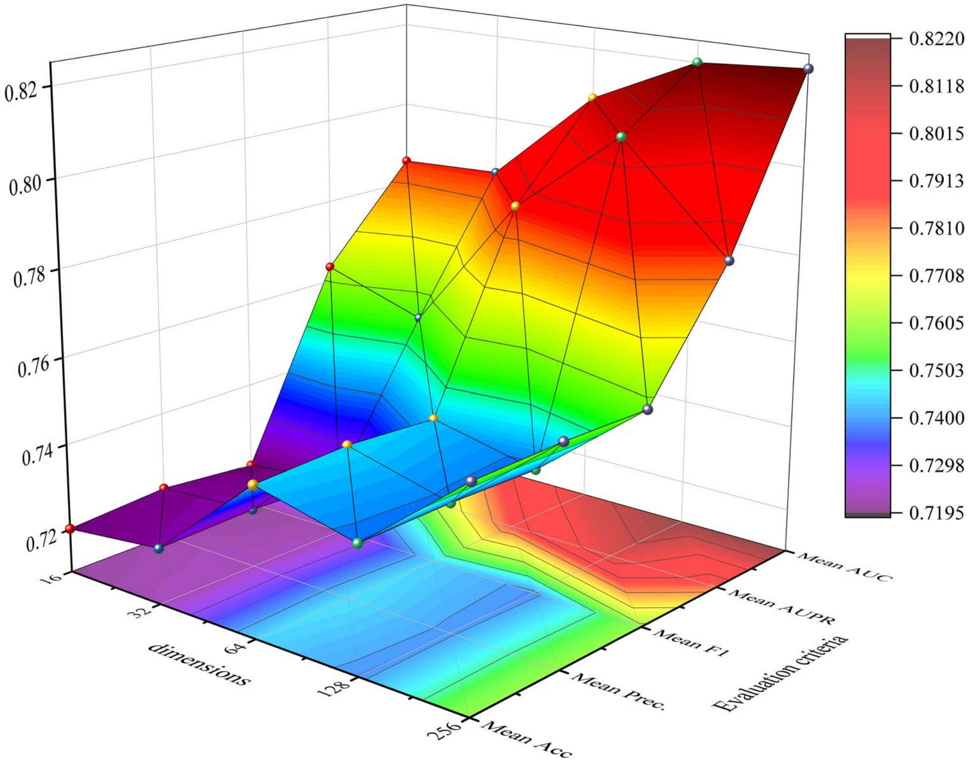


Figure 10. The prediction results of MCA under different DAE dimensions.

Table 6. Comparison results between CMCSG and SOTA based on the CMI-9905 dataset.

Methods	JSNDCMI	BioDGW-CMI	BCMCMI	DeepCMI	KS-CMI	BEROLECMI	RBNE-CMI	CMCSG
F1	0.8217	0.8322	0.8359	0.8232	0.8340	0.8392	0.8431	0.8451
Prec	0.8232	0.8346	0.8083	0.8290	0.8366	0.8427	0.8456	0.8486
ACC	0.8231	0.8325	0.8316	0.8244	0.8343	0.8395	0.8434	0.8454
AUC	0.9003	0.9026	0.9041	0.9054	0.9086	0.9104	0.9142	0.9205
AUPR	0.8999	0.8962	0.8990	0.8978	0.9144	0.9086	0.9144	0.9183

