Mathematical Biosciences
and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(12): 20553–20575.
DOI: 10.3934/mbe.2023909
Received: 29 August 2023
Revised: 17 October 2023
Accepted: 01 November 2023
Published: 14 November 2023

Research article

PCDA-HNMP: Predicting circRNA-disease association using heterogeneous network and meta-path

Lei Chen* and Xiaoyu Zhao

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

* **Correspondence:** Email: lchen@shmtu.edu.cn; Tel: +862138282825; Fax: +862138282800.

Abstract: Increasing amounts of experimental studies have shown that circular RNAs (circRNAs) play important regulatory roles in human diseases through interactions with related microRNAs (miRNAs). CircRNAs have become new potential disease biomarkers and therapeutic targets. Predicting circRNA-disease association (CDA) is of great significance for exploring the pathogenesis of complex diseases, which can improve the diagnosis level of diseases and promote the targeted therapy of diseases. However, determination of CDAs through traditional clinical trials is usually time-consuming and expensive. Computational methods are now alternative ways to predict CDAs. In this study, a new computational method, named PCDA-HNMP, was designed. For obtaining informative features of circRNAs and diseases, a heterogeneous network was first constructed, which defined circRNAs, mRNAs, miRNAs and diseases as nodes and associations between them as edges. Then, a deep analysis was conducted on the heterogeneous network by extracting meta-paths connecting to circRNAs (diseases), thereby mining hidden associations between various circRNAs (diseases). These associations constituted the meta-path-induced networks for circRNAs and diseases. The features of circRNAs and diseases were derived from the aforementioned networks via mashup. On the other hand, miRNA-disease associations (mDAs) were employed to improve the model's performance. miRNA features were yielded from the meta-path-induced networks on miRNAs and circRNAs, which were constructed from the meta-paths connecting miRNAs and circRNAs in the heterogeneous network. A concatenation operation was adopted to build the features of CDAs and mDAs. Such representations of CDAs and mDAs were fed into XGBoost to set up the model. The five-fold cross-validation yielded an area under the curve (AUC) of 0.9846, which was better than those of some existing state-of-the-art methods. The employment of mDAs can really enhance the model's performance and the importance analysis on meta-path-induced networks shown that networks produced by the meta-paths containing validated CDAs provided the most important contributions.

Keywords: circRNA; disease; circRNA-disease association; heterogeneous network; meta-path; network embedding algorithm; miRNA-disease association

1. Introduction

Circular RNA (circRNA) is a special type of endogenous non-coding RNA. Unlike linear RNAs, circRNAs are produced through the process of back splicing, and their 3' and 5' ends are connected by either exon or intron circularization to form a covalently closed continuous loop structure. In 1976, Sanger et al. first discovered circRNA in viroid and Sendai virus particles of infected plants by electron microscopy and other techniques [1]. In 1979, Hsu et al. observed the presence of circRNA in the cytoplasm of eukaryotic cells through electron microscopy [2]. After that, more and more circRNAs were discovered in eukaryotic cells and fungal cells [3–6]. Increasing evidence shows that circRNAs are widely distributed in animals and plants, and they play important roles in many biological processes. CircRNAs can function as miRNA sponges [7], regulators of RNA-binding proteins [8] and parental gene transcription [9]. In recent years, circRNA has also been used as a new clinical diagnostic marker and a potential target for human disease treatment [10–13]. For example, Xu et al. found that circRNA Cdrlas is a therapeutic target for diabetes, which affects insulin secretion in islet cells by tricking miR-7 [10]. Cui et al. reported that hsa_circRNA_103636 is a potential new biomarker for depression [12]. Lu et al. found that hsa_circ_0063425 and hsa_circ_0056891 are biomarkers for the early stages of type 2 diabetes [13]. The aforementioned facts indicate that some circRNAs have strong associations with certain diseases. On the one hand, investigations on such associations are helpful to elucidate the mechanisms of circRNA functional roles; on the other hand, such investigations explore the pathogenesis of complex diseases. Thus, these investigations improve disease diagnoses and promote disease-targeted therapies.

To date, only a few circRNA-disease associations (CDAs) have been determined. These limited associations have become a hinderance to further studying mechanisms of circRNA functional roles and uncover the pathogenesis of diseases. Evidently, the determination of numerous CDAs can improve our understanding on circRNAs and diseases. Traditional clinical trials can provide a solid determination of CDAs; however, they always need lots of time and are very expensive. It is urgent to design quick and cheap methods to detect CDAs. Computational methods are deemed as suitable alternative tools to identify latent CDAs. Generally, abundant information on circRNAs and diseases is needed for building the computational methods. Fortunately, in recent years, some public databases, such as Circ2Disease [14], CircR2Disease [15], etc., have been set up, which are facilitated to discover hidden CDAs. On the other hand, the recent development of computational methods [16–19], especially various machine learning methods, provides more powerful technical support. These methods can deeply analyze the validated data and infer reliable and novel knowledge. In the field of CDA prediction, some computational methods have been proposed. These methods can be roughly divided into three groups: network-based methods [20–23], machine learning-based methods [24–29] and recommendation systems [30,31]. These previous methods are introduced in Section 2. Although current methods provide a high performance for the prediction of CDAs, there still exists spaces for improvement. For example, the feature representation of CDAs is far from perfect, thus decreasing the prediction quality of machine learning-based methods.

In this study, we combined the network-based and machine learning-based methods to develop a powerful model, named PCDA-HNMP, to identify CDAs, which can be deemed as the continued work of [25]. The main idea of PCDA-HNMP was to extract informative features from constructed networks, which were learnt by the downstream classification algorithm to generate efficient classification patterns. The contributions of this study were as below:

1) We proposed a prediction model to identify novel CDAs; the model was named PCDA-HNMP.

2) A heterogeneous network containing circRNAs, diseases, miRNAs and mRNAs was constructed using the currently known associations between above objects. Furthermore, the meta-paths for connecting circRNAs (diseases) were extracted from this heterogeneous network to further mine hidden associations between various circRNAs (diseases), inducing the networks for circRNAs (diseases). The powerful network embedding algorithm, mashup [32], was applied to the above networks to generate informative circRNA (disease) features.

3) To improve the learning quality, miRNA-disease associations (mDAs) were employed in PCDA-HNMP to help predict CDAs. The informative miRNAs features were obtained from bipartite networks of miRNAs and circRNAs, which were constructed based on the meta-paths connecting circRNAs and miRNAs in the heterogeneous network.

4) The features of circRNAs and diseases were combined to represent CDAs; mDAs were represented in a similar manner. PCDA-HNMP adopted the powerful classification algorithm, XGBoost (XGB) [33], to understand classification patterns from the aforementioned representations of CDAs and mDAs.

