BMC Bioinformatics

RESEARCH

Open Access

BEROLECMI: a novel prediction method to infer circRNA-miRNA interaction from the role definition of molecular attributes and biological networks

Xin-Fei Wang[1], Chang-Qing Yu[1*], Zhu-Hong You[2*], Yan Wang[3,4*], Lan Huang[3], Yan Qiao[5], Lei Wang[6,7] and Zheng-Wei Li[6]

*Correspondence:
xaycq@163.com;
zhuhongyou@nwpu.edu.cn;
wy6868@jlu.edu.cn

[1] School of Information Engineering, Xijing University, Xi'an, China
[2] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[3] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China
[4] School of Artificial Intelligence, Jilin University, Changchun, China
[5] College of Agriculture and Forestry, Longdong University, Qingyang, China
[6] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
[7] Guangxi Academy of Sciences, Nanning, China

Abstract

Circular RNA (CircRNA)–microRNA (miRNA) interaction (CMI) is an important model for the regulation of biological processes by non-coding RNA (ncRNA), which provides a new perspective for the study of human complex diseases. However, the existing CMI prediction models mainly rely on the nearest neighbor structure in the biological network, ignoring the molecular network topology, so it is difficult to improve the prediction performance. In this paper, we proposed a new CMI prediction method, BEROLECMI, which uses molecular sequence attributes, molecular self-similarity, and biological network topology to define the specific role feature representation for molecules to infer the new CMI. BEROLECMI effectively makes up for the lack of network topology in the CMI prediction model and achieves the highest prediction performance in three commonly used data sets. In the case study, 14 of the 15 pairs of unknown CMIs were correctly predicted.

Keywords: Competing endogenous RNA, circRNA–miRNA interaction, Association prediction, Network embedding, Biomarker discovery

Introduction

CircRNA (circular RNA) is a kind of non-coding RNA with a closed-loop structure that exists in eukaryotic cells [1]. CircRNA was first discovered in the 1980s [2] and is considered to be a noise or by-product in the process of transcription. However, with the development of high-throughput sequencing technology, researchers have re-focused on circRNA and found that they have a variety of biological functions. By screening and analyzing the differentially expressed circRNA, researchers revealed the expression patterns of circRNA in different tissues, developmental stages, and disease states [3]. CircRNA affects cell function by regulating miRNA activity or interacting with RNA-binding proteins. Therefore, it plays a role as a potential biomarker [4].

MicroRNA (miRNA) is a kind of non-coding RNA, which can regulate gene expression by binding to the 3' untranslated region (3'UTR) of the target gene [5]. In the CMI, circRNA binds to miRNA through the "miRNA response element (MRE)" with complementary sequences to form a circRNA-miRNA complex [6, 7]. This binding occurs mainly through two main mechanisms: the sponge effect and the competitive binding mechanism. In the sponge effect, circRNA absorbs multiple miRNA molecules as the "sponge" of miRNA, thus reducing the binding between miRNA and target mRNA, and affecting the regulation of target genes by miRNA. In the competitive binding mechanism, circRNA and miRNA compete directly with the 3'UTR binding sites of the same target gene to hinder the effect of miRNA on the target gene. CMI plays an important role in the regulation of gene expression, cell proliferation, and tumorigenesis, which provides new hope for the diagnosis, treatment, and prognosis of diseases [8, 9].

Although circRNA has been proven to play an important role in biology, its specific function is not completely clear, so there is an urgent need for a comprehensive study of circRNA. At present, a large number of circRNA have been identified, which brings new opportunities and challenges to the related data mining work. Limited by manpower, materials, and resources, it is impractical to match a large number of CMI data aimlessly. Therefore, the use of computing technology to provide a high probability pre-selection range for wet experiments has become the main means of CMI discovery. CMI inference methods based on computing technology are mainly divided into two kinds: one is based on sequence matching, such as miRanda [10], and TargetScan [11], and the other is the prediction model based on known association. Based on the principle of complementary binding sites, the method based on sequence matching can achieve large-scale CMI prediction, but it will produce too many false positive samples, and only relying on a single feature will cause the known molecular association and biological structural features to be ignored; the use of advanced prediction model can effectively make up for the shortcomings of sequence matching methods, but it is often affected by data scale, network expansion, attribute collection and so on.

Currently, models that use computational methods to predict associations between biomedical entities are constantly being proposed, such as drug–drug interaction prediction [12], drug–target interaction prediction [13, 14], LncRNA-disease prediction [15–17], lncRNA–protein interaction [18, 19], circRNA-disease prediction [20–22], disease-associated Piwi-interacting RNAs prediction [23], cell–cell communication inference [24–26] and phage-host interaction prediction [27]. However, there are still few studies on prediction models for circRNA–miRNA interactions. Most of the existing CMI prediction models use the nearest-neighbor relationship in the biological network for modeling. For example, Wang et al. [28] proposed the KGDCMI method, which uses the HOPE method to embed the node association structure in the CMI network to predict the unknown CMI; Guo et al. [29] used the SDNE method to embed the similarity between nodes and the similarity of neighbor sets in the CMI network to predict the new CMI. Yu et al. [30] and He et al. [31] use graph convolution neural network to aggregate the feature prediction of nearest neighbor nodes in high-dimensional space CMI; Qian et al. [32] proposed CMASG, by extracting linear and nonlinear features from the CMI network to predict unknown CMI. Wang et al. used the signed graph convolutional neural network to aggregate friend and foe relationships in circRNA-miRNA-cancer

networks to predict unknown CMI [33]. These methods are used to model and extract the nearest neighbor relationship in the CMI network to realize the effective prediction of CMI. However, as a kind of community relationship in the network, the application of the nearest neighbor relationship in the biological network has some limitations: 1. CMI biological network is sparse, most nodes have little or no nearest neighbor relationship; 2. The nearest neighbor relationship will be bound to the node representation, so it is difficult to achieve network expansion; 3. Limited by the capture mode of neighboring nodes, the similarity (structural similarity) between nodes with similar network topology features but not adjacent nodes is ignored. This makes it difficult to achieve a breakthrough in the accuracy of CMI prediction. As a kind of network structure, the local topological structure information in the network has been applied to the CMI prediction model. Wang et al. added local topological structure similarity to the model as a supplement to obtain reliable prediction performance [34]; Wang et al. used the wavelet diffusion mode to extract the topological structure feature of nodes in the CMI network and obtain the highest prediction performance [35]. This means that the topological structure feature is effective in model feature extraction, but these methods only use the local topological structure in the network as a supplement to the features and do not clearly define the topological structure, which limits the value of local topological structure features.

