BMC Bioinformatics

RESEARCH Open Access

# CMAGN: circRNA–miRNA association prediction based on graph attention auto-encoder and network consistency projection

Anhui Yin[1], Lei Chen[1*], Bo Zhou[2,3] and Yu-Dong Cai[4*]

*Correspondence: chen_lei1@163.com; cai_yud@126.com

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China
[2] Institute of Wound Prevention and Treatment, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[3] School of Basic Medical Sciences, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[4] School of Life Sciences, Shanghai University, Shanghai 200444, People's Republic of China

## Abstract

Background: As noncoding RNAs, circular RNAs (circRNAs) can act as microRNA (miRNA) sponges due to their abundant miRNA binding sites, allowing them to regulate gene expression and influence disease development. Accurately identifying circRNA–miRNA associations (CMAs) is helpful to understand complex disease mechanisms. Given that biological experiments are time consuming and labor intensive, alternative computational methods to predict CMAs are urgently needed. Results: This study proposes a novel computational model named CMAGN, which incorporates several advanced computational methods, for predicting CMAs. First, similarity networks for circRNAs and miRNAs are constructed according to their sequences. Graph attention autoencoder is then applied to these networks to generate the first representations of circRNAs and miRNAs. The second representations of circRNAs and miRNAs are obtained from the CMA network via node2vec. The similarity networks of circRNAs and miRNAs are reconstructed on the basis of these new representations. Finally, network consistency projection is applied to the reconstructed similarity networks and the CMA network to generate a recommendation matrix. Conclusion: Five-fold cross-validation of CMAGN reveals that the area under ROC and PR curves exceed 0.96 on two widely used CMA datasets, outperforming several existing models. Additional tests elaborate the reasonability of the architecture of CMAGN and uncover its strengths and weaknesses.

**Keywords:** CircRNA-miRNA associations, Graph attention auto-encoder, Network consistency projection, Node2vec

## Introduction

The significance of noncoding RNAs (ncRNAs) in cellular processes has become increasingly prominent. Different from conventional RNA molecules, ncRNAs do not directly encode proteins. Mounting evidence indicates that they perform crucial within-cell biological functions that are closely linked to health and diseases. Since the first discovery of ncRNAs in the 1960s, the field of transcriptomics has steadily expanded and

uncovered the involvement of ncRNAs in numerous cellular processes [1]. Among them, circular RNAs (circRNAs) and microRNAs (miRNAs) have garnered particular attention. Many studies have reported close relationships between diseases and circRNAs or miRNAs [2–9]. Thus, investigations on circRNAs and miRNAs can improve our understanding of various diseases.

MiRNAs are a class of small noncoding RNA molecules composed of approximately 20–25 nucleotides. The first miRNA, lin-4, was identified in *Caenorhabditis elegans* [10]. MiRNAs also regulate the expression of target genes by binding to mRNAs, playing a crucial role in several processes, such as cellular differentiation, development, and disease progression. CircRNAs play a unique role in regulating gene expression and cellular processes due to their distinctive circular structure and diverse functional modes. They can also adsorb miRNAs, thereby regulating gene expression [7, 11]. CircRNAs can act as miRNA sponges, blocking the degradation of miRNAs on downstream target genes. Therefore, circRNAs and miRNAs are closely connected [12–16], and determining these circRNA–miRNA associations (CMAs) is essential. Biological experiments are traditional ways of identifying novel CMAs; however, they always have low efficiency and high cost. Given the large number of circRNAs and miRNAs, biological experiments are no longer suitable for determining the associations. Therefore, designing quick and cheap techniques to detect CMAs is urgently needed.

An increasing number of circRNA and miRNA properties have been discovered and collected in public databases, such as circBank [17], CircBase [18], and miRbase [19], providing strong data support for identifying novel CMAs. Many advanced computational methods have also been designed to efficiently process the data of circRNAs and miRNAs. Designing computational methods to identify CMAs has become feasible because of the above materials and methods. To our knowledge, some computational methods have already been proposed. Fang and Lei [20] proposed the model KRWRMC to predict CMAs. It establishes a heterogeneous circRNA–miRNA network and employs a K-nearest neighbor algorithm based on the random walk restart heterogeneous to predict potential associations between circRNAs and miRNAs. The other methods are all based on machine learning and designed in a similar way. They construct some networks for circRNAs, miRNAs, or both and apply existing feature extraction and fusion algorithms, such as graph neural networks [21–23], network embedding algorithms [24–27], word embedding [27], or newly designed algorithms [28–30], to generate informative circRNA and miRNA features. These features are then fed into downstream prediction engines, including LightGBM [21, 29, 30], inner product operation [22, 23], weighted neighborhood regularized logistic matrix factorization [24], deep neural network [25, 27], inductive matrix completion algorithm [26], and gradient boosting decision tree [28], to make a prediction. Research on CMA prediction is still at the early stage, and the performance of the above models is mediocre, with the area under the receiver operating characteristic (ROC) curve not exceeding 0.95. Hence, there is a great space for improvement.

In this study, we proposed a novel computational model named CMAGN for predicting CMAs. The model first constructed two similarity networks based on the sequences of circRNAs and miRNAs. Together with the raw representations derived from the circRNA–miRNA adjacency matrix, these networks were fed into a graph attention

autoencoder (GATE) [31] to access the first representations of circRNAs and miRNAs. Then, node2vec [32] was applied to the CMA network to generate the second representations of circRNAs and miRNAs. With these novel representations, the similarity networks for circRNAs and miRNAs were reconstructed. Finally, network consistency projection (NCP) was applied to the reconstructed similarity networks and the CMA network to predict latent CMAs. Five-fold cross-validation revealed that the area under ROC and precision–recall (PR) curves on two widely used CMA datasets exceed 0.96, which is higher than those of previous models. Additional tests were conducted to elaborate the reasonability of each part of the proposed model and uncover its strengths and weaknesses.

## Materials and methods

### Dataset

The original CMAs were retrieved from the circBank database (http://www.circbank.cn/) [17], a comprehensive database containing human annotated circRNA from different sources. The file named "circbank_miRNA_all_v1.zip" comprised a large number of CMAs, with each association containing one circRNA, one miRNA, and one score. Following the same operation in previous studies [21–23, 25, 28–30], associations with scores no less than 1000 were selected, yielding 9589 CMAs covering 2115 different circRNAs and 821 different miRNAs. This dataset is same as those used in previous studies [21–23, 25, 28–30].

Efficient computational methods can be designed on the basis of these known CMAs. For formulation, these associations can be represented by a circRNA–miRNA adjacency matrix denoted by $AS \in \Re^{m\times n}$, where $m$ is the number of miRNAs and $n$ is the number of circRNAs ($m=821$ and $n=2115$ in this study). If the $i$-th miRNA and $j$-th circRNA constituted an association, then $AS(i,j)$ was set to 1; otherwise, it was defined as 0. The presence of zeros in $AS$ did not necessarily indicate the absence of association between the corresponding circRNA and miRNA; rather, it meant that the association has not been observed or confirmed at present. The goal of this study was to design an efficient computational method to pinpoint latent CMAs in $AS$. In addition to the adjacency matrix, bipartite graph is a common way to represent CMAs. This graph has two node groups—one containing the circRNA nodes and the other consisting of the miRNA nodes. Each edge in this graph indicates a CMA, i.e., one circRNA node and one miRNA node are connected if and only if they can constitute an association. In the following text, this bipartite graph was denoted by $N_{CM}$ and called the CMA network.

Another CMA dataset collected by Wang et al. [25] was employed in the present work, which can be obtained at https://github.com/1axin/KGDCMI. This dataset has 9905 CMAs, including 2346 cirRNAs and 962 miRNAs, and has also been used in several CMA prediction studies [23, 27–30]. Here, it was adopted to further test the performance of CMAGN. For a clear distinction between the two datasets, this dataset was labeled as CMA-9905, and the above-mentioned dataset was labeled as CMA-9589.