5) The five-fold cross-validation of the high performance PCDA-HNMP had an area under the curve (AUC) of 0.9846, which was better than those yielded by some existing state-of-the-art methods. Further tests showed that the employment of mDAs enhanced the model's performance; networks yielded by meta-paths containing validated CDAs provided the most important contributions to the PCDA-HNMP.

2. Related work

In this study, we investigated the problem of CDA prediction. To our knowledge, several methods have been proposed in this field, which can be roughly divided into three groups. Here, the brief descriptions on these methods were provided.

2.1. Network-based method

The network-based methods always construct networks for circRNAs, diseases or both and employ powerful network algorithms to build the methods. Two methods constructed a heterogeneous network defining circRNAs and diseases as nodes [21,23]. The edges in this network were determined by circRNA associations, disease associations and validated CDAs. Based on this heterogeneous network, the methods assigned a score to each pair of circRNAs and diseases used either KATZ [21] method or path weighted algorithms [23]. A high score indicates that the corresponding pair can be a latent CDA. Deng et al. improved the aforementioned method by employing inferred CDAs, which were obtained by validated circRNA-protein and protein-disease associations, to the heterogeneous network [22]. The KATZ method was used to measure the strength of the associations between circRNAs and diseases. The last network-based method was proposed by Li et al. [20], which was designed in a different way. Based on the validated CDAs, a bipartite network regarding circRNAs and diseases was constructed, which was fed into the DeepWalk software to yield the circRNA and disease features. The disease and circRNA topological similarities were calculated based on their features. Network consistency projection was applied to the similarity networks on circRNAs and diseases and the aforementioned bipartite network to evaluate the associations of circRNAs and diseases.

2.2. Machine learning-based method

Some methods in this group adopted traditional machine learning algorithms to build the methods. First, they extracted informative features for circRNAs and diseases, then fed them into classification algorithms to learn the classification patterns. Zhang et al. proposed the iCDA-CGR to identify CDAs [24]. First, it constructed multiple similarity matrices for either circRNAs or diseases based on circRNA sequences, circRNA-gene associations, disease semantic and validated CDA information and combined them into one similarity matrix for circRNAs (diseases). The similarity scores in the matrix were picked up as the features of circRNAs (diseases), which were fed into a support vector machine to build the iCDA-CGR. Kouhsar et al. proposed the CircWalk method, which adopted a different scheme to generate the features of circRNAs (diseases) [25]. It constructed a heterogeneous network containing circRNAs, diseases, miRNAs and mRNAs and employed DeepWalk to generate the features of circRNAs (diseases) from this network. The XGB was applied to these features to understand classification patterns.

With the successful applications of deep learning algorithms, some methods to identify CADs adopted deep learning algorithms. The functions of deep learning in these methods were either to extract high-level features or to learn deep patterns for classification. GCNCDA, which was designed by Wang et al., used similarity scores as features of circRNAs (diseases) [26]. These raw features were refined by the fast learning with graph convolutional networks (FastGCN) algorithm and the output features were fed into the forest by penalizing attributes (Forest PA) algorithm to make a prediction. CDASOR employed the k-mers embedding in circRNA sequences to represent circRNAs [27], and were refined by a 1-D convolutional neural network (CNN) and bi-directional long short-term memory (BiLSTM). As for disease representation, it adopted an ontology embedding. A fully connected layer was applied to the above representation for making a prediction. Deng et al. presented the MSPCD method to identify CDAs [28]. The association scores were collected as the raw features of circRNAs (diseases), which were used to extract high-order features by fully connected layers. Finally, the deep neural network (DNN) was adopted to make a prediction. DMFCDA used the association relationship between circRNAs and diseases as the raw features [29], which were refined by three fully connected layers. The refined circRNA and disease features were combined and fed into two fully connected layers to make a prediction.

2.3. Recommendation system

Recommendation systems always construct one or more kernels for circRNAs and diseases, respectively. These kernels were combined with an adjacency matrix of circRNAs and diseases to yield the recommendation matrix. iCircDA-MF, which was proposed by Wei and Liu, first constructed the circRNA kernel by integrating the CircRNA gene-related similarity and the Gaussian interaction profile (GIP) kernel similarity [30]. For the disease kernel, it integrated the disease semantic similarity and GIP kernel similarity. The adjacency matrix was reformulated by the weighted K nearest known neighbors (WKNKN) algorithm. Finally, it adopted a non-negative matrix factorization and solved an optimization problem to discover the optimal factorization, thereby generating the recommendation matrix. Li et al. used a similar method to build the kernels of circRNAs and diseases, though it solved a different optimization problem to produce the recommendation matrix [31].

3. Materials and methods

3.1. Benchmark dataset

An effective dataset is important to build efficient prediction models. Here, the validated CDAs were retrieved from a previous study [25]. These CDAs were collected from four public datasets: Circ2Disease [14], CircR2Disease [15], CTD [34], and CircAtlas [35]. The original dataset in [25] contained 575 validated CDAs, involving 474 circRNAs and 64 diseases. After combining different names of the same disease (e.g., hirschsprung disease and hirschsprung's disease), a total of 61 diseases was investigated in this study. Approximately 571 CDAs related to these diseases were maintained, which still covered 474 circRNAs. We termed 571 CADs as positive samples in this study. From these CADs, a matrix, denoted by A_{cd}, was constructed. $A_{cd}(j,k) = 1$ if and only if the j-th circRNA and the k-th disease comprised an association.

Generally, the negative samples are necessary to construct binary classification models. Additionally, 575 unlabeled pairs of circRNA and diseases used in [25] were employed. After removing four pairs, we obtained unlabeled pairs equivalent to the number of positive samples. In fact, these pairs were obtained by randomly pairing circRNAs and diseases. Although, some pairs may be latent CADs, their probabilities are very low, and the proportion of latent CADs is very small. This random selection of unlabeled pairs is widely used in an association prediction [26,36–38]. Accordingly, 571 unlabeled pairs were regarded as negative samples, which were combined with positive samples to constitute the dataset, denoted as *DS*.

In addition, in recent years, mDA predictions have become a hot topic [39–42]. Since circRNA and miRNA are two special types of RNAs and several studies have reported their special associations [43–45], mDAs may be helpful in predicting CDAs. Thus, for the aforementioned 61 diseases, we extracted their related miRNAs from Circ2Disease [14], HMDD [46], and Mir2Disease [47]. Approximately 912 mDAs were obtained. Additionally, these associations were termed as positive samples. From these mDAs, we also constructed a matrix, denoted by A_{di}, where $A_{di}(j,k) = 1$ if and only if the j-th disease and the k-th miRNA constituted an association. Likewise, 912 negative samples for mDAs were also generated by randomly selecting miRNAs and diseases. These positive and negative samples constituted the dataset DS_m.