In this paper, we propose a CMI prediction method, BEROLECMI, which defines role attributes for each molecule through molecular attribute features, molecular self-similarity networks, and molecular network features for advanced prediction tasks. Specifically, BEROLECMI first uses the pre-trained Bidirectional Encoder Representations from the Transformers model for DNA language in genome (DNABERT) [36] to extract attribute features from RNA sequence, then constructs RNA self-similarity networks through Gaussian kernel function and sigmoid kernel function respectively, and the high-level representation is learned by sparse autoencoder (SAE) [37]. Next, the proposed model uses the structured embedding method to extract the role similarity of each molecule in the CMI network. Finally, these features are organically integrated as molecular exclusive role features and sent to the classifier for training and prediction tasks. In the performance verification, the prediction performance of the BEROLECMI in common data sets exceeds that of all known models. In the case study, the proposed method accurately predicts 14 of 15 pairs of CMIs. The flow chart of BEROLECMI is shown in Fig. 1.

Materials and methods

Molecular attribute feature construction

In this part, the BEROLECMI method combines the sequence structure of molecules and self-similarity network to construct molecular attribute descriptors. Specifically, we first cut the sequence of RNA molecules to form effective sequence fragments, and then use the pre-training model, DNABERT, to learn the potential features in RNA sequences. Next, we use the Gaussian kernel function and sigmoid kernel function to construct molecular self-similarity networks and use SAE to learn the advanced representation of molecular self-similarity. Finally, the RNA sequence feature and molecule

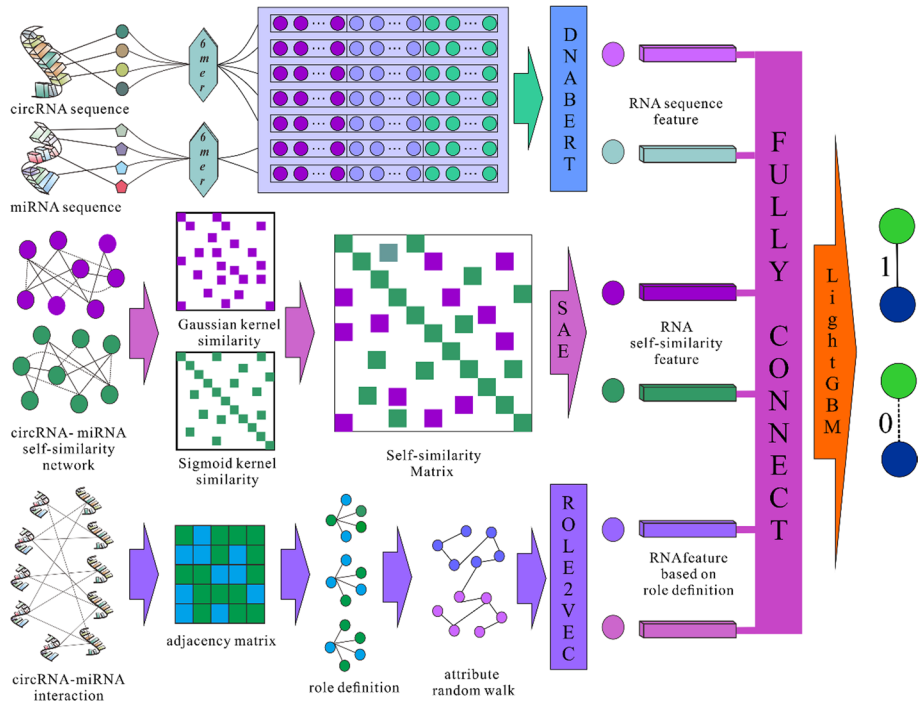

Fig. 1 The flow chart of BEROLECMI

self-similarity feature are fully connected, and the exclusive attribute feature is assigned to each molecule.

RNA sequence feature extraction

The Bidirectional Encoder Representations from Transformers (BERT) model has achieved great success in the field of natural language processing (NLP), which provides a new pattern for the use of large-scale preprocessing models. However, due to the unique biological characteristics, and physical and chemical structure of DNA sequences, the traditional NLP algorithm is difficult to effectively apply to the modeling of DNA data. Therefore, through the improvement and adjustment of BERT, Ji et al.[36] developed a DNABERT model suitable for DNA sequences. In this part, we introduce the DNABERT model to learn the sequence features of RNA. Specifically, BEROLECMI first cuts the RNA sequence according to the sequence aggregate of 6-mer, which is determined by the training corpus of the DNABERT model; then, the obtained RNA sequence aggregate is used as the corpus, and the pre-trained DNABERT model is used to capture the local and global representation of the DNA sequence. The feature extraction process of the RNA sequence is shown in Fig. 2. In addition, detailed information about the DNABERT pre-training model is referred to in the research of Ji et al. [36].

Molecular self-similarity feature construction

RNA molecular self-similarity features are based on the functional similarity hypothesis, that is, molecules with similar binding targets may have the same functions. BEROLECMI introduces the Gaussian kernel function and sigmoid kernel function to construct RNA self-similarity in the CMI network. For the dataset CMI-9905 stored in

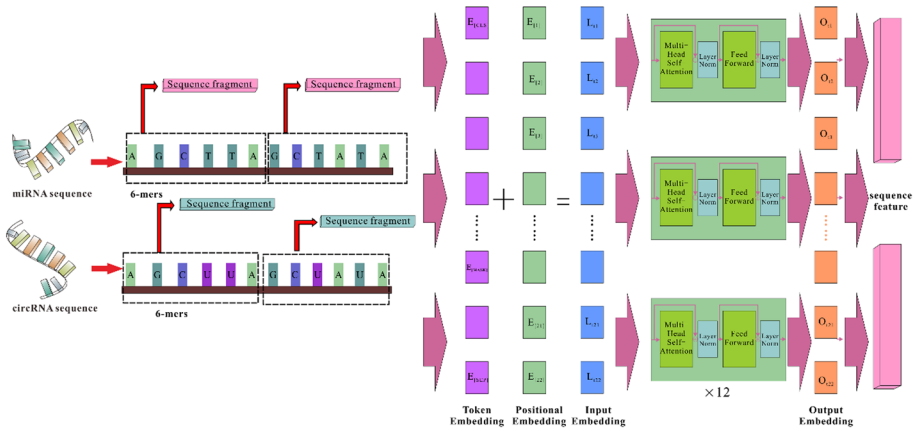

Fig. 2 The feature extraction process of the RNA sequence

the adjacency matrix M, when circRNA a interacted with miRNA b, M_{ab} is 1, otherwise 0. The circRNA Gaussian kernel representation in the matrix can be calculated as:

$$K_{circRNA}(U_a, U_b) = \exp(-\lambda||LP(U_a) - LP(U_b)||^2) \tag{1}$$

where U_a and U_b represent circRNA a and circRNA b respectively, $K_{circRNA}(U_a, U_b)$ denotes the GIP kernel similarity between circRNA a and b, λ is a parameter controlling the bandwidth of the GIP kernel function, calculated as:

$$\lambda_U = \lambda_U{\prime}/\left(\frac{1}{n}\sum_{L=1}^{n}||LP(U_i)||^2\right) \tag{2}$$