### Similarity network construction

The similarity between circRNAs or miRNAs is essential information to construct prediction methods; similar circRNAs are always associated with similar miRNAs, and vice

versa. The most basic similarity between circRNAs or miRNAs can be discerned from their sequences. Here, the Levenshtein distance was employed to measure the difference between two sequences and then calculate the similarity between them.

CircRNA and miRNA sequences were downloaded from three databases: CircFunBase (https://bis.zju.edu.cn/CircFunBase/) [33], CircBase (http://www.circbase.org/) [18], and miRbase (http://www.mirbase.org/) [19]. The Levenshtein distance is an edit distance commonly used to measure dissimilarity between two strings and is defined as the minimum number of edit operations required to transform the source string into the target string, allowing only three single-character operations: insertion, deletion, and substitution. In this study, the circRNA and miRNA sequences were deemed as strings to generate the Levenshtein distance between two circRNAs or two miRNAs. For formulation, the distance between two circRNAs $c_i$ and $c_j$ was denoted by $D_{circRNA}(c_i, c_j)$, and that between two miRNAs $m_i$ and $m_j$ was represented by $D_{miRNA}(m_i, m_j)$. Afterward, the sequence similarity of two circRNAs or miRNAs can be computed as follows:

$$CSS(c_i, c_j) = \frac{l(c_i) + l(c_j) - D_{circRNA}(c_i, c_j)}{l(c_i) + l(c_j)} \tag{1}$$

$$MSS(m_i, m_j) = \frac{l(m_i) + l(m_j) - D_{miRNA}(m_i, m_j)}{l(m_i) + l(m_j)} \tag{2}$$

where $CSS \in \Re^{n \times n}$ is the sequence similarity matrix of circRNAs; $MSS \in \Re^{m \times m}$ is the sequence similarity matrix of miRNAs; $l(c_i)$ and $l(c_j)$ are the sequence length of circRNAs $c_i$ and $c_j$, respectively; and $l(m_i)$ and $l(m_j)$ are the sequence length of miRNAs $m_i$ and $m_j$, respectively.

The similarity networks for circRNAs and miRNAs were constructed on the basis of the matrices *CSS* and *MSS*, respectively, and 2115 circRNAs and 821 miRNAs were defined as nodes. The edges in these two networks were determined by the corresponding elements in *CSS* and *MSS*. Threshold $T$ was set to *CSS* and *MSS* to discard the linkages with weak similarities. In *CSS*, if the element at the $i$-th row and $j$-th column was larger than the threshold, then the $i$-th and $j$-th circRNAs were connected by an edge. Under this operation, a similarity network for circRNAs was formed and denoted by $N_{\text{circRNA}}$. The network for miRNAs indicated by $N_{\text{miRNA}}$ was also constructed using the same operation.

#### Representations of circRNAs and miRNAs

Correct and complete representations of circRNAs and miRNAs are essential for building efficient models to predict CMAs. Here, each representation contained two parts. The first part was derived from the adjacency matrix *AS*, where each row indicated the raw representation of one miRNA and each column stood for the raw representation of one circRNA. These raw representations were refined by GATE [31] to access the high-level features of circRNAs and miRNAs. The second part was obtained from the CMA network $N_{CM}$ via a popular network embedding algorithm, node2vec [32]. These two parts contained different types of information for circRNAs and miRNAs. The first part focused on the local relationships of circRNAs or miRNAs because the raw representations contained only the information of the direct neighbors of circRNAs or miRNAs.

For the second part, the entire topological structure of the CMA network $N_{CM}$ was considered, including the global relationships of circRNAs or miRNAs. The combination of these two parts can fully represent circRNAs and miRNAs.

#### *Representations of circRNAs and miRNAs yielded by GATE*

The raw feature vectors of circRNAs derived from the adjacency matrix $AS$ were rudimentary and can be improved by some advanced computational methods. The similarity network $N_{\text{circRNA}}$ in Sect. "Similarity network construction" indicated the relationships between circRNAs. Fusing the above two forms can access to the high-level features of circRNAs. The same problem occurred for miRNAs. In this study, GATE [31] was employed to reconstruct the feature vectors of circRNAs and miRNAs by using the raw feature vectors of circRNAs (miRNAs) derived from $AS$ and circRNA (miRNA) similarity network as the input. The reconstructed feature vectors contained essential information regarding the raw feature vectors and the topological structure of the similarity networks.

GATE is an unsupervised learning method that can learn new representations of nodes from the topological structure of a network and the raw representations of its nodes. It involves two procedures, namely, encoder and decoder, which are described briefly as follows:

*Encoder* The encoder comprises several layers. The first layer adopts the raw features of nodes as input, and other layers use the output of the former layer as input. For clear descriptions, the raw feature vector of the $i$-th node is denoted by $F_i = h_i^{(0)}$, and the output of the $k$-th layer for the $i$-th node is represented by $h_i^{(k)}$. The main idea of GATE is to aggregate the feature vectors of neighbors for a given node to generate the new feature vector of the node. However, it further considers the weights of neighbors, that is, it assigns different weights to different neighbors. To determine the weights, GATE first calculates the correlation between any two nodes as

$$b_{ij}^{(k)} = Sigmoid\left(V_s^{(k)^T}\sigma\left(W^{(k)}h_i^{(k-1)}\right) + V_r^{(k)^T}\sigma\left(W^{(k)}h_j^{(k-1)}\right)\right) \tag{3}$$

where $W^{(k)} \in \Re^{d(k)\times d(k-1)}$, $V_s^{(k)} \in \Re^{d(k)}$, and $V_r^{(k)} \in \Re^{d(k)}$ are three matrices that consist of trainable parameters at the $k$-th layer; $d(k)$ is the dimension of the output feature vectors of the $k$-th layer; and $\sigma$ is an activation function. The above correlations were then normalized by the softmax function defined by

$$a_{ij}^{(k)} = Softmax(b_{ij}^{(k)}) = \frac{\exp(b_{ij}^{(k)})}{\sum_{l\in N(i)}\exp(b_{il}^{(k)})} \tag{4}$$

where $N(i)$ stands for the closed neighbor set of the $i$-th node. The outcomes of **Eq.** 4 were regarded as the neighbors' weights. The aggregation for the $i$-th node was conducted as

$$h_i^{(k)} = \sum_{l\in N(i)} a_{il}^{(k)}\sigma(W^{(k)}h_l^{(k-1)}) \tag{5}$$

where $W^{(k)}$ is same as that in **Eq.** 3. If the encoder contains $L$ layers, then $h_i^{(L)} = h_i$ is the output of the encoder for the $i$-th node. In this study, the feature vectors of circRNAs (miRNAs) derived from adjacency matrix $AS$ were used as the raw feature vectors of nodes and fed into the encoder procedure of GATE. The output of the last layer was selected as the novel feature vectors of circRNAs (miRNAs) and used in the following step to construct the model. The reconstructed feature vector for the $i$-th circRNA and miRNA is denoted by $C_i^g$ and $M_i^g$, respectively.

*Decoder* As an unsupervised learning method, GATE contains the decoder that recovers the raw features of nodes, thereby testing the quality of reconstructed feature vectors yielded by the encoder. The decoder consists of the same number of layers as the encoder. The output of the encoder is selected as the input of the decoder, denoted by $\tilde{h}_i^{(L)}$, and the output of the $k$-th layer is indicated by $\tilde{h}_i^{(k-1)}$. Here, $\tilde{h}_i^{(k-1)}$ was updated using the following equations:

$$\tilde{b}_{ij}^{(k)} = Sigmoid\left(\tilde{V}_s^{(k)^T}\sigma\left(\tilde{W}^{(k)}\tilde{h}_i^{(k)}\right) + \tilde{V}_r^{(k)^T}\sigma\left(\tilde{W}^{(k)}\tilde{h}_j^{(k)}\right)\right) \tag{6}$$

$$\tilde{a}_{ij}^{(k)} = Softmax(\tilde{b}_{ij}^{(k)}) = \frac{\exp(\tilde{b}_{ij}^{(k)})}{\sum_{l\in N(i)}\exp(\tilde{b}_{il}^{(k)})} \tag{7}$$

$$\tilde{h}_i^{(k-1)} = \sum_{l\in N(i)} \tilde{a}_{il}^{(k)}\sigma(\tilde{W}^{(k)}\tilde{h}_l^{(k)}) \tag{8}$$

where $\tilde{W}^{(k)} \in \Re^{d(k-1)\times d(k)}$, $\tilde{V}_s^{(k)} \in \Re^{d(k-1)}$, and $\tilde{V}_r^{(k)} \in \Re^{d(k-1)}$ are three matrices that contain trainable parameters; and $d(k-1)$ is the dimension of the output feature vectors of the $k$-th layer. The outcome of the last layer for the $i$-th node is defined as $\tilde{F}_i$, i.e., $\tilde{F}_i = \tilde{h}_i^{(0)}$.