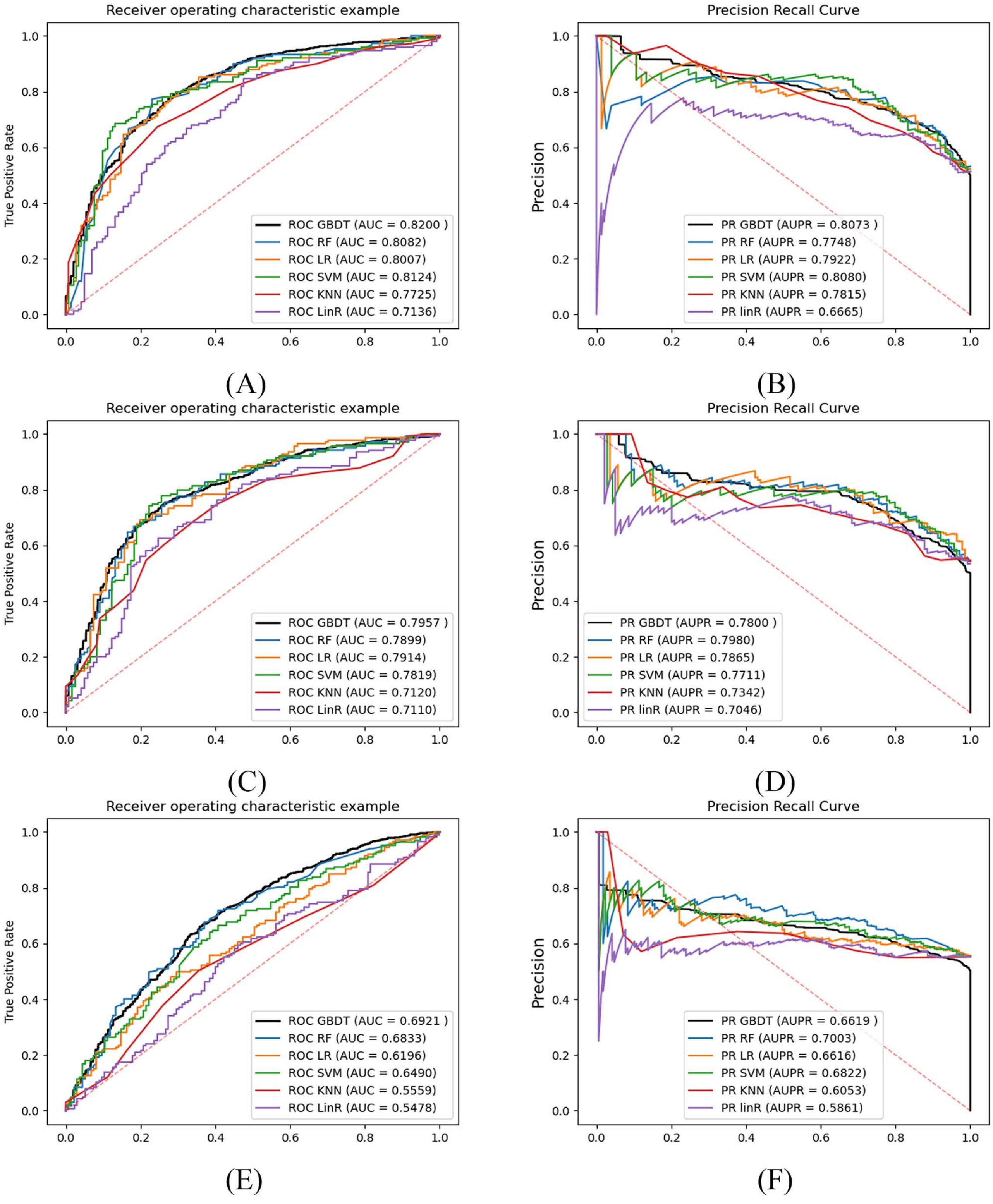
embeddings to extract network features of nodes for CMI prediction. KS-CMI uses a balance theory-based signed graph convolutional network to extract node features and predict CMIs. BEROLECMI extracts node features to predict CMI by defining molecular roles. RBNE-CMI utilizes an incomplete attribute network embedding framework to extract molecular embedding features for CMI prediction. The predictive performance of the models based on the CMI-9905 dataset is derived from the work of Yu et al., and the comparison results are summarized in Table 6.