Similarly, the miRNA Gaussian kernel representation can be calculated as:

$$K_{miRNA}(V_a, V_b) = \exp(-\lambda||LP(V_a) - LP(V_b)||^2) \tag{3}$$

$$\lambda_V = \lambda_V{\prime}/\left(\frac{1}{n}\sum_{L=1}^{n}||LP(V_i)||^2\right) \tag{4}$$

The circRNA sigmoid kernel representation is defined as:

$$S_{circRNA}(U_a, U_b) = \tanh\{\eta[G(U_a)] \times \mu[G(U_b)]\} \tag{5}$$

where $\eta = 1/N$, N is the dimension of the input data.

Similarly, the miRNA sigmoid kernel representation is defined as:

$$S_{miRNA}(V_a, V_b) = \tanh\{\eta[G(V_a)] \times \mu[G(V_b)]\} \tag{6}$$

Molecular self-similarity integration

BEROLECMI organically integrates Gaussian kernel similarity and sigmoid kernel similarity of molecules to obtain a highly representative self-similarity matrix. For circRNA, the similarity matrix can be calculated as:

$$F_U(U_a, U_b) = \begin{cases} S_{circRNA}(U_a, U_b) & K_{circRNA}(U_a, U_b) < 0.1, S_{circRNA}(U_a, U_b) > 0.1 \\ K_{circRNA}(U_a, U_b) & otherwise \end{cases} \tag{7}$$

For miRNA, the similarity matrix can be calculated as:

$$F_V(V_a, V_b) = \begin{cases} S_{miRNA}(V_a, V_b) & K_{miRNA}(V_a, V_b) < 0.1, S_{miRNA}(V_a, V_b) > 0.1 \\ K_{miRNA}(V_a, V_b) & otherwise \end{cases} \tag{8}$$

The Gaussian kernel similarity, sigmoid kernel similarity, and self-similarity matrix of circRNA and miRNA are shown in Fig. 3.

Molecular self-similarity feature enhancement

BEROLECMI performs dimensionality reduction and feature enhancement on molecular self-similarity networks by introducing SAE. The SAE has the same encoding and decoding structure as the ordinary autoencoder, and the input layer maps the input data to the hidden layer L for encoding:

$$L_1 = \sigma(WX(L_0) + b) \tag{9}$$

where $X(L_0)$ is the input data, W is the parameter of the hidden layer.

The difference is that the SAE adds sparsity constraints during the training process. By adding sparsity constraints to the neurons in the hidden layer, the hidden layer keeps a low activation value to obtain highly representative key features. The sparse penalty term T can be calculated as:

$$T = \sum_{i=1}^{N} KL(P||F) \tag{10}$$

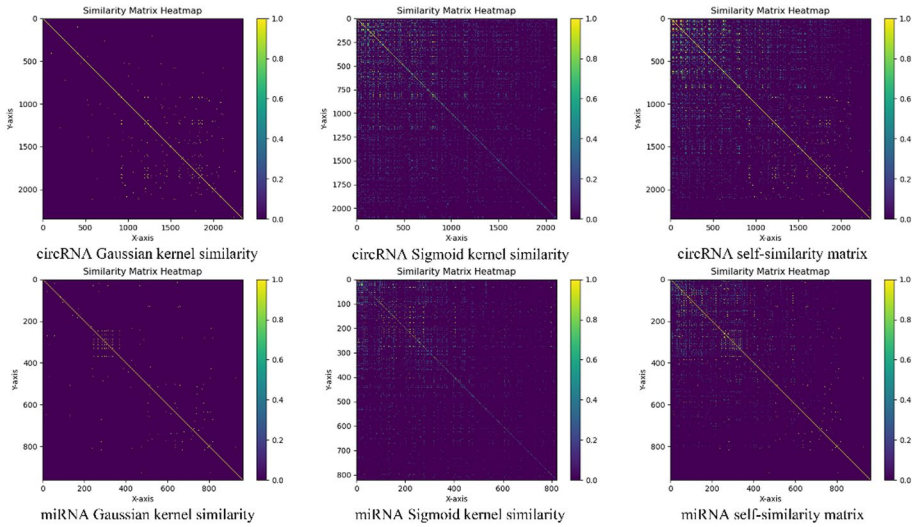

Fig. 3 The self-similarity of molecular

where T represents the sum of the degree to which the penalty item P deviates from F, and N represents the number of neurons in the hidden layer. In this experiment, the KL divergence (Kullback–Leibler) is used to calculate as follows:

$$KL(P||F) = P\log\frac{P}{F} + (1-P)\log\frac{1-P}{1-F} \tag{11}$$

where P is the sparse parameter of KL, the closer F is to P, the smaller the value of KL.

By adding sparsity constraints in the encoding layer, SAE can learn a low-dimensional robust representation of the original features, thereby improving model performance.

Network embedding based on role definition

To effectively extract molecular structured embeddings in biological networks, BEROLECMI introduces the Role2vec algorithm [38] combined with attribute random walks to capture molecular structure similarities in networks. For an undirected graph G < N, E > composed of N nodes and E edges constructed based on the CMI biological network, the Role2vec algorithm defines different roles for nodes according to the network topology, such as motifs, graphlets, etc., as shown in Fig. 4. The structure type is flexibly selected via parameters.

The detailed description of the Role2vec algorithm is shown in Algorithm 1. Role2vec uses graph G, sub-graph structure M, embedded dimension d, node walk r, walk length l, and context window size z as the algorithm input. In steps 1–3, the subgraph structure in graph G is first extracted and transformed into a specific representation, and then the subgraph structure is mapped to x by the function f. In step 4, the transfer probability p is calculated. In steps 6–10, the nodes were reordered in the reconstructed graph $G^{'}$, and the attribute walk to extract the feature representation a of node n. Then, add a to the set A. Finally, the role embedding representation of each node is obtained by using the stochastic gradient descent algorithm.