*Loss Function* The whole GATE uses the raw feature vector of the $i$-th node $F_i$ as input. The encoder reconstructs $F_i$ as $h_i$, and the decoder recovers $h_i$ into $\tilde{F}_i$. GATE considers two types of loss to examine the quality of the reconstructed feature vector $h_i$. The first is the distance between $F_i$ and $\tilde{F}_i$ measured by

$$\sum_{i=1}^{N}\left\|F_i - \tilde{F}_i\right\|_2 \tag{9}$$

where $N$ is the total number of nodes in the network. The connections in the network should also be considered, and the neighboring nodes should be assigned to similar feature vectors. The second type of loss is formulated as

$$-\sum_{i=1}^{N}\sum_{l\in N(i)}\log\left(\frac{1}{1+\exp(-h_i^T h_l)}\right) \tag{10}$$

where $N(i)$ stands for the closed neighbor set of the $i$-th node. Accordingly, the final loss function combines the losses in **Eqs.** 9 and 10 as

$$\text{Loss} = \sum_{i=1}^{N} (\left\| F_i - \tilde{F}_i \right\|_2 - \lambda \sum_{l \in N(i)} \log(\frac{1}{1 + \exp(-h_i^T h_l)})) \tag{11}$$

where $\lambda$ is a parameter balancing the two types of loss, $F_i$ is the raw feature vector of the $i$-th node, $\tilde{F}_i$ is the recovered feature vector (output of the decoder procedure) of the $i$-th node, $h_i$ is the new feature vector of the $i$-th node (output of the encoder procedure), and $h_l$ is the new feature vector of the $l$-th node that is the neighbor of the $i$-th node. New high-quality representations of nodes can be accessed by minimizing the loss.

#### *Representations of circRNAs and miRNAs yielded by node2vec*

The above representation of circRNAs (miRNAs) were derived from their local relationships to miRNAs (circRNAs) and the circRNA (miRNA) similarity network. However, the global relationships of circRNAs to miRNAs were ignored. The same problem occurred for miRNAs. Therefore, additional information was employed to evaluate the global associations between circRNAs and miRNAs. Known CMAs are essential for predicting novel associations. Thus, using them to measure the associations between circRNAs and miRNAs can help in setting up efficient models. As mentioned in Sect. "Dataset", the known CMAs are represented by the CMA network $N_{CM}$, from which the representation of circRNAs and miRNAs can be accessed. The powerful network embedding algorithm, node2vec [32], was employed for this task.

Node2vec is the improved version of DeepWalk [34]. For a given network, it produces many paths starting from each node. Suppose the starting node is $u$ and the current endpoint of the path starting from $u$ is $n_{i\text{-}1}$, which is the ($i$-1)-th node in this path. This path is extended to the $i$-th node, denoted by $n_i$ by selecting one neighbor of $n_{i\text{-}1}$. However, the selection probability is not equal. The probability is determined using the following equation:

$$P(n_i = w | n_{i-1} = v) = \begin{cases} \pi_{vw}/Z & \textit{if } w \text{ is adjacent to } v \\ 0 & \textit{Otherwise} \end{cases} \tag{12}$$

,

where $\pi_{vw}$ is the transition probability from $v$ to $w$ and computed as

$$\pi_{vw} = \alpha_{pq}(t, w) \cdot w_{vw} \tag{13}$$

$$\alpha_{pq}(t, w) = \begin{cases} 1/p & \textit{if } d_{tw} = 0 \\ 1 & \textit{if } d_{tw} = 1 \\ 1/q & \textit{if } d_{tw} = 2 \end{cases} \tag{14}$$

where $w_{vw}$ stands for the weight on edge connecting $v$ and $w$, $t$ is the ($i$-2)-th node in the path, and $d_{tw}$ denotes the distance between $t$ and $w$. The symbol $Z$ in **Eq. 12** is the sum of the transition probabilities from $v$ to its neighbors. Evidently, the paths sampled by node2vec are produced by biased selection. The next node is not selected with equal probability. This operation can efficiently capture the structure traits of the network. The selection of the next node is continuously executed until the current path reaches the

predefined length. After the paths starting from each node have been sampled, these paths are deemed as sentences. Meanwhile, the nodes in the paths are considered as words. This information is fed into word2vec with skip-gram to produce the feature vectors of the nodes.

This study applied the node2vec program downloaded from https://snap.stanford.edu/node2vec/ to the CMA network $N_{CM}$ to generate the feature vectors of circRNAs and miRNAs. For formulation, the feature vectors of the $i$-th circRNA and miRNA are denoted by $C_i^n$ and $M_i^n$, respectively.

For the $i$-th circRNA, two representations $C_i^g$ and $C_i^n$ were obtained and then combined to generate the final representation of the circRNA. The same operation was conducted for each miRNA. The final representations of the $i$-th circRNA and miRNA, denoted by $\hat{C}_i$ and $\hat{M}_i$, respectively, were formulated as

$$\hat{C}_i = [C_i^g, C_i^n] \tag{15}$$

$$\hat{M}_i = [M_i^g, M_i^n] \tag{16}$$

#### Reconstruction of similarity networks

With these new representations of circRNAs and miRNAs, their corresponding informative similarity network can be built. Here, cosine kernel was adopted to measure the associations between circRNAs (miRNAs) on the basis of their new representations. For circRNAs $c_i$ and $c_j$, their similarity was reformulated as.

$$CS(c_i, c_j) = \frac{\hat{C}_i \cdot \hat{C}_j^T}{\left|\hat{C}_i\right| \cdot \left|\hat{C}_j\right|}, \tag{17}$$

where $CS$ is the new circRNA similarity matrix collecting the outcomes of **Eq. 17**. The circRNA similarity network was built using $CS$ as the weighted adjacency matrix.

The similarity network for miRNAs was reconstructed in the same way. For miRNAs $m_i$ and $m_j$, their similarity was measured as follows:

$$MS(m_i, m_j) = \frac{\hat{M}_i \cdot \hat{M}_j^T}{\left|\hat{M}_i\right| \cdot \left|\hat{M}_j\right|} \tag{18}$$

where $MS$ is the miRNA similarity matrix that stores the outcomes of **Eq. 18**. The miRNA similarity network was then established using $MS$ as the weighted adjacency matrix.

#### NCP

With the above circRNA similarity network (represented by $CS$), miRNA similarity network (represented by $MS$), and CMA network (represented by adjacency matrix $AS$), a simple and efficient network-based method named NCP was utilized to calculate the

association score for circRNAs and miRNAs. NCP has wide applications in association predictions [35–39]. The score yielded by NCP integrates two subscores, which are obtained by projecting the respective circRNA and miRNA similarity networks onto the CMA network. For circRNA $c_i$ and miRNA $m_j$, the subscore for projecting the circRNA similarity network onto the CMA network was calculated as.

$$CSP(c_i, m_j) = \frac{AS(i,:) \times CS(:,j)}{|AS(i,:)|} \tag{19}$$

where $AS(i,:)$ represents the $i$-th row of $AS$, $CS(:,j)$ denotes the $j$-th column of $CS$, and $|AS(i,:)|$ is the length of the vector $AS(i,:)$. Meanwhile, the subscore for projecting the miRNA similarity network onto the CMA network was computed as.

$$MSP(c_i, m_j) = \frac{MS(i,:) \times AS(:,j)}{\left|AS(:,j)\right|} \tag{20}$$

where $MS(i,:)$ is the $i$-th row of $MS$, $AS(:,j)$ indicates the $j$-th column of $AS$, and $\left|AS(:,j)\right|$ indicates the length of $AS(:,j)$. Finally, NCP integrated the above subscores as follows:

$$NSP(c_i, m_j) = \frac{CSP(c_i, m_j) + MSP(c_i, m_j)}{|CS(i,:)| + \left|MS(:,j)\right|} \tag{21}$$

where $NSP$ is the final recommendation matrix collecting all the outcomes of **Eq. 21**.