Table 6 demonstrates that CMCSG achieved competitive performance on the CMI-9905 dataset. Notably, the SOTA model includes RNA sequences as attribute features in its feature engineering process. In contrast, CMCSG, which is designed for circRNA-miRNA-cancer prediction tasks, does not incorporate targeted attribute feature extraction strategies. Despite this, CMCSG still achieved the highest predictive performance, proving the superiority of the proposed model. Additionally, the experimental results suggest that the feature extraction method based on local and global neighborhood structure embedding and signed graph representation learning is not only advantageous in multi-source networks but also remains a strong choice for modeling and prediction tasks of the single association. This is because the use of multi-structure analysis and feature extraction deepens the model's understanding of node interactions within the network, thereby enhancing the value of the extracted features.

Case study

To validate the practicality of the CMCSG method in cancer biomarker detection, we predicted bladder cancer-related miRNA biomarkers and cancer-related circRNA-miRNA interactions based on the CMI-753 dataset. We identified 15 miRNA biomarkers for bladder cancer and 10 cancer-related circRNA-miRNA interactions, with the results presented in Tables 7 and 8, respectively.

Table 7 shows that among the top 15 miRNA biomarkers, 12 were confirmed by the circR2Cancer database. Additionally, the three unconfirmed miRNAs do not indicate prediction errors, but rather represent high-probability cancer biomarkers yet to be discovered. Table 8 reveals that out of the ten predicted promising cancer-related circRNA-miRNA interactions, eight were confirmed by the circR2Cancer database, while the remaining two are potential pathogenic circRNA-miRNA associations. The case study results indicate that CMCSG is a promising computational model candidate for predicting potential cancer biomarkers.



(A) (B)

(C) (D)

(E) (F)

Figure 11. The ROC and PR curves of different classifiers ((A) and (B) are the ROC and PR curves of CMCSG in MCA prediction; (C) and (D) are the ROC and PR curves of CMCSG in CCA prediction; (E) and (F) are CMCSG ROC and PR curves in CMI prediction).

Conclusion

In the past few years, ceRNA has been proven to be an important regulator, which leads to a variety of diseases through the mediation of miRNA. The study of the ceRNA network can provide new opportunities for the diagnosis, treatment, and prognosis of complex diseases. The use of calculation methods can effectively promote the discovery of unknown associations at a low cost. The current methods are influenced by reductionism, and most of them independently study specific parts of biological systems or binary associations such as miRNA-disease association,

Table 7. Case study of miRNA marker prediction based on bladder cancer.

Num	Cancer	miRNA	Evidence	detection method
1	bladder cancer	miR-17	PMID:29386015	qRT-PCR
2	bladder cancer	miR-132-3p	PMID:30983072	qPCR;western blots
3	bladder cancer	miR-3666	PMID:30984788	qRT-PCR
4	bladder cancer	miR-191-5p	PMID:31802888	RT-qPCR
5	bladder cancer	miR-103a-3p	PMID:27484176	qPCR; Western blot etc.
6	bladder cancer	miR-570-3p	PMID:32072011	qRT-PCR
7	bladder cancer	miR-499a-3p	Unconfirmed	None
8	bladder cancer	miR-1305	PMID:32019579	qPCR
9	bladder cancer	miR-29a-3p	PMID:27363013	qRT-PCR; microarray
10	bladder cancer	miR-940	Unconfirmed	None
11	bladder cancer	miR-142-5p	PMID:31777254	qRT-PCR;microarray
12	bladder cancer	miR-181a-5p	PMID:30999937	qRT-PCR
13	bladder cancer	miR-1184	PMID:31758655	qRT-PCR
14	bladder cancer	miR-200b-3p	Unconfirmed	None
15	bladder cancer	miR-1178-3p	PMID:30458784	qRT-PCR

Table 8. Case study on the prediction of cancer-related circRNA-miRNA interactions.