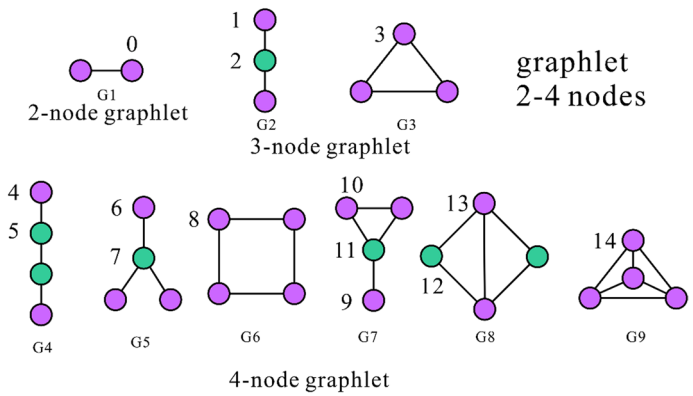

Fig. 4 Different roles for nodes according to the network topology

Algorithm 1 Role2vec

Input: G<V, E>, Subgraph structure(graphlet) M, embedding dimension d, walk length l, window size z, node walks r
Output: embedding F of each node
Extract subgraph structure(graphlet) to M
Transform feature in M
Map vertices to roles function $f : m \to x$
Calculate the transition probability p
$G' < V, E, p >$
Walk node for i=1, 2, …, r do
  Randomly arrange the node set N
  For each node $n \in N$
    $a = attribute\ random\ walk(G', M, n, f, l)$
    Add a to set A
F = stochastic gradient descent(z, d, A)
Return embedding F

Different from the traditional random walk, the walk method with the structure type as the node attribute only focuses on the local topology of the node, so it can effectively define roles for different nodes. This means that the nodes only rely on structural features for embedding, which has high scalability and can be extended to nodes in distant or even different graphs.

Results

Evaluation criteria

In this study, we introduce five-fold cross-validation (five-fold CV) to evaluate the performance of the proposed method. five-fold CV divides the CMI data into five subsets at random, each time four subsets are used as the training set, one subset is used as the test set, and five experiments are performed until the predicted score of each subset. We comprehensively evaluate the predictive performance of the proposed model by combining multiple evaluation criteria including Accuracy. (Acc.), Precision. (Prec.), Recall and F1-score. The evaluation criteria can be represented as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\Pr ec. = \frac{TP}{TP + FP} \tag{13}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - score = 2 \times \frac{Precision \times \mathrm{Re}\, call}{Precision + \mathrm{Re}\, call} \tag{15}$$

Among them, TP and TN respectively represent the number of positive samples and negative samples predicted correctly by the model; FP and FN respectively represent the number of negative samples and positive samples predicted incorrectly by the model. In addition, we introduce the receiver operating characteristic curve (ROC), and precision-recall curve (PR).

In this work, we evaluate model performance through the five-fold CV based on three commonly used datasets in the field of CMI prediction. The CMI-9905 dataset was compiled by Wang et al. [28], including 9905 interactions between 2346 circRNA and 962 miRNA. The data set contains CMI with high confidence, which is used as the benchmark data set in this study. The CMI-9589 dataset comes from the circBank database [39], and we select 9589 interactions of CMI between 2115 circRNA and 821 miRNA with high confidence as training data. The CMI-753 dataset is collected from the circR-2Cancer database [40]. Through strict screening and processing, we have obtained 753 interactions of CMI between 515 circRNA and 469 miRNA in the latest version of the data, all of which are supported by experiments. In this experiment, we use this data for a case study.

In addition, we construct negative samples to balance the data set based on the uniqueness principle of sequence complementarity. Since miRNA has the response components (MRE), endogenous RNAs with a common MRE regulate each other's expression by competitively binding to miRNA. This theory is called the competitive endogenous RNA hypothesis [41]. If a circRNA is determined to contain an MRE, it may be a potential target of a miRNA, and vice versa. In this study, negative samples are defined as interacting pairs that do not share common MREs, and we adopt specific negative sample construction methods for different data sets. Specifically, for the CMI-9905 dataset and CMI-9589 dataset, we construct all possible interactions between circRNA and miRNA, then delete the CMI with confidence score (score > 0) in the circBank database, and finally randomly select the same number of CMI as negative samples to participate in model training; For the CMI-753 data set based on real cases, since known interaction pairs are reported by experiments or papers, we select interaction pairs that have not been reported in existing studies as negative samples. Using this method can effectively avoid the potential CMI as a negative sample and ensure the reliability of the model performance.

Table 1 The prediction result of BEROLECMI based on the benchmark dataset

Test set	Acc	Prec	Recall	F1-score	AUROC	AUPR
1	0.8465	0.8491	0.8465	0.8463	0.9171	0.9104
2	0.8301	0.8333	0.8301	0.8297	0.9040	0.9024
3	0.8367	0.8401	0.8367	0.8363	0.9090	0.9115
4	0.8304	0.8320	0.8304	0.8302	0.9050	0.8983
5	0.8541	0.8590	0.8541	0.8536	0.9170	0.9204
Mean	0.8395	0.8427	0.8396	0.8392	0.9104	0.9086
Std	0.0093	0.0101	0.0094	0.0093	0.0056	0.0076