#### Outline of CMAGN

This work designed a new model named CMAGN for predicting CMAs. The architecture of this model is illustrated in Fig. 1. CMAGN consists of several modules. First, the similarity networks for circRNAs and miRNAs were constructed on the basis of their sequence similarities in module A as illustrated in Fig. 1(A). Second, the raw feature vectors of circRNAs (miRNAs) were extracted from the circRNA–miRNA adjacency matrix and fed into GATE to access new representations of circRNAs (miRNAs) in module B, as illustrated in Fig. 1(B). Third, the feature vectors of circRNAs and miRNAs were generated by applying node2vec to CMA network in module C, as displayed in Fig. 1(C). Fourth, two representations of circRNAs (miRNAs) were combined to reconstruct the circRNA (miRNA) similarity network in module D, as shown in Fig. 1(D). Finally, NCP was applied to the reconstructed similarity networks and the CMA network to yield the recommendation matrix in module E, as shown in Fig. 1(E).

## Results and discussion

### Evaluation metrics

In this study, five-fold cross-validation [40] was adopted to evaluate the performance of the models. When conducting the cross-validation, we denoted 9589 validated CMAs as positive samples. Meanwhile, the negative samples were randomly selected

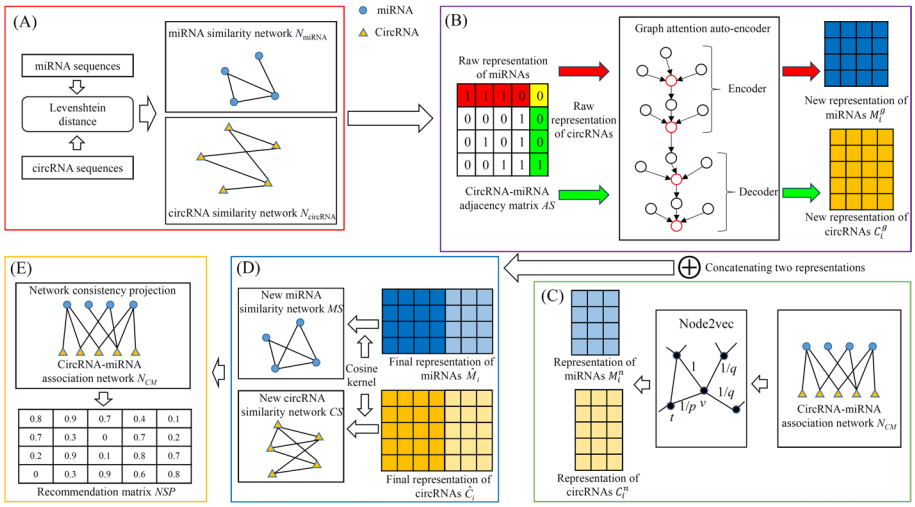



**Fig. 1** Entire procedures of CMAGN. **A** The circRNA (miRNA) similarity network is built on the basis of sequence similarities between them. **B** The similarity network and raw representations of circRNAs (miRNAs) derived from the adjacency matrix are fed into the GATE to generate new representations of circRNAs (miRNAs). **C** Representations of circRNAs (miRNAs) are extracted from the CMA network via node2vec. **D** The above two representations are combined to encode circRNAs (miRNAs), which are used to reconstruct circRNA (miRNA) similarity network. **E** The reconstructed similarity networks and association network are analyzed by NCP to produce a recommendation matrix

from unlabeled associations, and their number was the same as that of the positive samples. All samples were randomly and equally divided into five sets. Each set was selected as a test set one by one, and the remaining sets constituted the training set. In every round of the five-fold cross-validation, the corresponding positions of the singled-out positive samples in the adjacency matrix $AS$ were replaced with zero, and the corresponding edges in the CMA network were removed.

ROC and PR curves were adopted to assess the cross-validation results [41–44]. For this task, a group of thresholds for predicting positive samples should be set. For a given threshold, if the association score is higher than the threshold, then the corresponding sample is predicted to be positive; otherwise, it is predicted to be negative. On the basis of such results, the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) can be counted. Subsequently, the true positive rate (TPR, same as recall), false positive rate (FPR), and precision [45–48] can be computed as

$$\begin{cases} TPR = \dfrac{TP}{TP + FN} \\ FPR = \dfrac{FP}{FP + TN} \\ Precision = \dfrac{TP}{TP + FP} \end{cases} \tag{22}$$

After a set of thresholds are set, a group of TPR, FPR, and precision values can be obtained. The ROC curve can be plotted by setting TPR as the Y-axis and FPR as the X-axis, and the PR curve is plotted by defining precision as the Y-axis and recall as the X-axis. The area under these two curves are key measurements to evaluate a model's

performance, which were denoted as AUROC and AUPR in this study. In general, a high AUROC (AUPR) corresponds to high performance.

### Parameter settings

CMAGN has several parameters distributed in raw similarity network construction, GATE, and node2vec. These parameters were determined as follows, with the final settings listed in Table 1.

In the construction of the raw similarity network of circRNAs (miRNAs), threshold $T$ was employed to determine the associations between circRNAs (miRNAs) according to their sequences as mentioned in Sect. "Similarity network construction". As suggested in [49], $T$ was set to 0.8.

In GATE, the numbers of layers in the encoder and decoder were the two key parameters and were generally set to two, that is, two layers in the encoder and decoder. The sizes of each layer are also important. We set 128 and 64 neurons in the two encoder layers and corresponding decoder layers. The learning rate was set to 0.001. The parameter dropout, which is a probability to temporarily inactive neurons, was set to 0.5. Finally, the parameter $\lambda$ in the loss function (**Eq.** 11), which balances the contributions of two types of loss, was set to its default value of 1.

When using node2vec to access the second representations of circRNAs and miRNAs, we set the feature dimension as 128, which is the default value in node2vec. The other parameters were also set to their default values: the number of paths starting from each node was set to 10, the length of the path was set to 80, and parameters $p$ and $q$ were set to 1.

### Performance of CMAGN on the CMA-9589 dataset

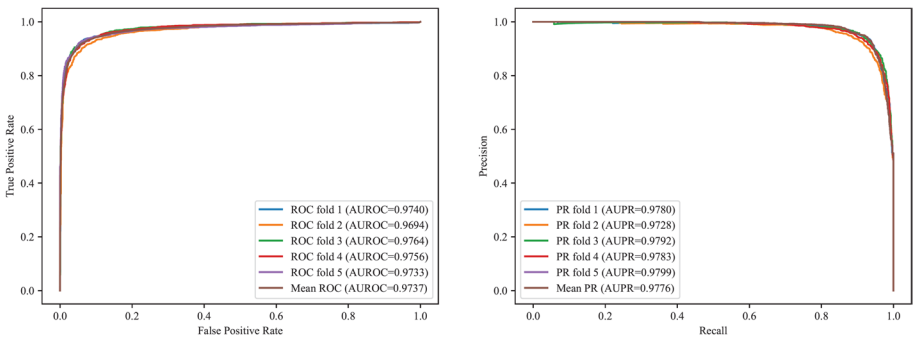
According to Sect. "Parameter settings", the selected parameters of CMAGN were set, and this model was evaluated by fivefold cross-validation on the CMA-9589 dataset. Table 2 lists the AUROC and AUPR under each fold, along with the mean AUROC and AUPR. The ROC and PR curves are illustrated in Fig. 2. The AUROC values across the five folds were 0.9740, 0.9694, 0.9764, 0.9756, and 0.9733, with a mean AUROC of 0.9737. The AUPR values across the five folds were 0.9780, 0.9728, 0.9792, 0.9783, and