Num	circRNA	miRNA	Evidence	detection method
1	BCRC-3	miR-182-5p	PMID:30285878	qRT-PCR
2	circ_ANKIB1	miR-19b	PMID:31667786	qRT-PCR
3	circ_ARF3	miR-1299	PMID:32240746	qRT-PCR
4	circ_ANKIB1	miR-1271	Unconfirmed	None
5	circ_SPECC1	miR-526b	PMID:31349968	qPCR
6	circABCC2	miR-665	PMID:31417632	qRT-PCR
7	circ-ACACA	miR-1183	PMID:32236577	qPCR
8	circ-ACACA	miR-370-3p	Unconfirmed	None
9	circASAP1	miR-326	PMID:31838741	qRT-PCR; FISH
10	circASAP1	miR-532-5p	PMID:31838741	qRT-PCR; FISH

circRNA-disease association, and circRNA-miRNA interaction. Because the specific network in the organism works as a whole, it is more suitable to model the specific association prediction by using the large-scale biological network. From the point of feature extraction, most of the existing methods use a single type of association as training and prediction data, which may lead to incomplete molecular feature extraction or label leakage.

To solve the above problems, we put forward the CMCSG method in this work, which solves the above problems from two aspects. Firstly, CMCSG constructs a circRNA-miRNA-cancer network (CMCN) based on 72 kinds of cancer regulatory networks, which contains rich association information, makes the proposed method capable of multi-task prediction; for the possible problems of incomplete feature extraction and label leakage, CMCSG uses known association information to extract all molecular features and then predicts unknown association sub-network inference model, which effectively avoids the existence of this problem. In addition, we also embed the molecular network structure with the local and global topological structure of molecules to obtain the structural representation of nodes; for sparse graph representation learning, we introduce the signed graph attention network to maximize molecular links by combining status theory and balance theory. Finally, the proposed model achieved leading predictive performance and showed the potential to be a powerful tool for predicting cancer marker associations in case studies.

However, this study still has some limitations. In the heterogeneous graph representation of the learning module, due to the need to sample as many as 38 kinds of motifs for each node, it greatly increases the time complexity of CMCSG, and it is difficult to deal with larger graphs, which increases the computational overhead. With the progress of graph representation learning methods, combining more advanced graph representation learning methods (such as deep attribute graph clustering [55] and hypergraphs [56]) for feature learning is also an important basis for subsequent research and development. In the follow-up research, we will further improve this problem to obtain an excellent prediction model with higher efficiency and more accurate performance.

Key Points

- We propose an embedding method based on neighborhood structure, which embeds the global and local topologies of nodes in CMCN networks, and effectively preserves the structural representation of nodes in the network.
- Using signed graph attention network propagation and aggregation node features based on balance theory and status theory, which can make effective use of positive and negative samples and maximize the positive and negative links of molecules.
- The sub-network inference mode is used to extract the features of all molecules by using binary associations

in ternary heterogeneous networks to predict unknown types of associations and effectively avoid label leakage.

Author contributions

H-L, X-FW, W-Y: conceptualization, methodology, software. H-L, W-Y, R-CG: resources, and data curation. X-FW, S-N, XP-X, W-L, and Z-QZ: validation. All authors contributed to the manuscript revision and approved the submitted version.

Conflict of interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by the National Natural Science Foundation of China (No.62072212), the Jilin Provincial Scientific and Technological Development Program (No.20230201065GX, 20240101364JC) and the Jilin Provincial Key Laboratory of Big Data Intelligent Cognition (No.YDZJ202402075CXJD).

Data availability

The datasets for this paper can be found in the CircR2Cancer http://www.biobdlab.cn:8000/. CMCSG and CMCSG can be found at https://github.com/1axin/CMCSG.