3.2. Outline of the PCDA-HNMP

In this study, a new prediction model for CDAs was designed; the model was called PCDA-HNMP. Its construction procedures are illustrated in Figure 1. First, several types of associations between circRNAs, diseases, miRNAs and mRNAs were obtained from some public datasets, and a heterogeneous network containing these associations was constructed. Second, several types of meta-paths were extracted from the above-constructed heterogeneous network, which were further adopted to set up circRNA, disease and circRNA-miRNA networks. Then, the powerful network embedding algorithm, mashup, was employed to extract circRNA, disease and miRNA features. Finally, the feature representations of CDAs in *DS* and mDAs in DS_m were obtained based on above features, which were fed into a binary classification algorithm to train the classification model. In the following sections, the detailed descriptions of each stage are given.

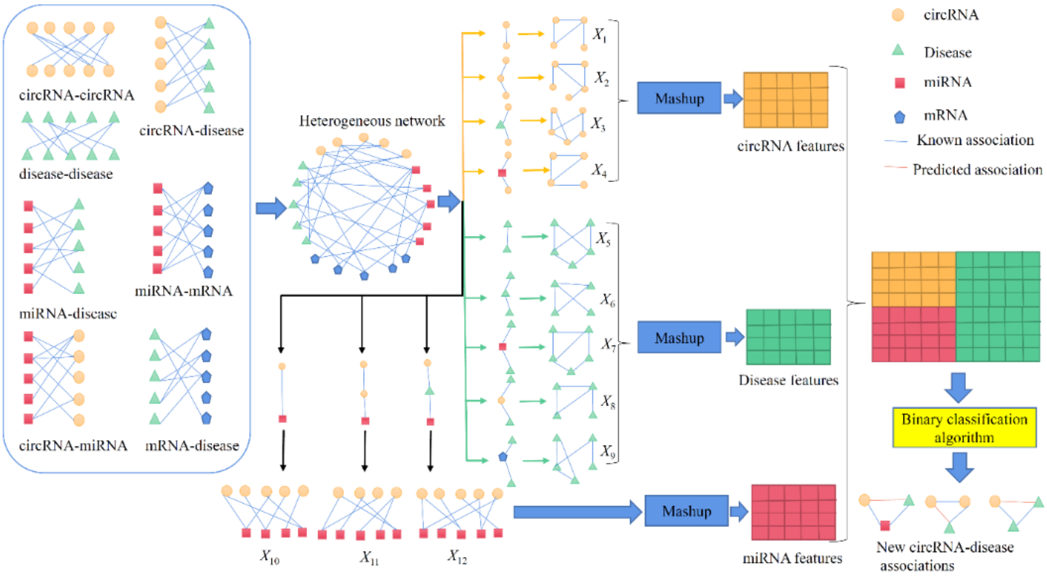

Figure 1. Entire procedures of PCDA-HNMP. Seven types of associations between circRNAs, diseases, miRNAs and mRNAs are obtained from public databases, which are used to construct a heterogeneous network. From this network, meta-paths for circRNAs, diseases and miRNAs are extracted. Based on these paths, meta-path-induced networks are built, from which circRNA, disease and miRNA features are generated by mashup. These features are refined to represent CDAs and mDAs, which are fed into one binary classification algorithm to build the model.

3.3. Heterogeneous network construction

In recent years, the use of multi-sources to design classification models is quite popular. The abundant information behind the multi-sources is helpful to improve the performance of the models. This study attempted to design a model for predicting CDAs. It is known that miRNAs and mRNAs all have special associations with various diseases; the employment of them can help to discover latent CADs. On the other hand, a network is a powerful form, which can organize all objects at a system level. Thus, a heterogeneous network was constructed, which defined circRNAs, diseases, mRNAs and miRNAs as nodes. Its construction was similar to that in [25]. Besides nodes, edges are another component of the network. As there were four types of nodes in the network, we defined seven types of edges, each of which corresponded to one type of association between circRNAs, diseases, mRNAs and miRNAs. The CDAs and mDAs are mentioned in Section 3.1. The remaining five types of associations are described below.

circRNA-circRNA association. The sequence similarity of two circRNAs was adopted to measure their association. The sequences of 474 circRNAs were obtained from CircBase [48]. These sequences were fed into the BioPython package [49] to yield the similarity of any two circRNAs. The average similarity score was set as the threshold to determine the association of two circRNAs. As a result, 46 circRNA-circRNA associations were obtained.

circRNA-miRNA association. This type of association was obtained from RAID [50] and StarBase [51]. Approximately, 1313 associations were contained in the network.

miRNA-mRNA association. This type of association was sourced from miRTarbase [52], Circ2Disease [14], and StarBase [51]. Approximately, 1194 associations were obtained.

mRNA-disease association. DisGeNet is a public database that collects the related genes of various diseases. From this database, the mRNA-disease associations were downloaded. Approximately, 1931 associations were finally used and included in the network.

Disease-disease association. The semantic similarity of any two diseases was calculated based on the tree structure of diseases in the medical subject heading (MeSH) [53] and Wang et al.'s method [54]. After setting the threshold to 0.8, 162 disease-disease associations were obtained.

From the aforementioned five types of associations, we constructed five matrices, denoted by A_{cc}, A_{ci}, A_{im}, A_{md} and A_{dd}, respectively, using the same method for constructing A_{cd} and A_{di}. The descriptions of these symbols are listed in Table 1. Furthermore, the aforementioned five types of associations and those mentioned in Section 3.1 were all integrated in the constructed heterogeneous network. This network contained 2532 nodes and 6129 edges. The detailed information of this network is provided in Table 2. For convenience, this network is denoted by *HN*.

Table 1. Descriptions of symbols used in this study.

Symbol	Descriptions
A_{cd}	Association matrix between circRNAs and diseases
A_{di}	Association matrix between diseases and miRNAs
A_{cc}	Association matrix between circRNAs
A_{ci}	Association matrix between circRNAs and miRNAs
A_{im}	Association matrix between miRNAs and mRNAs
A_{md}	Association matrix between mRNAs and diseases
A_{dd}	Association matrix between diseases
V_j^i	Raw feature vector of the *i*-th node in the *j*-th network yielded by RWR algorithm
X^i	Final feature vector of the *i*-th node
W_j^i	Context feature vector of the *i*-th node in the *j*-th network
A_j^i	Reconstructed feature vector of the *i*-th node in the *j*-th network
A_{jk}^i	The *k*-th component of the reconstructed feature vector of the *i*-th node in the *j*-th network

Table 2. Statistics for the heterogeneous network.