**Table 1** Parameter settings of the CMAGN model

| Module | Parameter | Setting |
|---|---|---|
| Similarity network construction | Threshold $T$ | 0.8 |
| GATE | Number of layers | 2 |
| | Number of neurons | 128, 64 |
| | Learning rate | 0.001 |
| | Dropout | 0.5 |
| | $\lambda$ | 1 |
| Node2vec | Dimension | 128 |
| | Number of paths | 10 |
| | Length of paths | 80 |
| | $p$ | 1 |
| | $q$ | 1 |

**Table 2** Five-fold cross-validation results of CMAGN on the CMA-9589 dataset

| Measurement | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean |
|---|---|---|---|---|---|---|
| AUROC | 0.9740 | 0.9694 | 0.9764 | 0.9756 | 0.9733 | 0.9737 |
| AUPR | 0.9780 | 0.9728 | 0.9792 | 0.9783 | 0.9799 | 0.9776 |




(A) (B)

**Fig. 2** ROC and PR curves to illustrate the performance of CMAGN on the CMA-9589 dataset. **A** ROC curves; **B** PR curves. The ROC and PR curves are nearly perfect, and the AUROC and AUPR values vary within a small range

**Table 3** Comparison of the models when one part is removed or replaced

| Original item | Replacement item/Removed | AUROC | AUPR |
|---|---|---|---|
| GATE + Node2vec | Removed | 0.9019 | 0.9026 |
| | GATE | 0.9346 | 0.9546 |
| | Node2vec | 0.9678 | 0.9711 |
| Cosine kernel | GIP kernel | 0.9659 | 0.9715 |
| | Jaccard kernel | 0.8788 | 0.8769 |
| Network consistency projection | KATZ | 0.9567 | 0.9599 |
| | WKNKN | 0.9584 | 0.9625 |

0.9799, with a mean AUPR of 0.9776. These results indicate that CMAGN performs well in predicting CMAs and exhibits almost the same performance on different folds (similar AUROC and AUPR values on different folds), suggesting that it was stable throughout the validation.

### Utility of node2vec and GATE

CMAGN adopted two powerful methods (node2vec and GATE) to access the representations of circRNAs and miRNAs. We conducted some ablation tests to assess their key constructions for building the CMAGN.

First, node2vec and GATE were removed from CMAGN. The raw similarity networks of circRNAs and miRNAs derived from their sequences were directly fed into NPC to yield the recommendation matrix. This model was evaluated by five-fold cross-validation. The results are listed in Table 3, and the corresponding ROC and PR curves are illustrated in Fig. 3. The AUROC and AUPR of this model were 0.9019 and 0.9026,

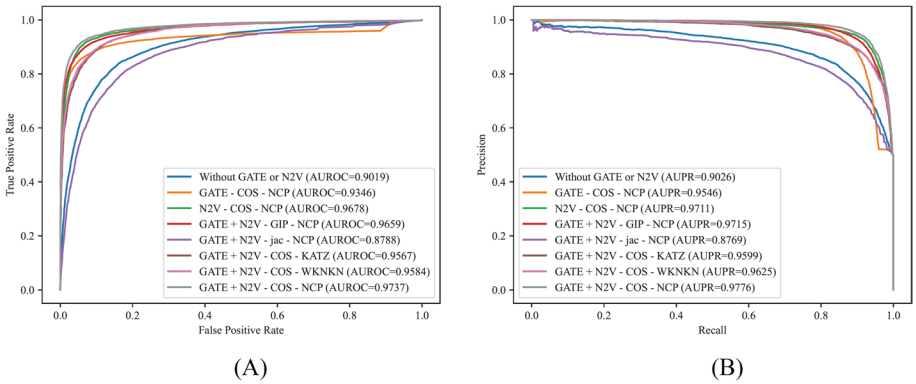



**Fig. 3** ROC and PR curves to illustrate the performance of CMAGN when one part is removed or replaced. **A** ROC curves; **B** PR curves. The model "Without GATE or N2V" indicates the model by removing GATE and node2vec (N2V) from CMAGN. Three parts of CMAGN are considered, namely, GATE + N2V, cosine (COS) kernel, and NCP, in other models, which are removed or replaced with other methods. Their names in cutline consist of the used methods, connected by the symbol "-". Each part provides key contributions for building CMAGN because removal or replacement decreases the performance

respectively, which were lower than those of CMAGN. The ROC and PR curves were always under those of CMAGN, and gaps between the ROC curves or PR curves of the two models were evident. All these findings suggested that this model exhibits lower performance than CMAGN, proving the valuable contributions of node2vec and GATE. These two methods aid the construction of accurate similarity networks of circRNAs and miRNAs, thereby supporting NCP in yielding an accurate recommendation matrix.

Second, we removed node2vec from CMAGN. In this case, the similarity networks of circRNAs and miRNAs were constructed only from the representations yielded by GATE. The model was also evaluated by five-fold cross-validation. The AUROC and AUPR for this model are listed in Table 3, and the corresponding ROC and PR curves are illustrated in Fig. 3. The AUROC and AUPR values were 0.9346 and 0.9546, respectively, which are lower than those of CMAGN. This finding suggests that the removal of node2vec decreases the performance of CMAGN. Furthermore, the ROC and PR curves were always under those of CMAGN, further confirming the above conclusion. All these results indicated the importance of node2vec.

Third, the importance of GATE was investigated in a similar way: by removing it from CMAGN. The similarity networks were set up using only the features yielded by node2vec. Five-fold cross-validation yielded the AUROC of 0.9678 and AUPR of 0.9711 as listed in Table 3, which are also lower than those of CMAGN. The ROC and PR curves in Fig. 3 further confirmed that the removal of GATE weakens the performance. These results suggest the importance of GATE.

On the basis of the above arguments, node2vec and GATE provide essential contributions for building CMAGN. This result was reasonable because they generated the features of circRNAs and miRNAs from different aspects. GATE generated the features focusing on the local relationships of circRNAs (miRNAs) to miRNAs (circRNAs) and the relationships between circRNAs (miRNAs), and node2vec produced the features containing the global relationships of circRNAs (miRNAs) to miRNAs (circRNAs). The combination of these features can fully describe the essential information of circRNAs

(miRNAs), allowing the model to exhibit high performance. Compared with that when GATE was removed, the model showed lower AUROC and AUPR when node2vec was removed, implying that node2vec was more important than GATE. Accordingly, the features yielded by node2vec can capture more essential properties of circRNAs and miRNAs compared with those yielded by GATE.

#### Utility of cosine kernel when reconstructing similarity networks

The cosine kernel was adopted to measure the linkage of circRNAs (miRNAs) when reconstructing similarity networks. To prove that this selection is reasonable, we replaced it with other two popular kernels: Gaussian interaction profile (GIP) and Jaccard kernels. The models using these two kernels were assessed by five-fold cross-validation. As illustrated in Fig. 3, the obtained ROC and PR curves were always under the corresponding curves of CMAGN, which used cosine kernel. The AUROC and AUPR are listed in Table 3. The model using GIP kernel yielded an AUROC of 0.9659 and an AUPR of 0.9715. Using the Jaccard kernel generated lower AUROC and AUPR of 0.8788 and 0.8769, respectively. All these values were lower than those of CMAGN, proving that using the cosine kernel to measure the similarity of circRNAs (miRNAs) is reasonable.

#### Utility of NCP

In CMAGN, the recommendation matrix was generated by NCP. We also used other powerful methods to access the recommendation matrix, including KATZ [50] and weighted K-nearest known neighbor [51]. The models that use these methods to generate recommendation matrix were constructed and evaluated by five-fold cross-validation. The cross-validation results are listed in Table 3. The AUROC values for these models were 0.9567 and 0.9584, and the AUPR values were 0.9599 and 0.9625. All these values were lower than those of the model that uses NCP to produce the recommendation matrix (i.e., CMAGN). The ROC and PR curves illustrated in Fig. 3 also verified the relatively low performance of these two models. These results confirmed the utility value of NCP.