References

1. Djebali S, Davis CA, Merkel A. *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8. https://doi.org/10.1038/nature11233.
2. Kasinski AL, Slack FJ. MicroRNAs en route to the clinic: Progress in validating and targeting microRNAs for cancer therapy. *Nat Rev Cancer* 2011;**11**:849–64. https://doi.org/10.1038/nrc3166.
3. Stahlhut C, Slack FJ. MicroRNAs and the cancer phenotype: Profiling, signatures and clinical implications. *Genome Med* 2013;**5**: 111–2. https://doi.org/10.1186/gm516.
4. Bartel DP. MicroRNAs: Target recognition and regulatory functions. *Cell* 2009;**136**:215–33. https://doi.org/10.1016/j.cell.2009.01.002.
5. Seitz H. Redefining microRNA targets. *Curr Biol* 2009;**19**:870–3. https://doi.org/10.1016/j.cub.2009.03.059.
6. Salmena L, Poliseno L, Tay Y. *et al.* A ceRNA hypothesis: The Rosetta stone of a hidden RNA language? *Cell* 2011;**146**:353–8. https://doi.org/10.1016/j.cell.2011.07.014.
7. Franco-Zorrilla JM, Valli A, Todesco M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* 2007;**39**:1033–7. https://doi.org/10.1038/ng2079.
8. Poliseno L, Salmena L, Zhang J. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010;**465**:1033–8. https://doi.org/10.1038/nature09144.
9. Karreth FA, Tay Y, Perna D. *et al.* In vivo identification of tumor-suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* 2011;**147**:382–95. https://doi.org/10.1016/j.cell.2011.09.032.
10. Calin GA, Dumitru CD, Shimizu M. *et al.* Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci* 2002;**99**: 15524–9. https://doi.org/10.1073/pnas.242606799.
11. Austin TBM. First microRNA mimic enters clinic. *Nat Biotechnol* 2013;**31**:577. https://doi.org/10.1038/nbt0713-577.
12. Manterola L, Guruceaga E, Pérez-Larraya JG. *et al.* A small noncoding RNA signature found in exosomes of GBM patient serum as a diagnostic tool. *Neuro Oncol* 2014;**16**:520–7. https://doi.org/10.1093/neuonc/not218.
13. Guo JU, Agarwal V, Guo H. *et al.* Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* 2014;**15**:1–14. https://doi.org/10.1186/s13059-014-0409-z.
14. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: Towards systematic evaluation of computational models. *Brief Bioinform* 2022;**23**:bbac407. https://doi.org/10.1093/bib/bbac407.
15. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: Experimental results, databases, webservers and data fusion. *Brief Bioinform* 2022;**23**:bbac397. https://doi.org/10.1093/bib/bbac397.
16. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: Taxonomy, trends and challenges of computational models. *Brief Bioinform* 2022;**23**:bbac358. https://doi.org/10.1093/bib/bbac358.
17. Long S, Tang X, Si X. *et al.* TriFusion enables accurate prediction of miRNA-disease association by a tri-channel fusion neural network. *Communications Biology* 2024;**7**:1067. https://doi.org/10.1038/s42003-024-06734-0.
18. Wang C-C, Li T-H, Huang L. *et al.* Prediction of potential miRNA–disease associations based on stacked autoencoder. *Brief Bioinform* 2022;**23**:bbac021. https://doi.org/10.1093/bib/bbac021.
19. Li Z, Wan L, Wang L. *et al.* HHOMR: A hybrid high-order moment residual model for miRNA-disease association prediction. *Brief Bioinform* 2024;**25**:bbae412. https://doi.org/10.1093/bib/bbae412.
20. Zhao BW, He YZ, Su XR. *et al.* Motif-aware miRNA-disease association prediction via hierarchical attention network. *IEEE J Biomed Health Inform* 2024;**28**:4281–94. https://doi.org/10.1109/JBHI.2024.3383591.
21. Wang L, Wong L, You ZH. *et al.* AMDECDA: Attention mechanism combined with data ensemble strategy for predicting CircRNA-disease association. *IEEE Trans Big Data* 2024;**10**:320–9. https://doi.org/10.1109/TBDATA.2023.3334673.
22. Lan W, Dong Y, Zhang H. *et al.* Benchmarking of computational methods for predicting circRNA-disease associations. *Brief Bioinform* 2023;**24**:bbac613. https://doi.org/10.1093/bib/bbac613.
23. Wang L, Li Z-W, You Z-H. *et al.* GSLCDA: An unsupervised deep graph structure learning method for predicting CircRNA-disease association. *IEEE J Biomed Health Inform* 2023;**28**:1742–51. https://doi.org/10.1109/JBHI.2023.3344714.
24. Wang X-F, Yu CQ, You ZH. *et al.* A feature extraction method based on noise reduction for circRNA-miRNA interaction prediction combining multi-structure features in the association networks. *Brief Bioinform* 2023;**24**:bbad111. https://doi.org/10.1093/bib/bbad111.
25. Wang X-F, Yu CQ, You ZH. *et al.* KS-CMI: A circRNA-miRNA interaction prediction method based on the signed graph neural network and denoising autoencoder. *Iscience* 2023;**26**:107478. https://doi.org/10.1016/j.isci.2023.107478.
26. Wang X-F, Yu CQ, Li LP. *et al.* KGDCMI: A new approach for predicting circRNA–miRNA interactions from multi-source information extraction and deep learning. *Front Genet* 2022;**13**:958096. https://doi.org/10.3389/fgene.2022.958096.