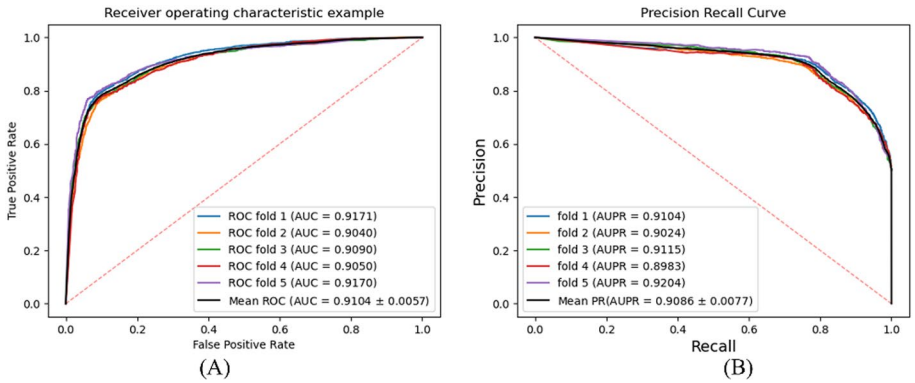

Fig. 5 The ROC curve (**A**) and PR curve (**B**) of the BEROLECMI

Table 2 The prediction result of BEROLECMI based on commonly used datasets

	Acc	Prec	Recall	F1-score	AUROC	AUPR
CMI-9589						
1	0.8835	0.8856	0.8835	0.8833	0.9528	0.9499
2	0.8723	0.8759	0.8723	0.8720	0.9492	0.9409
3	0.8743	0.8755	0.8743	0.8742	0.9454	0.9367
4	0.8832	0.8848	0.8832	0.8831	0.9507	0.9465
5	0.8756	0.8777	0.8756	0.8754	0.9475	0.9416
Mean	0.8777	0.8799	0.8778	0.8776	0.9491	0.9431
Std	0.0046	0.0043	0.0047	0.0047	0.0025	0.0046
CMI-753						
1	0.6987	0.6987	0.6987	0.6987	0.7781	0.7758
2	0.7409	0.7461	0.7409	0.7393	0.8133	0.7932
3	0.7309	0.7364	0.7309	0.7292	0.7856	0.7591
4	0.7375	0.7432	0.7375	0.7361	0.8002	0.7503
5	0.6811	0.6876	0.6811	0.6785	0.7766	0.7289
Mean	0.7178	0.7224	0.7178	0.7163	0.7907	0.7614
Std	0.0236	0.0243	0.0237	0.0237	0.0140	0.0219

Performance evaluation

In this section, we use the CMI-9905 as the benchmark dataset for performance evaluation. The data of the model in the five-fold CV is objectively recorded in Table 1.

The data in Table 1 shows that in the five-fold CV based on the benchmark data set, the average values of the six evaluation criteria of BEROLECMI are 0.8395, 0.8427, 0.8396, 0.8392, 0.9104, and 0.9086, respectively, which means that the proposed model can efficiently complete the prediction task of CMI. The ROC and PR curves of the BEROLECMI are shown in Fig. 5.

Performance on different datasets

To reflect the generalization ability of BEROLECMI in CMI prediction, we perform prediction tasks based on all commonly used datasets in the field of CMI prediction (CMI-9589, CMI-753). According to our statistics, more than 80% of CMI prediction models

use the dataset adopted in this work as benchmark data. The experimental results based on commonly used datasets are shown in Table 2.

The data in Table 2 shows that in all common data sets in the field of CMI prediction, the AUROC of BEROLECMI based on the CMI-9589 data set exceeds 90%, and the AUROC based on the CMI-753 data set exceeds 75%, which means that the proposed model can effectively complete the CMI prediction task in commonly used data sets, and is expected to become a reliable candidate tool for CMI prediction.

The validity of the model feature extraction

In this section, we verify the effectiveness of feature extraction for each part of the BEROLECMI through independent experiments. Specifically, we divide BEROLECMI into three modules: sequence feature extraction (BE-A), self-similarity feature extraction (BE-B), and structured embedding (BE–C), and then use the three modules separately for feature extraction and perform prediction tasks to evaluate the proposed model for each module feature extraction effectiveness. The experimental results are objectively recorded in Table 3. To facilitate comparison, we use histograms to visualize the data in Table 3, as shown in Fig. 6.

The data in Table 3 shows that all the feature modules of the BEROLECMI can effectively complete the CMI prediction, which shows the effectiveness of the feature extraction strategy of the proposed method; among all three modules, the sequence feature

Table 3 Predicted result of different modules of BEROLECMI

	Acc	Prec	Recall	F1-score	AUROC	AUPR
BE-A						
1	0.6663	0.7022	0.6663	0.6509	0.7104	0.7266
2	0.6769	0.7088	0.6769	0.6641	0.7376	0.7521
3	0.6681	0.711	0.6681	0.6503	0.7539	0.7606
4	0.6772	0.6983	0.6772	0.6683	0.7375	0.7534
5	0.658	0.6874	0.6580	0.6440	0.7209	0.7364
Mean	0.6693	0.7015	0.6693	0.6555	0.7320	0.7458
Std	0.0071	0.0084	0.0072	0.0091	0.0150	0.0124
BE-B						
1	0.8243	0.8262	0.8243	0.8241	0.8873	0.8803
2	0.8261	0.829	0.8261	0.8257	0.8934	0.8921
3	0.8152	0.8175	0.8152	0.8149	0.8769	0.8715
4	0.8147	0.8168	0.8147	0.8144	0.8830	0.8812
5	0.8311	0.8326	0.8311	0.8310	0.8884	0.8847
Mean	0.8222	0.8244	0.8223	0.8220	0.8858	0.8819
Std	0.0063	0.0062	0.0064	0.0064	0.0055	0.0066
BE-C						
1	0.8009	0.8021	0.8009	0.8007	0.8850	0.8869
2	0.8089	0.8106	0.8089	0.8087	0.8867	0.8739
3	0.8059	0.8081	0.8059	0.8056	0.8885	0.8835
4	0.8102	0.8120	0.8102	0.8099	0.8857	0.8782
5	0.8001	0.8036	0.8001	0.7995	0.8808	0.8660
Mean	0.8052	0.8072	0.8052	0.8048	0.8853	0.8777
Std	0.0040	0.0038	0.0041	0.0041	0.0025	0.0073