Type	Entry	Number
Nodes	circRNA	474
	Disease	61
	miRNA	632
	mRNA	1365
Association	circRNA⟵⟶ disease	571
	circRNA⟵⟶circRNA	46
	circRNA⟵⟶miRNA	1313
	miRNA⟵⟶disease	912
	miRNA⟵⟶mRNA	1194
	mRNA⟵⟶disease	1931
	disease⟵⟶disease	162

3.4. Meta-path-induced networks

A meta-path is a widely used concept when constructing models based on networks [55–57]. The hidden association information can be extracted from a given network using a meta-path. From the heterogeneous network *HN* constructed above, we can further infer the circRNA-circRNA associations based on the paths connecting two circRNAs. For example, two circRNAs, c_1 and c_2, were not directly connected in *HN* but have a common neighbor, say miRNA mi_1. The meta-path c_1-mi_1-c_2 indicated the special association between them. In a similar manner, the inferred associations for other objects can be extracted. By employing these inferred associations, more abundant information for circRNAs, diseases and miRNAs can be obtained, thereby helping construct more efficient classification models.

Based on the length and inner nodes of meta-paths, we extracted several types of meta-paths for circRNAs, diseases and miRNAs. Since the number of meta-paths sharply increases with an increase in path length, we only considered the meta-paths with a length less than three. With these meta-paths, the induced networks for circRNAs, diseases and miRNAs were constructed.

3.4.1. Meta-path-induced networks for circRNAs

For circRNA, four types of meta-paths were extracted from *HN*: circRNA-circRNA (meta-path-1); circRNA-circRNA-circRNA (meta-path-2); circRNA-disease-circRNA (meta-path-3); and circRNA-miRNA-circRNA (meta-path-4). The adjacency matrices for the induced networks of the aforementioned meta-paths can be obtained as follows:

$$\begin{cases} X_1 = A_{cc} & for\ meta-path-1 \\ X_2 = \chi_{Z^+}(A_{cc} \times A_{cc}) & for\ meta-path-2 \\ X_3 = \chi_{Z^+}(A_{cd} \times A_{cd}^T) & for\ meta-path-3 \\ X_4 = \chi_{Z^+}(A_{ci} \times A_{ci}^T) & for\ meta-path-4 \end{cases} \tag{1}$$

where $\chi_{Z^+}(\cdot)$ is a characteristic function of the positive integer set Z^+ on each member in the matrix, and A_{cc}, A_{cd}, and A_{ci} are association matrices between circRNAs, circRNAs and diseases, and circRNAs and miRNAs, respectively (see Table 1 for detailed descriptions).

3.4.2. Meta-path-induced networks for diseases

Likewise, five types of meta-paths were extracted from *HN* for diseases: disease-disease (meta-path-5); disease-disease-disease (meta-path-6); disease-miRNA-disease (meta-path-7); disease-circRNA-disease (meta-path-8); and disease-mRNA-disease (meta-path-9). The adjacency matrices for the induced networks of above meta-paths can be accessed by the following:

$$\begin{cases} X_5 = A_{dd} & for\ meta-path-5 \\ X_6 = \chi_{Z^+}(A_{dd} \times A_{dd}) & for\ meta-path-6 \\ X_7 = \chi_{Z^+}(A_{di} \times A_{di}^T) & for\ meta-path-7 \\ X_8 = \chi_{Z^+}(A_{cd}^T \times A_{cd}) & for\ meta-path-8 \\ X_9 = \chi_{Z^+}(A_{md}^T \times A_{md}) & for\ meta-path-9 \end{cases} \tag{2}$$

where $\chi_{Z^+}(\cdot)$ is a characteristic function of the positive integer set Z^+ on each member in the matrix, and A_{dd}, A_{di}, A_{cd}, and A_{md} are association matrices between diseases, diseases and miRNAs, circRNAs and diseases, and mRNAs and diseases, respectively (see Table 1 for detailed descriptions).

3.4.3. Meta-path-induced networks for miRNAs

The meta-paths for miRNAs were different from those for circRNA and diseases because miRNA was an aid for the prediction of CDAs. In the training procedure, the role of miRNA was the same as circRNA. The meta-paths for miRNAs should be related to circRNAs. Thus, we adopted the following three types of meta-paths: circRNA-miRNA (meta-path-10); circRNA-circRNA-miRNA (meta-path-11); and circRNA-disease-miRNA (meta-path-12). Similarly, the adjacency matrices of their corresponding induced networks were constructed by the following:

$$\begin{cases} X_{10} = A_{ci} & for\ meta-path-10 \\ X_{11} = \chi_{Z^+}(A_{cc} \times A_{ci}) & for\ meta-path-11 \\ X_{12} = \chi_{Z^+}(A_{cd} \times A_{di}) & for\ meta-path-12 \end{cases} \tag{3}$$

where $\chi_{Z^+}(\cdot)$ is a characteristic function of the positive integer set Z^+ on each member in the matrix, and A_{ci}, A_{cc}, A_{cd}, and A_{di} are association matrices between circRNAs and miRNAs, circRNAs, circRNAs and diseases, and diseases and miRNAs, respectively (see Table 1 for detailed descriptions).

3.5. Feature extraction

The meta-path-induced networks included abundant information of circRNAs, diseases and miRNAs. It was challenging to quantify this information and extract the informative features for them. Some network embedding algorithms have been proposed to extract the essential features for nodes from one or more networks, such as DeepWalk [58], node2vec [59], etc. These algorithms can abstract linkages in the network and assign features to each node. As multiple networks were built for each of the circRNAs, diseases and miRNAs, we selected the powerful network embedding algorithm, mashup [32], which is the only one network embedding algorithm that can tackle more than one networks at one time. Its brief description is described below.

Given *m* networks, denoted as $N_1, N_2, \cdots, N_m$, the random walk with restart (RWR) algorithm [60,61] is applied on each network, where each node is selected one by one as a seed node. When the i-th node in N_j is set as the seed node, the RWR algorithm assigns a probability to each node in N_j. The vector collecting these probabilities is picked up as the raw feature vector of such node, denoted by V_j^i. Such raw feature vectors always have large dimensions and multiple vectors can be obtained for the same node, which are derived from different networks. Thus, the following procedures from mashup fuse different feature vectors for the same node into a unified vector and reduce the dimension. Let X^i be the unified feature vector of the *i*-th node in all networks and W_j^i be the context feature vector of the *i*-th node in the *j*-th network. Based on X^i and W_j^i, the vector for the *i*-th node in the *j*-th network can be constructed and formulated by the following:

$$A_j^i = \left(A_{j1}^i, A_{j2}^i, \cdots, A_{jn}^i\right)^T \tag{4}$$

where *n* is the total number of nodes and its *k*-th component A_{jk}^i $(1 \leq k \leq n)$ is defined by the following:

$$A_{jk}^i = \frac{\exp\left((X^i)^T W_j^k\right)}{\sum_{k'} \exp\left((X^i)^T W_j^{k'}\right)} \tag{5}$$

where X^i is the unified feature vector of the *i*-th node and W_j^k is the context feature vector of the *k*-th node in the *j*-th network. Generally, A_j^i should be similar to V_j^i as much as possible so that the best values in X^i and W_j^i can be obtained. Thus, mashup solves the following optimization problem:

$$\underset{X^i, W_j^i}{Minimize} \ \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} D_{KL}(V_j^i || A_j^i) \tag{6}$$

where D_{KL}represents the function of KL-divergence (relative entropy), and *n* and *m* are the number of nodes and networks, respectively.