#### Analysis of CMAGN for different circRNA–miRNA associations

CMAGN exhibited a high overall performance. However, its performance for different CMAs may not be the same, that is, CMAGN can provide reliable predictions for some CMAs but unsatisfying predictions for other CMAs. To uncover the strengths and weaknesses of CMAGN in this regard, we equally divided the circRNAs (miRNAs) into two groups according to the number of CMAs that they involved. The first group contained circRNAs (miRNAs) with high numbers, and the second group contained the rest of the circRNAs (miRNAs) with low numbers. For convenience, these two groups were called high and low groups. Accordingly, all the CMAs were divided into four groups, namely, high (circRNA)–high (miRNA), high (circRNA)–low (miRNA), low (circRNA)–high (miRNA), and low (circRNA)–low (miRNA) groups. The high (circRNA)–high (miRNA) group consisted of CMAs involving the circRNAs and miRNAs from the high groups. The other three CMA groups were constructed in the same way. For the cross-validation results of CMAGN, the AUROC and AUPR were individually counted for the above four CMA groups as listed in Table 4. The corresponding ROC and PR curves are displayed

**Table 4** Performance of CMAGN on different CMA groups

| CircRNA-miRNA association group | AUROC | AUPR |
|---|---|---|
| High (circRNA)-high (miRNA) group | 0.9792 | 0.9867 |
| High (circRNA)-low (miRNA) group | 0.9779 | 0.9858 |
| Low (circRNA)-high (miRNA) group | 0.9595 | 0.9483 |
| Low (circRNA)-low (miRNA) group | 0.9610 | 0.9471 |

in Fig. 4. The performance of CMAGN in the high (circRNA)–high (miRNA) and high (circRNA)–low (miRNA) groups was evidently higher than in the other two groups. The AUROC values on these two groups were higher than 0.97, and the AUPR values exceeded 0.98. By contrast, the AUROC and AUPR values for the other two groups were approximately 0.96 and 0.94, respectively. This finding indicated that the model yielded reliable predictions for circRNAs that can interact with many miRNAs. However, its predictions for circRNAs interacting with few miRNAs had relatively low credibility. For miRNAs, such influence was not as evident. Given the circRNAs in the same group, CMAGN's performance on the miRNAs in the high or low groups was almost the same. These results suggest that CMAGN is more sensitive to circRNAs than to miRNAs.

### Computation time analysis

According to Sect. "Outline of CMAGN", CMAGN has five modules (modules A–E), whose computation times must be analyzed to understand the bottleneck of this model. For one five-fold cross-validation, the computation time of the five modules is listed in Table 5. Module A needed the most time, occupying 91.6%. The function of module A is to construct miRNA and circRNA similarity networks. A similar calculation was conducted on the basis of the Levenshtein distance. In general, dynamic programming is

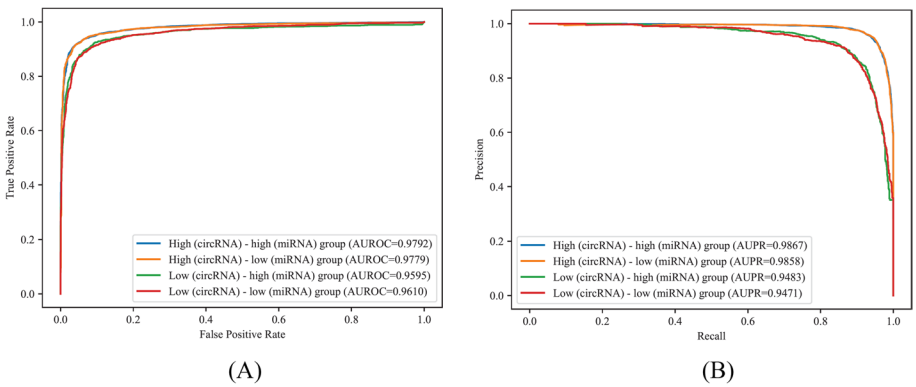



**Fig. 4** ROC and PR curves to illustrate the performance of CMAGN on different CMA groups. **A** ROC curves; **B** PR curves. The performance of CMAGN in the high (circRNA)-high (miRNA) and high (circRNA)-low (miRNA) groups is higher than in the other two groups

**Table 5** Computation time of each module in CMAGN[a]

| Module | Module A | Module B | Module C | Module D | Module E |
|---|---|---|---|---|---|
| Time | 5018.66 s | 131.12 s | 326.22 s | 0.32 s | 2.29 s |

a: The time listed in this table is obtained by a computer with CPU i7-12,700 and 40 GB memory. The modules A-E are explained in Sect. "Outline of CMAGN"

used to compute the Levenshtein distance between two strings. Its computational complexity is $O(m \times n)$, where $m$ and $n$ are the lengths of two strings. Given that most circRNAs are extremely long, computing the Levenshtein distance between two circRNAs is time consuming. This phenomenon was the main reason module A required so much time. Meanwhile, modules B and C contained GATE and node2vec, respectively. Their computation procedures can be parallelized and the input networks were not very large; hence, these two modules did not take much time. Modules D and E involved easy computation that required minimal time. Thus, reducing the computation time for the Levenshtein distance is the key to improving the efficiency of CMAGN.

#### Comparison of existing models

Several models have been proposed to predict CMAs, as introduced in Sect. "Introduction". We compared CMAGN with previous models to prove the superiority of our model. For pairwise comparison, we selected the following seven models set up on the CMA-9589 dataset and evaluated by five-fold cross-validation: CMASG [21], KGDCMI [25], GCNCMI [22], SGCNCMI [23], JSNDCMI [28], BioDGW-CMI [29], and BEROLECMI [30]. The AUROC and AUPR values of the above models, which were directly obtained from their corresponding papers, are listed in Table 6. GCNCMI, JSNDCMI, BioDGW-CMI, and BEROLECMI outperformed the other three models, yielding AUROC and AUPR values higher than 0.93. KGDCMI and SGCNCMI were in the second rank, generating AUROC and AUPR of approximately 0.90. CMASG showed the lowest performance, with AUROC and AUPR lower than 0.89. For easy comparison, the performance of our model CMAGN is also listed in Table 6. CMAGN exhibited higher performance than all the above models, with AUROC and AUPR values that were all at least 2.4% higher than those of these previous models, proving its superiority. CMAGN adopted two representations of circRNAs and miRNAs to fully describe the relationships between circRNAs, miRNAs, and circRNAs–miRNAs. The similarity network of circRNAs (miRNAs) can widely measure their associations based on such representations. The accurate and powerful method NCP was applied to the circRNA and miRNA similarity networks to access a reliable recommendation matrix. All these steps resulted in the high performance of CMAGN.

A paired Student's t-test was also conducted on CMAGN and previous models. All the p-values listed in Table 7 were lower than the 0.05 confidence level, suggesting

**Table 6** Performance of different models for predicting CMAs on the CMA-9589 dataset

| Model | AUROC | AUPR |
|---|---|---|
| CMASG [21] | 0.8804 | 0.8629 |
| KGDCMI [25] | 0.9041 | 0.8937 |
| GCNCMI [22] | 0.9320 | 0.9396 |
| JSNDCMI [28] | 0.9415 | 0.9403 |
| SGCNCMI [23] | 0.9015 | 0.9011 |
| BioDGW-CMI [29] | 0.9476 | 0.9416 |
| BEROLECMI [30] | 0.9491 | 0.9431 |
| CMAGN | <u>0.9737</u> | <u>0.9776</u> |

**Table 7** Paired Student's t-test results of the CMAGN and other models on the CMA-9589 dataset under five-fold cross-validation

| Model | P-value | |
|---|---|---|
| | AUROC | AUPR |
| KGDCMI [25] | $4.85\times10^{-39}$ | $6.48\times10^{-40}$ |
| GCNCMI [22] | $7.79\times10^{-35}$ | $1.89\times10^{-33}$ |
| JSNDCMI [28] | $1.02\times10^{-32}$ | $2.68\times10^{-33}$ |
| SGCNCMI [23] | $2.42\times10^{-39}$ | $3.70\times10^{-39}$ |
| BioDGW-CMI [29] | $5.32\times10^{-31}$ | $5.20\times10^{-33}$ |
| BEROLECMI [30] | $1.62\times10^{-30}$ | $1.15\times10^{-32}$ |

**Table 8** Five-fold cross-validation results of CMAGN on the CMA-9905 dataset

| Measurement | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean |
|---|---|---|---|---|---|---|
| AUROC | 0.9614 | 0.9540 | 0.9616 | 0.9613 | 0.9674 | 0.9611 |
| AUPR | 0.9725 | 0.9655 | 0.9728 | 0.9717 | 0.9735 | 0.9712 |

significant differences between CMAGN and previous models. This finding proved that CMAGN performs significantly better than previous models do.