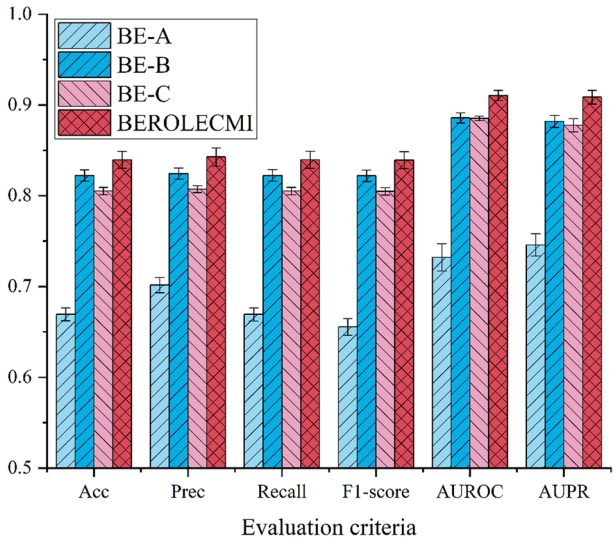

Fig. 6 Performance comparison of different modules

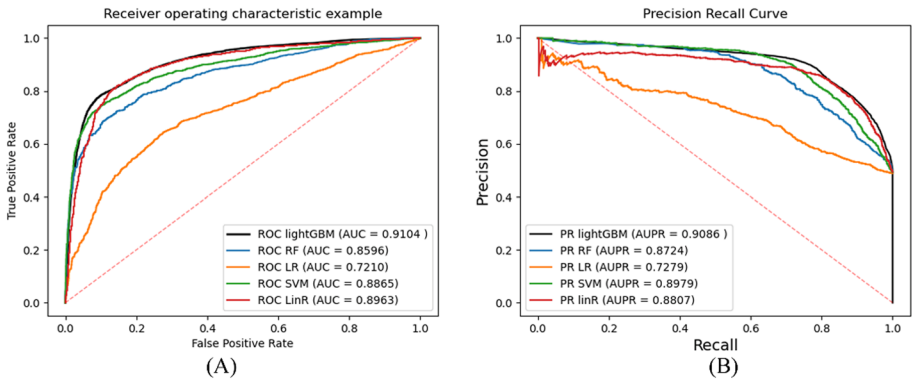

Fig. 7 Comparison of prediction results of different classifiers (**A** is the comparison of AUC results, **B** is the comparison of AUPR results)

extraction module has the lowest prediction results, which shows that the sequence feature is useful complements to model features; self-similarity features and structured embeddings achieve high predictive results, which means that combining functional similarity assumptions and role-defined structural embeddings can effectively improve the predictive performance of the model. Through the organic integration of the three feature extraction modules, we achieved the highest model prediction performance, verifying the effectiveness of the model construction.

Optimal classification strategy

In this study, we conduct prediction tasks based on different classifiers to determine the best classification strategy for the proposed method. In prediction tasks based on CMI-9905 datasets, we use the lightGBM [42], Random forest (RF) [43], Logistic Regression (LR) [44], Support Vector Machine (SVM) [45], Linear Regression (LinR) [46] for CMI

prediction tasks, and the best classification strategy was selected by comparing the performance of the proposed methods. The prediction results are shown in Fig. 7.

The data in Fig. 7 shows that the model using the lightGBM classifier achieves the highest predictive performance. LightGBM (Light Gradient Boosting Machine) [34] is an ensemble learning classifier based on Gradient Boosting Decision Tree (GBDT), which performs well in prediction tasks.LightGBM adopts the ensemble learning method to build a powerful prediction model by iteratively training multiple weak prediction models. It repeatedly optimizes the loss function, and each iteration builds a new decision tree on the residual of the previous model and then combines multiple decision trees to generate the final prediction result. LightGBM has the advantages of high efficiency, low memory consumption, accuracy, and support for large-scale data sets. Through the comparison of various classifiers, we finally choose LightGBM as the final classification strategy of the model.

Compared with the existing models

To evaluate the advantages of BEROLECMI in the CMI prediction, we compared the proposed model with other models in the CMI prediction field based on three commonly used data sets.

Lan et al. proposed the NECMA model, which combines circRNA-miRNA association, circRNA Gaussian kernel similarity and miRNA Gaussian kernel similarity to construct a heterogeneous network, then uses the NetMF algorithm based on matrix decomposition to extract hidden features in the heterogeneous network, and finally uses weighted neighborhood regularized logistic matrix decomposition and inner product obtain the circRNA-miRNA association probability [47]; Qian et al. proposed the CMIVGSD model, using the singular value decomposition algorithm and variational autoencoder to extract linear and nonlinear features from the circRNA-miRNA interaction network to predict unknown circRNA-miRNA interactions [32]; Wang et al. proposed the KGDCMI model, which combines RNA sequence feature and CMI network behavior feature to predict unknown CMI; Yu et al. proposed the first comprehensive prediction model of circRNA, SGCNCMI, which uses the graph neural network based on the contributing mechanism aggregates multi-modal information of molecules in biological networks and can achieve multiple predictions of circRNA-miRNA interactions, circRNA-gene interactions, and circRNA-cancer associations [30]; Guo et al. proposed the WSCD model, combined with the word2vec algorithm in natural language processing to process RNA sequences, used the SDNE algorithm to extract behavior features in the CMI network, and finally used a deep neural network to predict CMI [29]; He et al. proposed the GCNCMI model, using graph convolutional neural networks to aggregate node information to predict potential circRNA-miRNA interaction; Wang et al. proposed the JSNDCMI model, which for the first time combined denoising methods and local topological structure information in the CMI network for molecular feature extraction to predict unknown CMI [34]; Yao et al. proposed the IIMCCCMA model, which combined matrix factorization and improved inductive matrix completion algorithms predict unknown CMI [48]. Wang et al. proposed BioDGW-CMI, which combines BERT and wavelet diffusion to extract sequence and association network structure information of RNA molecules to predict potential CMI [35]. These models have achieved exciting results in CMI

Table 4 Model performance comparison with different CMI prediction models

	AUROC	AUPR
CMI-9905		
KGDCMI	0.8930	0.8767
WSCD	0.8923	0.8935
SGCNCMI	0.8942	0.8887
JSNDCMI	0.9003	0.8999
BioDGW-CMI	0.9026	0.8962
BEROLECMI	0.9104	0.9086
CMI-9589		
CMIVGSD	0.8804	0.8629
SGCNCMI	0.9015	0.9011
KGDCMI	0.9041	0.8937
GCNCMI	0.9320	0.9396
JSNDCMI	0.9415	0.9403
BioDGW-CMI	0.9476	0.9416
BEROLECMI	0.9491	0.9431
CMI-753		
NECMA	0.4989	0.0003
GCNCMI	0.5679	0.0004
CMIVGSD	0.5755	0.0007
IIMCCMA	0.6702	0.0009
BioDGW-CMI	0.7821	0.7688
BEROLECMI	0.7907	0.7614

Table 5 Paired t-test results of the BEROLECMI and other models under five-fold cross-validation

t-test	WSCD	KGDCMI	SGCNCMI	JSNDCMI
p value	0.0011	0.0005	0.0071	0.0233

prediction, and we compare the BEROLECMI with these models to reflect the superior performance of the proposed model.

The comparison data are recorded in Table 4. It is worth noting that all the comparison data in this study use the same data and verification methods as the comparison models, and the number of comparison models exceeds 70% of all models in the field of CMI prediction.