The current study adopted the mashup program obtained from http://cb.csail.mit.edu/cb/mashup/, which was performed with its default parameters. This program was applied to the meta-path-induced networks of circRNAs, diseases, and miRNAs to generate features of circRNAs, diseases, miRNAs, respectively. Particularly, for the feature vectors derived from meta-path-induced networks for miRNAs, we only picked up the feature vectors of miRNAs and discarded the features for circRNAs.

3.6. Binary classification algorithm

The features of circRNAs, diseases, miRNAs were yielded by mashup. For CDAs in *DS*, the features of circRNAs and diseases were combined, whereas the features of miRNAs and diseases were aggregated for mDAs in DS_m. The feature vectors of CDAs and mDAs, along with their labels (1 for positive and 0 for negative), were collected as the underlying dataset. A classification algorithm was necessary to build the binary classification model on the above dataset. In this study, six classification algorithms were attempted, including support vector machine (SVM) [62], logistic regression (LR) [63], random forest (RF) [64], AdaBoost and random forest as base classifier (ABRF) [65], XGB [33] and multilayer perceptron (MP) [66]. Based on different classification algorithms, the models were tested by a cross-validation method [67], thereby building the optimal classification model. The corresponding Python packages in scikit-learn [68] were adopted to implement the aforementioned six algorithms.

3.7. Evaluation method and measurements

In this study, we adopted a five-fold cross-validation [67] to evaluate the performance of all models. In the original five-fold cross-validation, samples are divided into five parts. Each part was selected as a test dataset one by one, whereas the remaining parts constituted the training dataset. The model built on the training dataset is applied to the test dataset. Finally, each sample is tested exact once. This study slightly changed this procedure as mDAs in DS_m were used to improve the prediction quality and were always poured into the training dataset. In detail, CDAs in *DS* were randomly and equally divided into five parts. Each part was singled out as test dataset one by one, whereas the rest four parts and DS_m comprised the training dataset. In this case, only the CDAs in *DS* were tested.

The predicted results of a binary classification model can be counted as four entries: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Based on these entries, some measurements can be computed. This study adopted sensitivity (SN) (same as recall), specificity (SP), precision, accuracy (ACC) and F1-measure [69–74], which can be calculated by the following:

$$\begin{cases} SN = \frac{TP}{TP+FN} \\ SP = \frac{TN}{TN+FP} \\ Precision = \frac{TP}{TP+FP} \\ ACC = \frac{TP+TN}{TP+TN+FP+FN} \\ F1 - measure = \frac{2\times Precision\times Recall}{Precision+Recall} \end{cases} \tag{7}$$

The above measurements only evaluate the performance under a fixed threshold for determining the positive sample, which is generally set to 0.5. A receiver operating characteristic (ROC) curve can reflect the performance of models under various thresholds. By setting different values of the threshold, a group of SN and SP can be obtained. The ROC curve sets SN as the Y-axis and 1-SP as the X-axis in a coordinate system. The area under the curve (AUC) is a key measurement to assess the performance of the models. In general, the higher the AUC, the higher the performance. This study adopted AUC as the major measurement.

4. Results

This study designed a new computational model for the identification of CDAs, named PCDA-HNMP. The entire construction procedures are illustrated in Figure 1. This section performed tests on this model and proved its superiority. To execute the tests, the mashup [32] program retrieved from http://cb.csail.mit.edu/cb/mashup/ and Python packages of six classification algorithms downloaded from scikit-learn [68] were adopted. All codes used in this study are available at https://github.com/Zxy-zxy0/PCDA-HNMP.git.

4.1. Performance of models with different classification algorithms

The proposed model, PCDA-HNMP, extracted features from meta-path-induced networks. In this study, six classification algorithms were attempted to discover a proper classification algorithm for tackling these features, as mentioned in Section 3.6. The main parameters of these algorithms were tuned, and the final optimal parameters were obtained and provided in https://github.com/Zxy-zxy0/PCDA-HNMP/tree/master/Code/classifier. As for the feature dimension yielded by mashup, we also tried various values, including 10, 20, 30, 40 and 50. All models with different parameters and feature dimensions were evaluated by a five-fold cross-validation. Generally, the feature dimension may influence the performance of the model. Figure 2 shows the performance of the models with six different classification algorithms under various dimensions, which was measured by AUC. It can be observed that the influence of the feature dimension for different classification algorithms was different. SVM, LR and MP were influenced by a feature dimension more than the other three algorithms and their performance was evidently lower than ABRF, XGB and RF. The performance of models with ABRF, XGB and RF under different dimensions was almost at the same level. After simple comparisons, we can obtain the optimal feature dimension for each classification algorithm. The optimal feature dimensions for MP, SVM and LR were all 50, whereas such dimensions for XGB, ABRF and RF were all 20. The detailed performance for different classification algorithms (under their optimal feature dimensions) is listed in Table 3. Evidently, the model with XGB provided the best ACC, F1-measure, and SN, and the highest precision and SP was accessed by the models with either ABRF or RF, respectively. The ROC curves of these six models are illustrated in Figure 3. It can be observed that the model with XGB yielded the highest AUC of 98.46%, followed by the models with ABRF, RF, MP, SVM and LR. From these results, we can conclude that the model with XGB yielded the highest performance. Accordingly, the PCDA-HNMP adopted XGB as the classification algorithm and the feature dimension was set to 20.