#### *Performance of CMAGN on the CMA-9905 dataset*

To fully examine the performance of CMAGN, we further employed another CMA dataset called CMA-9905, as mentioned in Sect. "Dataset". The five-fold cross-validation results of CMAGN on this dataset are shown in Table 8. The five AUROC values were 0.9614, 0.9540, 0.9616, 0.9613, and 0.9674, and the mean AUROC was 0.9611. The AUPR values across five folds were 0.9725, 0.9655, 0.9728, 0.9717, and 0.9735, yielding a mean AUPR of 0.9712. Such performance was slightly lower than that on the CMA-9589 dataset but was nevertheless still very high, proving the powerful prediction ability of CMAGN.

Previous CMA prediction models were also applied to the CMA-9905 dataset, including KGDCMI [25], WSCD [27], SGCNCMI [23], JSNDCMI [28], BioDGW-CMI [29], and BEROLECMI [30]. Their five-fold cross-validation results (AUROC and AUPR) are listed in Table 9. These previous models exhibited almost similar performance. In detail, the AUROC values changed between 0.89 and 0.92, and the AUPR values varied in the interval [0.87, 0.91]. BEROLECMI gave the highest performance among these previous methods. The performance of CMAGN is also provided in Table 9 for easy comparison. CMAGN evidently outperformed all the above previous models, with an AUROC that was at least 5% higher and an AUPR that was at least 6% higher. This superiority was more evident compared with that on the CMA-9589 dataset. We also conducted a paired Student's t-test on CMAGN and the above previous models on the CMA-9905 dataset. The results are provided in Table 10. Similar to the findings in Table 7, all p-values were lower than the 0.05 confidence level. This outcome also proved that CMAGN is significantly superior to the previous models

**Table 9** Performance of different models for predicting CMAs on the CMA-9905 dataset

| Model | AUROC | AUPR |
|---|---|---|
| KGDCMI [25] | 0.8930 | 0.8767 |
| WSCD [27] | 0.8923 | 0.8935 |
| SGCNCMI [23] | 0.8942 | 0.8887 |
| JSNDCMI [28] | 0.9003 | 0.8999 |
| BioDGW-CMI [29] | 0.9026 | 0.8962 |
| BEROLECMI [30] | 0.9104 | 0.9086 |
| CMAGN | **0.9611** | **0.9712** |

The number with bold and underline is the highest number in the corresponding column

**Table 10** Paired Student's t-test results of the CMAGN and other models on the CMA-9905 dataset under five-fold cross-validation

| Model | P-value | |
|---|---|---|
| | AUROC | AUPR |
| KGDCMI [25] | $1.33\times10^{-40}$ | $1.93\times10^{-43}$ |
| WSCD [27] | $1.09\times10^{-40}$ | $8.20\times10^{-42}$ |
| SGCNCMI [23] | $1.86\times10^{-40}$ | $2.60\times10^{-42}$ |
| JSNDCMI [28] | $1.16\times10^{-39}$ | $4.26\times10^{-41}$ |
| BioDGW-CMI [29] | $2.43\times10^{-39}$ | $1.61\times10^{-41}$ |
| BEROLECMI [30] | $3.78\times10^{-38}$ | $5.16\times10^{-40}$ |

on the CMA-9905 dataset. Combined with the previous results on CMA-9589, these findings indicated that CMAGN is currently the best model for predicting CMAs.

#### *Limitations of this study*

The proposed model CMAGN exhibited high performance in predicting CMAs. However, some limitations must be addressed. The circRNA–miRNA adjacency matrix $AS$ was directly fed into NCP to generate the recommendation matrix. The members in this matrix were either 0 or 1. Given that the strengths of different CMAs may not be the same, this representation was not perfect. Our model can be improved by employing quantitative representations of CMAs to the adjacency matrix. As a recommendation system, CMAGN can only predict novel CMAs from the involved circRNAs and miRNAs. For a circRNA or miRNA that is not in this system, CMAGN cannot identify its related CMAs. We had to discard the current model and rebuild a model that includes this circRNA or miRNA. The applicability of CMAGN was not good. In the future, we will expand this study by focusing on the above two aspects to design robust CMA prediction models.

## Conclusions

This study proposed a novel computational model called CMAGN to predict CMAs. It contained a complex procedure to extract informative representations of circRNAs and miRNAs, thereby building extensive similarity networks. These networks were then projected onto the CMA network by NCP to generate a reliable recommendation matrix. Ablation tests confirmed that the main parts of CMAGN are essential. The performance

of CMAGN was superior to that of existing models on two widely used CMA datasets, and it was more sensitive to circRNAs than miRNAs. This model can be a useful tool to identify potential CMAs. The codes and data used in this study can be accessed at https://github.com/brilliant-plus/CMAGN.

**Abbreviations**

| | |
|---|---|
| circRNA | Circular RNA |
| miRNA | MicroRNA |
| CMA | CircRNA-miRNA association |
| ncRNA | Non-coding RNA |
| GATE | Graph attention auto-encoders |
| NCP | Network consistency projection |
| ROC | Receiver operating characteristic |
| PR | Precision–recall |
| TP | True positive |
| FP | False positive |
| TN | True negative |
| FN | False negative |
| TPR | True positive rate |
| FPR | False positive rate |
| GIP | Gaussian interaction profile |

**Acknowledgements**
Not applicable.

**Author contributions**
L.C. and Y.D.C. designed the research; A.Y., L.C. and B.Z. conducted the experiments; A.Y. and B.Z. analyzed the results. All authors have read and approved the manuscript.

**Funding**
Not applicable.

**Availability of data and materials**
The original circRNA-miRNA associations in CMA-9589 dataset are available at circBank database (http://www.circbank.cn/) and those in CMA-9905 dataset are available at https://github.com/1axin/KGDCMI. The source codes and data in this study are available at https://github.com/brilliant-plus/CMAGN.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 22 January 2024 Accepted: 16 October 2024
Published online: 24 October 2024

**References**

1. Papageorgiou N, Tslamandris S, Giolis A, Tousoulis D. MicroRNAs in cardiovascular disease: perspectives and reality. Cardiol Rev. 2016;24(3):110–8.
2. Ha J. SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association. Knowl-Based Syst. 2023;263: 110295.
3. Ha J, Park C, Park C, Park S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. J Biomed Inform. 2020;102: 103358.
4. Chen X, Li TH, Zhao Y, Wang CC, Zhu CC. Deep-belief network for predicting potential miRNA-disease associations. Brief Bioinform. 2021;22(3):186. https://doi.org/10.1093/bib/bbaa186.
5. Ha J, Park S. NCMD: Node2vec-based neural collaborative filtering for predicting MiRNA-disease association. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(2):1257–68.
6. Chen L, Zhao X. PCDA-HNMP: Predicting circRNA-disease association using heterogeneous network and meta-path. Math Biosci Eng. 2023;20(12):20553–75.
7. Chen Y, Wang Y, Ding Y, Su X, Wang C. RGCNCDA: Relational graph convolutional network improves circRNA-disease association prediction by incorporating microRNAs. Comput Biol Med. 2022;143: 105322.