Data based on the CMI-753 dataset comes from the work of Yao et al. [48].

The data in Table 4 shows that BEROLECMI surpasses all known models in the performance comparison of all three datasets.

In addition, we conduct a paired t-test based on the CMI-9905 data set and known advanced models in the CMI field to evaluate the statistical difference between the proposed model and the state-of-the-art (SOTA) model. The experimental results are recorded in Table 5.

The data in Table 5 shows that in the verification of the proposed method with the SOTA model, the P values were less than 0.05 confidence level, which means that there is a significant difference between the proposed model and the comparison model, and

it has better statistical validity predict performance. There is no doubt that the proposed method is currently the most competitive in the field of CMI prediction.

Case study

To verify the practicability of BEROLECMI, we conducted a case study based on the CMI-753 dataset. The data in this dataset are all manually collected from existing literature and research, and all have experimental support.

In the case study, we perform a prediction task based on 15 pairs of CMIs to simulate the prediction performance of the proposed model in unknown CMIs. Specifically, we remove the interacting pairs for case studies in the CMI-753 data and then use the known 738 pairs of CMIs for model training to predict unknown CMIs. The prediction results are recorded in Table 6.

The data in Table 6 shows that among the 15 pairs of CMI used for prediction, 14 pairs were successfully predicted, which means that BEROLECMI can effectively predict CMI in real cases, and it is expected to be a powerful tool to provide pre-selection for wet experiments.

Conclusion

The circRNA-miRNA-mediated model provides new hope for the diagnosis and treatment of many complex diseases, but the biological properties of the mediated model are not completely clear, so there is an urgent need to speed up the discovery of CMI.

Influenced by the neighbor relationships in community networks, existing CMI prediction models typically rely on the nearest neighbor relationships in biological

Table 6 The prediction results in the case study

Num	circRNA	miRNA	Prediction score	Evidence	Cancer	Detection method
1	circ-ITCH	miR-10a	0.9999	30556849	Epithelial ovarian cancer	qPCR
2	circ-ITCH	miR-224	0.9999	29386015	Bladder cancer	qRT-PCR
3	circPVT1	miR-4663	0.9999	31636510	Esophageal carcinoma	qRT-PCR
4	circMTO1	miR-6893	0.9998	31226633	Cervical cancer	RT-qPCR;western blot
5	circMTO1	miR-92	0.9998	31456594	Glioblastoma	qRT-PCR
6	circ-ABCB10	miR-340-5p	0.9996	32196586	Hepatocellular carcinoma	qRT-PCR;Western blot.etc
7	circ-ABCB10	miR-452-5p	0.9996	32196586	Hepatocellular carcinoma	qRT-PCR;Western blot.etc
8	circ-ABCB10	let-7a-5p	0.9996	32273769	Breast cancer	qRT-PCR
9	circPVT1	miR-145	0.9985	30922567	Colorectal cancer	RT-qPCR
10	circ-PRMT5	miR-377	0.9969	31479715	non-small cell lung cancer	qPCR
11	circ-ABCB10	miR-1252	0.9724	31381507	Epithelial ovarian cancer	qRT-PCR
12	circMTO1	miR-19b-3p	0.9087	31886569	rectal cancer	qRT-PCR;Western blot
13	circMTO1	miR-9	0.7650	32207384	Renal cell carcinoma	qRT-PCR
14	circ-PRMT5	miR-498	0.7338	31479715	Non-small cell lung cancer	qPCR
15	circMTO1	miR-17-5p	0.2904	31713278	Prostate cancer	RT-qPCR

networks as the main modeling method. For example, KGDCMI and WSCD employ High-Order Proximity preserved Embedding (HOPE) and Structural Deep Network Embedding (SDNE) respectively, to capture high-order neighbor information and network structural features in heterogeneous CMI networks. Others like SGCNCMI, GCNCMI, and KS-CMI utilize graph convolutional neural networks to aggregate features of central nodes and neighbor nodes in the graph. In essence, these models use the nearest neighbor relationship set of the network or the nodes in the graph for feature extraction and aggregation. Although they can effectively predict CMI, only paying attention to a single structural feature may lead to difficult performance improvement. In addition, some models try to use other types of features in the CMI network for modeling, such as BioDGW-CMI uses the method based on wavelet diffusion to obtain the node network structure; JSNDCMI uses the multi-structure feature extraction framework to extract the topology features and functional similarity features of the nodes in the network. The prediction performance of these models is better than that of other models by adding a variety of molecular network structure features to feature extraction, which proves the advantage of multi-structure feature extraction in CMI prediction. However, although the existing methods have made progress, the existing methods still lack a targeted structure definition in the network. In this work, we propose a CMI prediction method BEROLECMI to simulate the molecular topology in the network. In this method, nine kinds of graphlet between each node and its adjacent four nodes in the CMI network are defined to extract topological features, then the attributes, self-similarity, and topological features of molecules are modeled to define unique role features for each molecule, thus inferring unknown CMIs.

BEROLECMI achieved the highest prediction performance in all commonly used datasets in the field of CMI prediction. Among them, in CMI-9905, the prediction results were 0.78% higher than the second-highest model; in the sparser CMI-753, the prediction results were higher than the second-highest model. 0.86%, which is much higher than other models by more than 10%. This means that topological structure features are an effective means of extracting network features for molecular association prediction, especially in the context of sparse relationships. In the case study, the proposed model accurately predicted 14 out of 15 pairs of CMIs. Excellent experimental results show that this method can effectively improve the performance of CMI prediction.

Although our method achieved promising results, there are still certain limitations that need to be addressed. In terms of feature extraction, although topological structures show higher advantages in sparse networks compared with nearest-neighbor structures, modeling of a single structure type shows limited capabilities in prediction tasks involving multi-type data. Therefore, the targeted use of multi-structure feature extraction will be the key to further improving prediction performance. In data construction, we use an equal amount of randomly generated negative samples for model training, which may cause potential CMI to be used as false negative samples, which can lead to a decline in model performance. We hope to continue to optimize the data used for training in subsequent research to further improve prediction performance. However, it is undeniable that BEROLECMI is currently the most competitive CMI prediction method and

has proven the effectiveness of the feature construction method based on role definition, which is expected to provide a reference for subsequent research.

Author contributions
X-FW, C-QY, Z-HY, Y-W: conceptualization, methodology, software, and data curation; L-H, Y-Q, L-W, Z-WL: conceptualization, validation, resources; X-FW wrote the main manuscript. All authors contributed to the manuscript revision and approved the submitted version.