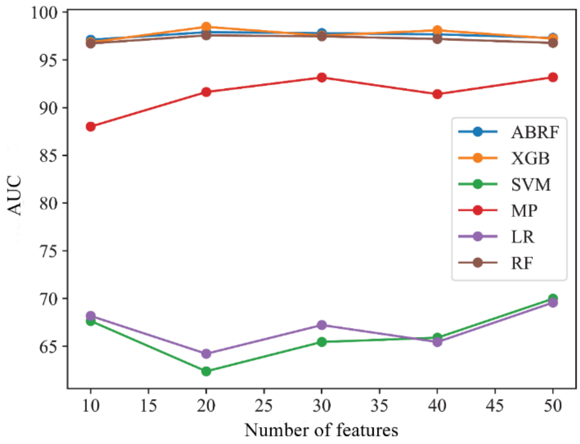

Figure 2. Trend of AUC of the models with six different classification algorithms under various feature dimensions. The AUC values of models with ABRF, XGB and RF are evidently higher and more stable than those of models with other three algorithms.

Table 3. Performance of the PCDA-HNMP with different classification algorithms.

Classification algorithm (feature dimension)	ACC (%)	F1-measure (%)	Precision (%)	SN (%)	SP (%)
XGB (20)	**94.40**	**94.39**	94.42	**94.40**	94.40
ABRF (20)	93.79	93.66	**95.51**	91.95	**95.62**
RF (20)	93.00	92.80	95.45	90.38	**95.62**
MP (50)	89.40	89.65	87.75	91.78	87.03
SVM (50)	67.86	68.43	67.22	69.71	66.02
LR (50)	64.01	63.91	64.14	63.75	64.28

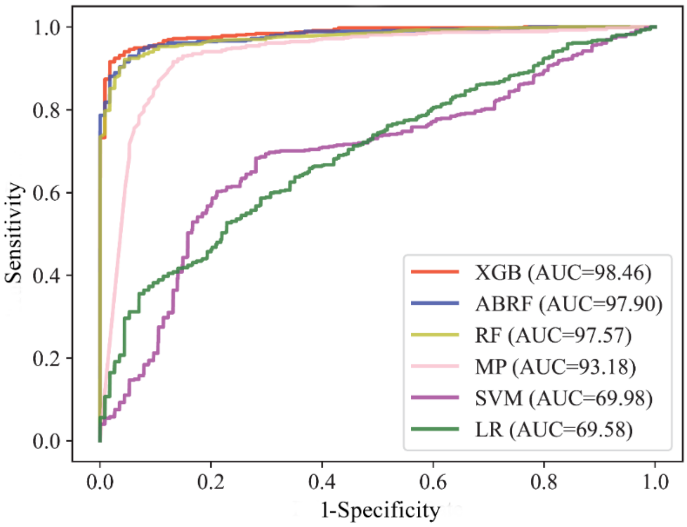

Figure 3. ROC curves of the models with six classification algorithms under their optimal feature dimensions. Obviously, the model with XGB provides the highest AUC value.

4.2. Utility of mDAs

PCDA-HNMP was designed for the identification of CDAs. From its construction procedures, the information of mDAs was employed when training the model. The purpose was to improve the training quality, thereby enhancing the performance of PCDA-HNMP. This section provides the evidence to prove that the employment of mDAs was helpful.

By removing the mDAs, we built another model. Its parameter was similar to those in PCDA-HNMP. Additionally, this model was evaluated by a five-fold cross-validation. The evaluation results are listed in Table 4. For easy comparisons, the results for PCDA-HNMP are also provided in this table (last row of Table 4). It can be observed that all measurements of PCDA-HNMP were higher than those of the model removing mDAs. For example, PCDA-HNMP improved the AUC by 0.15%, and the ACC and F1-measure were 1.32% and 1.26% higher than those of the model removing mDAs, respectively. Based on these results, the employment of mDAs can really improve the performance of PCDA-HNMP. Since there are close relationship between circRNAs and miRNAs, the existing mDAs can help predict CDAs. For example, if one miRNA is associated with one disease, the circRNA that has close relationship with this miRNA may also be associated with this disease. This is an important reason as to why the employment of mDAs can enhance the performance of PCDA-HNMP.

Table 4. Performance of the models when mDAs or meta-path-induced networks are employed or not.

Addition of mDAs	Addition of meta-path-induced networks	ACC (%)	F1-measure (%)	Precision (%)	SN (%)	SP (%)	AUC (%)
×	√	93.08	93.13	92.48	93.88	92.30	98.31
√	×	91.24	91.16	92.17	90.20	92.29	97.52
√	√	**94.40**	**94.39**	**94.42**	**94.40**	**94.40**	**98.46**

4.3. Importance of meta-path-induced networks

In PCDA-HNMP, a number of meta-path-induced networks were constructed to generate the features of circRNAs, diseases and miRNAs. To indicate the importance of these networks, some tests were conducted.

The first test was to prove the necessity of meta-path-induced networks. In fact, from the heterogeneous network *HN*, the features of circRNAs, diseases and miRNAs can be directly obtained via mashup. Based on these features, the model for the identification of CDAs can be built. Additionally, this model was evaluated by a five-fold cross-validation. The results are listed in Table 4. The ACC, F1-measure, precision, SN, SP and AUC were 91.24%, 91.16%, 92.17%, 90.20%, 92.29% and 97.52%, respectively. Compared with the corresponding measurements of PCDA-HNMP (last row of Table 4), each measurement was reduced. In detail, the AUC decreased about 1%, whereas others declined greater than 2%. These results proved that meta-path-induced networks can further mine the relationships between circRNAs, diseases and miRNAs, thereby generating more informative features to enhance the performance of PCDA-HNMP. The employment of meta-path-induced networks was a good and effective choice.

Table 5. Performance of the models when one meta-path-induced network is removed.

Meta-path-induced network	ACC (%)	F1-measure (%)	Precision (%)	SN (%)	SP (%)
X_1	92.21	92.25	91.88	92.65	91.77
X_2	92.91	92.94	92.41	93.52	92.30
X_3	**66.11**	**65.41**	**66.84**	**64.10**	**68.13**
X_4	94.05	94.06	93.75	94.40	93.70
X_5	92.30	92.31	92.16	92.47	92.12
X_6	91.59	91.63	91.34	91.94	91.24
X_7	92.21	92.23	92.00	92.47	91.95
X_8	91.33	91.42	90.58	92.30	90.37
X_9	91.42	91.48	90.87	92.13	90.72
X_{10}	92.56	92.58	92.34	92.82	92.29
X_{11}	91.42	91.53	90.44	92.65	90.20
X_{12}	91.07	91.12	90.67	91.59	90.55

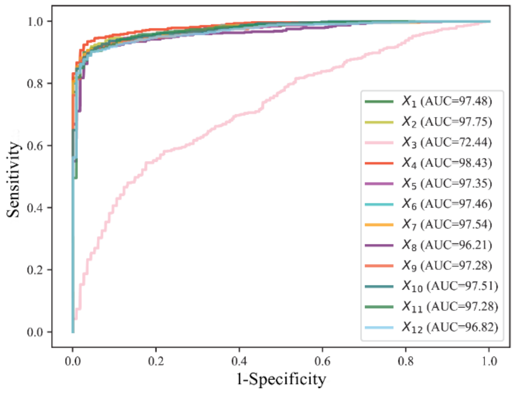

Figure 4. ROC curves of the model when one meta-path-induced network is removed. When the meta-path-induced network X_3 is removed, the AUC declined most, indicating such network is most important.