27. Zhou J, Wang X, Niu R. *et al.* Predicting circRNA-miRNA interactions utilizing transformer-based RNA sequential learning and high-order proximity preserved embedding. *Iscience* 2024;**27**:108592. https://doi.org/10.1016/j.isci.2023.108592.
28. Wei M, Wang L, Li Y. *et al.* BioKG-CMI: A multi-source feature fusion model based on biological knowledge graph for predicting circRNA–miRNA interactions. *Science China Information Sciences* 2024;**67**:1–2. https://doi.org/10.1007/s11432-024-4098-3.
29. Zhao B-W. *et al.* A heterogeneous information network learning model with neighborhood-level structural representation for predicting lncRNA-miRNA interactions. *Comput Struct Biotechnol J* 2024;**23**:2924–33. https://doi.org/10.1016/j.csbj.2024.06.032.
30. Sheng N, Wang Y, Huang L. *et al.* Multi-task prediction-based graph contrastive learning for inferring the relationship among lncRNAs, miRNAs and diseases. *Brief Bioinform* 2023;**24**:bbad276. https://doi.org/10.1093/bib/bbad276.
31. Lan W, Zhu M, Chen Q. *et al.* CircR2Cancer: A manually curated database of associations between circRNAs and cancers. *Database* 2020;**2020**:baaa085. https://doi.org/10.1093/database/baaa085.
32. Ahmed NK. *et al.* Learning role-based graph embeddings. *arXiv preprint arXiv*:180202896 2018.
33. Ribeiro LF, Saverese PH, Figueiredo DR. struc2vec: Learning node representations from structural identity. In:*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 385–94, 2017.
34. Vincent P, Larochelle H, Bengio Y. *et al.* Extracting and composing robust features with denoising autoencoders. In:*Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–103, 2008.
35. Leskovec J, Huttenlocher D, Kleinberg JJR. *et al.* Predicting positive and negative links in online social networks. In:*Proceedings of the 19th International Conference on World Wide Web*, pp. 641–50, 2010.
36. J. Huang, H. Shen, L. Hou, and X. Cheng. Signed graph attention networks. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks*. Munich, Germany: Springer International Publishing, 2019, pp. 566–77. https://doi.org/10.1007/978-3-030-30493-5_53.
37. Heider FJTJOP. *Attitudes and cognitive organization* 1946;**21**:107–12. https://doi.org/10.1080/00223980.1946.9917275.
38. J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 743–52. ACM, ed. ACM, 2012.
39. Milo R, Shen-Orr S, Itzkovitz S. *et al. Network motifs: simple building blocks of complex networks* 2002;**298**:824–7.
40. Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of statistics* 2001;1189–1232.
41. Breiman L. Random forests. *Machine learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324.
42. LaValley MP. Logistic regression. *Circulation* 2008;**117**:2395–9. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.