Funding
This work was supported by the National Natural Science Foundation of China (No.62273284), and in part by the NSFC Program, under Grant 62072378, and 62002297.

Data availability
The data and source code can be found at https://github.com/1axin/BEROLECMI.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 August 2023 Accepted: 1 August 2024
Published online: 10 August 2024

References

1. Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495:333–8.
2. Hsu M-T, Coca-Prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature. 1979;280:339–40.
3. Li Z, Huang C, Bao C, et al. Exon-intron circular RNAs regulate transcription in the nucleus. Nat Struct Mol Biol. 2015;22:256–64.
4. Kulcheski FR, Christoff AP, Margis R. Circular RNAs are miRNA sponges and can be used as a new class of biomarker. J Biotechnol. 2016;238:42–51.
5. Grishok A, Pasquinelli AE, Conte D, et al. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. Cell. 2001;106:23–34.
6. Seitz H. Redefining microRNA targets. Curr Biol. 2009;19:870–3.
7. Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. Nature. 2013;495:384–8.
8. Zhang Z, Yang T, Xiao J. Circular RNAs: promising biomarkers for human diseases. EBioMedicine. 2018;34:267–74.
9. Chen L, Shan G. CircRNA in cancer: fundamental mechanism and clinical potential. Cancer Lett. 2021;505:49–57.
10. John B, Enright AJ, Aravin A, et al. Human microRNA targets. PLoS Biol. 2004;2: e363.
11. Friedman RC, Farh KK-H, Burge CB, et al. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19:92–105.
12. Ren Z-H, You Z-H, Yu C-Q, et al. A biomedical knowledge graph-based method for drug–drug interactions prediction through combining local and global features with deep neural networks. Brief Bioinform. 2022;23:bbac363.
13. Li Y-C, You Z-H, Yu C-Q, et al. PPAEDTI: personalized propagation auto-encoder model for predicting drug-target interactions. IEEE J Biomed Health Inform. 2022;27:573–82.
14. Ren Z-H, You Z-H, Zou Q, et al. DeepMPF: deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis. J Transl Med. 2023;21:48.
15. Peng L, Huang L, Su Q, et al. LDA-VGHB: identifying potential lncRNA–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous Newton boosting machine. Brief Bioinform. 2024;25:bbad466.
16. Sheng N, Cui H, Zhang T, et al. Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA–disease association prediction. Brief Bioinform. 2021;22:bbaa067.
17. 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.
18. Peng L, Tan J, Tian X, et al. EnANNDeep: an ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. Interdiscip Sci: Comput Life Sci. 2022;14:209–32.
19. Wei M-M, Yu C-Q, Li L-P et al. LPIH2V: LncRNA-protein interactions prediction using HIN2Vec based on heterogeneous networks model. Front Genet. 2023;14.
20. Zhang H-Y, Wang L, You Z-H, et al. iGRLCDA: identifying circRNA–disease association based on graph representation learning. Brief Bioinform. 2022;23:bbac083.

21. Wang L, Wong L, Li Z, et al. A machine learning framework based on multi-source feature fusion for circRNA-disease association prediction. Brief Bioinform. 2022;23:bbac388.
22. Wang L, Wong L, You ZH, et al. NSECDA: natural semantic enhancement for CircRNA-disease association prediction. IEEE J Biomed Health Inform. 2022;26:5075–84.
23. Zheng K, Zhang X-L, Wang L, et al. Line graph attention networks for predicting disease-associated Piwi-interacting RNAs. Brief Bioinform. 2022;23:393.
24. Peng L, Xiong W, Han C, et al. Cell dialog: a computational framework for ligand-receptor-mediated cell–cell communication analysis. IEEE J Biomed Health Inform. 2024;28:580–91.
25. Peng L, Wang F, Wang Z, et al. Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. Brief Bioinform. 2022;23:bbac234.
26. Peng L, Tan J, Xiong W, et al. Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data. Comput Biol Med. 2023;163:107137.
27. Pan J, You W, Lu X, et al. GSPHI: a novel deep learning model for predicting phage-host interactions via multiple biological information. Comput Struct Biotechnol J. 2023;21:3404–13.
28. Wang X-F, Yu C-Q, Li L-P, et al. KGDCMI: a new approach for predicting circRNA–miRNA interactions from multi-source information extraction and deep learning. Front Genet. 2022;13:958096.
29. Guo L-X, You Z-H, Wang L, et al. A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. Brief Bioinform. 2022;23:bbac391.
30. Yu C-Q, Wang X-F, Li L-P, et al. SGCNCMI: a new model combining multi-modal information to predict circRNA-related miRNAs, diseases and genes. Biology. 2022;11:1350.
31. 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.
32. Qian Y, Zheng J, Jiang Y, et al. Prediction of circRNA-miRNA association using singular value decomposition and graph neural networks. IEEE/ACM Trans Comput Biol Bioinform. 2022;20:3461.
33. Wang X-F, Yu C-Q, You Z-H, et al. KS-CMI: A circRNA-miRNA interaction prediction method based on the signed graph neural network and denoising autoencoder. iScience. 2023;26:107478.
34. Wang X-F, Yu C-Q, You Z-H, 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.
35. 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.
36. Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
37. Ng A. Sparse autoencoder, CS294A Lecture notes 2011;72:1–19.
38. Ahmed NK, Rossi R, Lee JB et al. Learning role-based graph embeddings. 2018. arXiv preprint arXiv:1802.02896
39. Liu M, Wang Q, Shen J, et al. Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. 2019;16:899–905.
40. Lan W, Zhu M, Chen Q, et al. CircR2Cancer: a manually curated database of associations between circRNAs and cancers. Database. 2020;2020:baaa085.
41. Salmena L, Poliseno L, Tay Y, et al. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146:353–8.
42. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inform Process Syst. 2017;30:1.
43. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
44. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken: Wiley; 2013.
45. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565–7.
46. Su X, Yan X, Tsai CL. Linear regression. Wiley Interdiscip Rev: Comput Stat. 2012;4:275–94.
47. Lan W, Zhu M, Chen Q, et al. Prediction of circRNA-miRNA associations based on network embedding. Complexity. 2021;2021:6659695.
48. Yao D, Nong L, Qin M, et al. Identifying circRNA-miRNA interaction based on multi-biological interaction fusion. Front Microbiol. 2022;13:987930.

Publisher's Note