The second test further examined which meta-path-induced networks were more important (i.e., which networks provided more contributions for building PCDA-HNMP). To this end, each meta-path-induced network was removed one by one. We used rest networks to produce features and then built the model. The performance of each model is listed in Table 5 and its ROC curve, along with AUC, is provided in Figure 4. It can be found that the performance of these models were all lower than PCDA-HNMP. The removal of each meta-path-induced network reduced each measurement listed in Table 5. As for the AUC values, they also decreased compared with that of PCDA-HNMP. Each network gave positive contributions to PCDA-HNMP. Furthermore, the decline degree was not same when different networks were removed. Evidently, when the network X_3 was removed, the decline degree reached a maximum. The ACC, F1-measure and AUC dropped to 66.11%, 65.41% and 72.44%, respectively. The decline degree exceeded 25%. This result indicated that X_3 was most important for constructing PCDA-HNMP. The network X_3 was derived from meta-path-3 (circRNA-disease-circRNA). It was reasonable that the removal of currently known CDAs can provide the greatest influence to the model. As for other networks, they almost provide similar contributions. Relative speaking, networks X_8 and X_{12}, which were induced by meta-path-8 (disease-circRNA-disease) and meta-path-12 (circRNA-disease-miRNA), were more important than others. These networks were all related to associations between circRNAs and diseases. Therefore, this result was also reasonable.

4.4. Comparison with existing models

To date, several models have been designed for the identification of CDAs. To prove the superiority of PCDA-HNMP, we compared it with some existing state-of-the-art models, including DMFCDA [29], GCNCDA [26], CircWalk [25] and SIMCCDA [31]. Moreover, their performance was evaluated by a five-fold cross-validation. The obtained measurements, including the ACC, F1-measure, precision, SN and SP, are listed in Table 6. For easy comparisons, these measurements yielded by PCDA-HNMP are also provided in this table. Clearly, PCDA-HNMP provided the highest performance amongst all measurements. CircWalk gave the second highest performance. Each measurement of CircWalk was about 2.5% lower than that of PCDA-HNMP. As for the other three models (DMFCDA, GCNCDA and SIMCCDA), their performance was much lower than PCDA-HNMP. In detail, PCDA-HNMP yielded close to or more than 10% on each measurement. These results indicate that PCDA-HNMP evidently outperforms the above state-of-the-art models.

Table 6. Performance of various models under five-fold cross-validation.

Model	ACC (%)	F1-measure (%)	Precision (%)	SN (%)	SP (%)
PCDA-HNMP	**94.40**	**94.39**	**94.42**	**94.40**	**94.40**
CircWalk [25][$]	92.09	92.08	92.36	91.83	92.35
DMFCDA [29][$]	83.69	83.69	81.55	87.79	79.60
GCNCDA [26][$]	74.52	74.90	73.79	76.17	72.87
SIMCCDA [31][$]	83.36	16.40	9.10	84.54	83.34

$: The performance of these models was directly obtained from [25].

Additionally, we compared the ROC curves of PCDA-HNMP and the above four models, as illustrated in Figure 5. Similar to the results listed in Table 6, PCDA-HNMP yielded the highest AUC, which was higher than 98%. Circwalk still provided the second highest AUC, which was 97.77%. The AUC values of other three models were quite low, lower than 90%. Based on these results, we can further conclude that PCDA-HNMP was better than these existing models for the identification of CDAs.

According to the construction procedures of DMFCDA, GCNCDA and SIMCCDA, they employed limited sources. The objects of these models were only circRNAs and diseases. Few sources induced that they cannot completely describe circRNAs and diseases. This was the main reason why they were much more inferior to our model (PCDA-HNMP) and CircWalk. As for CircWalk, it employs miRNAs and mRNAs alongside circRNAs and diseases. However, it directly extracted circRNA and disease features from the heterogeneous network using DeepWalk [58]. Our model (PCDA-HNMP) gave a deep overview on the heterogeneous network and deeply mined the relationships between the circRNAs and diseases via meta-paths. Then, a more powerful network embedding algorithm, mashup, was adopted to extract features of circRNAs and diseases from the above-obtained relationships. Thus, the features used in PCDA-HNMP were more informative than those used in CircWalk. In addition, when training PCDA-HNMP, the known mDAs were employed. Since a close relationship exists between circRNAs and miRNAs, mDAs can help predict CDAs, thereby further improving the performance of PCDA-HNMP. Thus, it was logical that PCDA-HNMP outperformed CircWalk.

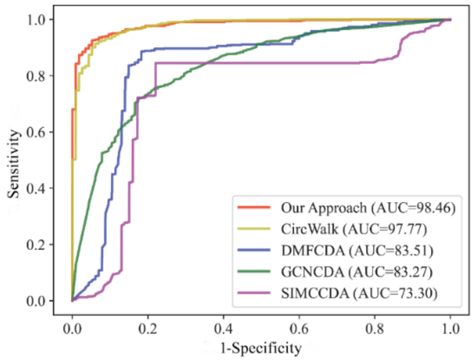

Figure 5. ROC curves of various models. Our approach, PCDA-HNMP, yields the highest AUC.

5. Discussion

In this study, we proposed a new model, PCDA-HNMP, for the identification of CADs. From the results listed in Section 4, it was found that the PCDA-HNMP had a high performance, with an AUC of 98.46%. The comparison results in Section 4.4 indicated that the model was better than four alternative methods. Compared with these methods, our model has three major advantages. The first advantage was the use of more related information of circRNAs and diseases. The heterogeneous network in PCDA-HNMP contained miRNAs and mRNAs alongside circRNAs and diseases. Several studies have been reported that miRNAs and mRNAs were highly related to circRNAs or diseases [40,41,43,45,75]. Their additions can enhance the changes to discover more hidden associations between circRNAs and diseases. Many alternative methods employed limited objects and were only used to measure the similarity between circRNAs and diseases (i.e., they were indirectly used in the methods). In our model, miRNAs and mRNAs were directly listed in the heterogeneous network, suggesting they can play more important roles in constructing the model. As for the second advantage, we employed the meta-paths to mine the hidden associations between circRNAs and diseases or circRNAs and miRNAs. The heterogeneous network was built based on the current knowledge on circRNAs, diseases, miRNAs and mRNAs. Evidently, such knowledge is far from complete. The employment of a meta-path can infer such unknown knowledge, thereby improving the model's performance, which has been proven in Section 4.3. One alternative method, CircWalk, adopted the same heterogeneous network but not apply a meta-path to mine hidden associations between circRNAs and diseases or circRNAs and miRNAs, thereby inducing a decreased performance compared to PCDA-HNMP. The last advantage is the employment of mDAs when training the model. Fusing the information of related objects was a common way to construct more efficient prediction models. However, most previous methods adopted these objects to obtain additional information of the main objects. For example, the miRNAs can be used to measure the associations between circRNAs or diseases. In general, they did not participate in the model training procedures. In our opinion, the validated mDAs were helpful to find out novel CDAs since miRNAs and circRNAs had close relationships, and the similar miRNAs and circRNAs may be related to the same disease, which can enhance the model's performance. The test results in Section 4.2 confirmed the above facts.