8. Deng L, Liu D, Li Y, Wang R, Liu J, Zhang J, Liu H. MSPCD: predicting circRNA-disease associations via integrating multi-source data and hierarchical neural network. BMC Bioinformatics. 2022;23(Suppl 3):427.
9. Yan C, Wang J, Wu FX. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. BMC Bioinformatics. 2018;19(Suppl 19):520.
10. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.
11. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014;56(1):55–66.
12. Zhang M, Huang N, Yang X, Luo J, Yan S, Xiao F, Chen W, Gao X, Zhao K, Zhou H, et al. A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis. Oncogene. 2018;37(13):1805–14.
13. He W, Shi X, Guo Z, Wang H, Kang M, Lv Z. Circ_0019693 promotes osteogenic differentiation of bone marrow mesenchymal stem cell and enhances osteogenesis-coupled angiogenesis via regulating microRNA-942-5p-targeted purkinje cell protein 4 in the development of osteoporosis. Bioengineered. 2022;13(2):2181–93.
14. Fan X, Yin X, Zhao Q, Yang Y. Hsa_circRNA_0045861 promotes renal injury in ureteropelvic junction obstruction via the microRNA-181d-5p/sirtuin 1 signaling axis. Ann Transl Med. 2021;9(20):1571.
15. Xie F, Li Y, Wang M, Huang C, Tao D, Zheng F, Zhang H, Zeng F, Xiao X, Jiang G. Circular RNA BCRC-3 suppresses bladder cancer proliferation through miR-182-5p/p27 axis. Mol Cancer. 2018;17(1):144.
16. Wang Z, Ma K, Pitts S, Cheng Y, Liu X, Ke X, Kovaka S, Ashktorab H, Smoot DT, Schatz M, et al. Novel circular RNA circNF1 acts as a molecular sponge, promoting gastric cancer by absorbing miR-16. Endocr Relat Cancer. 2019;26(3):265–77.
17. Liu M, Wang Q, Shen J, Yang BB, Ding X. Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. 2019;16(7):899–905.
18. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20(11):1666–70.
19. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2018;47(D1):D155–62.
20. Fang Z, Lei X. Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network. Big Data Mining and Analytics. 2019;2(4):261–72.
21. Qian Y, Zheng J, Jiang Y, Li S, Deng L. Prediction of circRNA-miRNA association using singular value decomposition and graph neural networks. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(6):3461–8.
22. He J, Xiao P, Chen C, Zhu Z, Zhang J, Deng L. GCNCMI: a graph convolutional neural network approach for predicting circRNA-miRNA interactions. Front Genet. 2022;13: 959701.
23. Yu CQ, Wang XF, Li LP, You ZH, Huang WZ, Li YC, Ren ZH, Guan YJ. SGCNCMI: a new model combining multi-modal information to predict circRNA-related miRNAs, diseases and genes. Biology (Basel). 2022;11(9):1350.
24. Lan W, Zhu M, Chen Q, Chen J, Ye J, Liu J, Peng W, Pan S. Prediction of circRNA-miRNA Associations based on network embedding. Complexity. 2021;2021:6659695.
25. Wang XF, Yu CQ, Li LP, You ZH, Huang WZ, Li YC, Ren ZH, Guan YJ. KGDCMI: a new approach for predicting circRNA-miRNA interactions from multi-source information extraction and deep learning. Front Genet. 2022;13: 958096.
26. Yao D, Nong L, Qin M, Wu S, Yao S. Identifying circRNA-miRNA interaction based on multi-biological interaction fusion. Front Microbiol. 2022;13: 987930.
27. Guo L-X, You Z-H, Wang L, Yu C-Q, Zhao B-W, Ren Z-H, Pan J. A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. Briefings in Bioinf. 2022;23(5):bbac391.
28. Wang X-F, Yu C-Q, You Z-H, Li L-P, Huang W-Z, Ren Z-H, Li Y-C, Wei M-M. A feature extraction method based on noise reduction for circRNA-miRNA interaction prediction combining multi-structure features in the association networks. Briefings Bioinf. 2023;24(3):bbad111.
29. Wang XF, Yu CQ, You ZH, Qiao Y, Li ZW, Huang WZ. 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.
30. Wang XF, Yu CQ, You ZH, Wang Y, Huang L, Qiao Y, Wang L, Li ZW. BEROLECMI: a novel prediction method to infer circRNA-miRNA interaction from the role definition of molecular attributes and biological networks. BMC Bioinf. 2024;25(1):264.
31. Salehi A, Davulcu H: **Graph attention auto-encoders**. *arXiv preprint* 2019.
32. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. InProceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 (pp. 855-864).
33. Meng X, Hu D, Zhang P, Chen Q, Chen M. CircFunBase: a database for functional circular RNAs. Database. 2019;2019:baz003.
34. Perozzi B, Al-Rfou R, Skiena S: **Deepwalk: Online learning of social representations**. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining: 2014*. 701–710.
35. Li G, Luo J, Wang D, Liang C, Xiao Q, Ding P, Chen H. Potential circRNA-disease association prediction using DeepWalk and network consistency projection. J Biomed Inform. 2020;112: 103624.
36. Ghulam A, Lei X, Zhang Y, Wu Z. Human drug-pathway association prediction based on network consistency projection. Comput Biol Chem. 2022;97: 107624.
37. Gu C, Liao B, Li X, Li K. Network consistency projection for human miRNA-disease associations inference. Sci Rep. 2016;6:36054.
38. Chen L, Xu J, Zhou Y. PDATC-NCPMKL: predicting drug's anatomical therapeutic chemical (ATC) codes based on network consistency projection and multiple kernel learning. Comput Biol Med. 2024;169: 107862.
39. Chen L, Huiru Hu. MBPathNCP: a metabolic pathway prediction model for chemicals and enzymes based on network consistency projection. Curr Bioinform. 2024. https://doi.org/10.2174/0115748936321359240827050752.
40. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: *International joint Conference on artificial intelligence: 1995*. Lawrence Erlbaum Associates Ltd: 1137–1145.
41. Davis J, Goadrich M: **The relationship between Precision-Recall and ROC curves**. In: *The 23rd international conference on machine learning: 2006*. 233–240.

42. Chen L, Zhang C, Xu J. PredictEFC: a fast and efficient multi-label classifier for predicting enzyme family classes. BMC Bioinf. 2024;25:50.
43. Chen L, Gu J, Zhou B. PMiSLocMF: predicting miRNA subcellular localizations by incorporating multi-source features of miRNAs. Briefings Bioinf. 2024;25(5):bbae386.
44. Chen L, Chen Y. RMTLysPTM: Recognizing multiple types of lysine PTM sites by deep analysis on sequences. Briefings Bioinf. 2024;25(1):bbad450.
45. Powers D. Evaluation: from precision recall and f-measure to roc, informedness markedness correlation. J Machi Learn Technol. 2011;2(1):37–63.
46. Ren J, Gao Q, Zhou X, Chen L, Guo W, Feng K, Huang T, Cai Y-D. Identification of key gene expression associated with quality of life after recovery from COVID-19. Med Biol Eng Compu. 2024;62(4):1031–48.
47. Ren J, Zhou X, Huang K, Chen L, Guo W, Feng K, Huang T, Cai Y-D. Identification of key genes associated with persistent immune changes and secondary immune activation responses induced by influenza vaccination after COVID-19 recovery by machine learning methods. Comput Biol Med. 2024;169: 107883.
48. Ren J, Chen L, Guo W, Feng K, Huang T, Cai Y-D. Patterns of gene expression profiles associated with colorectal cancer in colorectal mucosa by using machine learning methods. Comb Chem High Throughput Screening. 2024;27(19):2921–34.
49. Deng L, Liu Z, Qian Y, Zhang J. Predicting circRNA-drug sensitivity associations via graph attention auto-encoder. BMC Bioinf. 2022;23(1):160.
50. Zhu L, Duan G, Yan C, Wang J: **Prediction of microbe-drug associations based on Katz measure**. In: *2019 IEEE international conference on bioinformatics and biomedicine (BIBM): 2019*. IEEE: 183–187.
51. Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(3):646–56.


## Publisher's Note