43. Hearst MA, Dumais ST, Osuna E. *et al.* Support vector machines. *IEEE Intelligent Systems and their applications* 1998;**13**:18–28. https://doi.org/10.1109/5254.708428.
44. G. Guo, H. Wang, D. Bell. *et al.* KNN model-based approach in classification. In *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3–7, 2003. Proceedings*. Berlin Heidelberg: Springer, 2003, pp. 986–96 **2888**. https://doi.org/10.1007/978-3-540-39964-3_62.
45. Naseem I, Togneri R, Bennamoun M. Linear regression for face recognition. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:2106–12. https://doi.org/10.1109/TPAMI.2010.128.
46. Lan W, Zhu M, Chen Q. *et al.* Prediction of circRNA-miRNA associations based on network embedding. *Complexity* 2021;**2021**:6659695. https://doi.org/10.1155/2021/6659695.
47. He J, Xiao P, Chen C. *et al.* GCNCMI: A graph convolutional neural network approach for predicting circRNA-miRNA interactions. *Front Genet* 2022;**13**:959701. https://doi.org/10.3389/fgene.2022.959701.
48. Y. Qian, J. Zheng, Z. Zhang. *et al. CMIVGSD: circRNA-miRNA Interaction Prediction Based on Variational Graph Auto-Encoder and Singular Value Decomposition*. IEEE, 2021, pp. 205–10.
49. Yao D, Nong L, Qin M. *et al.* Identifying circRNA-miRNA interaction based on multi-biological interaction fusion. *Front Microbiol* 2022;**13**:987930. https://doi.org/10.3389/fmicb.2022.987930.
50. Wei M-M, Yu C-Q, Li L-P. *et al.* BCMCMI: A fusion model for predicting circRNA-miRNA interactions combining semantic and meta-path. *J Chem Inf Model* 2023;**63**:5384–94. https://doi.org/10.1021/acs.jcim.3c00852.
51. Wang X-F, Yu C-Q, You Z-H. *et al.* An efficient circRNA-miRNA interaction prediction model by combining biological text mining and wavelet diffusion-based sparse network structure embedding. *Comput Biol Med* 2023;**165**:107421. https://doi.org/10.1016/j.compbiomed.2023.107421.
52. Li Y-C. *et al.* DeepCMI: A graph-based model for accurate prediction of circRNA–miRNA interactions with multiple information. *Brief Funct Genomics* 2023;**23**:elad030.
53. Wang X-F, Yu CQ, You ZH. *et al.* BEROLECMI: A novel prediction method to infer circRNA-miRNA interaction from the role definition of molecular attributes and biological networks. *BMC bioinformatics* 2024;**25**:264. https://doi.org/10.1186/s12859-024-05891-7.
54. Yu C-Q, Wang XF, Li LP. *et al.* RBNE-CMI: An efficient method for predicting circRNA-miRNA interactions via multiattribute incomplete heterogeneous network embedding. *J Chem Inf Model* 2024;**64**:7163–72. https://doi.org/10.1021/acs.jcim.4c01118.
55. Yang Y, Su X, Zhao B. *et al.* Fuzzy-based deep attributed graph clustering. *IEEE Trans Fuzzy Syst* 2024;**32**:1951–64. https://doi.org/10.1109/TFUZZ.2023.3338565.
56. Feng Y, You H, Zhang Z. *et al. Hypergraph neural networks* 2019;**33**:3558–65. https://doi.org/10.1609/aaai.v33i01.33013558.