Besides the advantages, our model also had some disadvantages or limitations. The first disadvantage was the use of mDAs. Although the employment of mDAs can improve the model's performance, the improvement degree was limited (less than 0.2% on AUC). As miRNAs had the same roles to circRNAs when training the model, developing a method to extract miRNA features that were in the same space of the circRNA features was a challenging problem. In our model, we generated miRNA features from the bipartite networks of miRNAs and circRNAs. It was not clear whether this method was optimal. In our future work, we will design a more reasonable way to obtain miRNA features. We can potentially use the idea of graph convolution network to transfer the circRNA features to miRNAs so that circRNA and miRNA features were in the same space. PCDA-HNMP was a traditional machine learning model, which induced the second limitation. The feature extraction and classification procedures were completely separated. This meant that the features of CDAs and mDAs were not very special for the identification of CDAs. In deep learning, the end-to-end scheme gave us a new direction to build more powerful prediction models. In the future, we will fuse deep learning algorithms into our model to further enhance model's performance.

6. Conclusions

This study proposed a new model, PCDA-HNMP, for predicting circRNA-disease associations. To access informative features of circRNAs and diseases, a heterogeneous network was constructed and networks for circRNAs and diseases were built in terms of meta-paths extracted from the heterogeneous network. The test results indicated that the addition of meta-path-induced networks can really improve model's performance. Furthermore, we employed miRNA-disease associations when training the model, which can also improve the prediction quality. The superiority of PCDA-HNMP compared with some previous models indicated that it can be a useful tool to identify circRNA-disease associations. The codes and related data are available at https://github.com/Zxy-zxy0/PCDA-HNMP.git.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, A. K. Kleinschmidt, Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures, *Proc. Natl. Acad. Sci. USA*, **73** (1976), 3852–3856. https://doi.org/10.1073/pnas.73.11.3852
2. M. T. Hsu, M. Coca-Prados, Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells, *Nature*, **280** (1979), 339–340. https://doi.org/10.1038/280339a0
3. S. Memczak , M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, et al., Circular RNAs are a large class of animal RNAs with regulatory potency, *Nature*, **495** (2013), 333–338. https://doi.org/10.1038/nature11928
4. L. Chen, C. Huang, X. Wang, G. Shan, Circular RNAs in eukaryotic cells, *Curr. Genomics*, **16** (2015), 312–318. https://doi.org/10.2174/1389202916666150707161554
5. Q. Chu, X. Zhang, X. Zhu, C. Liu, L. Mao, C. Ye, et al., PlantcircBase: A database for plant circular RNAs, *Mol. Plant*, **10** (2017), 1126–1128. https://doi.org/10.1016/j.molp.2017.03.003
6. J. Salzman, R. E. Chen, M. N. Olsen, P. L. Wang, P. O. Brown, Cell-type specific features of circular RNA expression, *PLoS Genet.*, **9** (2013), e1003777. https://doi.org/10.1371/journal.pgen.1003777
7. T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, et al., Natural RNA circles function as efficient microRNA sponges, *Nature*, **495** (2013), 384–388. https://doi.org/10.1038/nature11993
8. Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, et al., Exon-intron circular RNAs regulate transcription in the nucleus, *Nat. Struct. Mol. Biol.*, **22** (2015), 256–264. https://doi.org/10.1038/nsmb.2959

9. J. T. Granados-Riveron, G. Aquino-Jarquin, The complexity of the translation ability of circRNAs, *Biochim. Biophys. Acta Gene Regul. Mech.*, **1859** (2016), 1245–1251. https://doi.org/10.1016/j.bbagrm.2016.07.009
10. H. Xu, S. Guo, W. Li , P. Yu, The circular RNA Cdr1as, via miR-7 and its targets, regulates insulin transcription and secretion in islet cells, *Sci. Rep.*, **5** (2015), 12453. https://doi.org/10.1038/srep12453
11. Q. Liu, X. Zhang, X. Hu, L. Dai, X. Fu, J. Zhang, et al., Circular RNA related to the chondrocyte ECM regulates MMP13 expression by functioning as a MiR-136 ‘Sponge’ in human cartilage degradation, *Sci. Rep.*, **6** (2016), 22572. https://doi.org/10.1038/srep22572
12. X. Cui, W. Niu, L. Kong, M. He, K. Jiang, S. Chen, et al., hsa_circRNA_103636: Potential novel diagnostic and therapeutic biomarker in Major depressive disorder, *Biomark. Med.*, **10** (2016), 943–952. https://doi.org/10.2217/bmm-2016-0130
13. Y. K. Lu, X. Chu, S. Wang, Y. Sun, J. Zhang, J. Dong, et al., Identification of circulating hsa_circ_0063425 and hsa_circ_0056891 as novel biomarkers for detection of type 2 diabetes, *J. Clin. Endocrinol. Metab.*, **106** (2021), e2688–e2699. https://doi.org/10.1210/clinem/dgab101
14. D. Yao, L. Zhang, M. Zheng, X. Sun, Y. Lu, P. Liu, Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease, *Sci. Rep.*, **8** (2018), 11018. https://doi.org/10.1038/s41598-018-29360-3
15. C. Fan, X. Lei, Z. Fang, Q. Jiang, F. X. Wu, CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases, *Database*, **2018** (2018), bay044. https://doi.org/10.1093/database/bay044
16. R. Sheikhpour, K. Berahmand, S. Forouzandeh, Hessian-based semi-supervised feature selection using generalized uncorrelated constraint, *Knowledge-Based Syst.*, **269** (2023), 110521. https://doi.org/10.1016/j.knosys.2023.110521
17. S. Forouzandeh, K. Berahmand, R. Sheikhpour, Y. Li, A new method for recommendation based on embedding spectral clustering in heterogeneous networks (RESCHet), *Expert Syst. Appl.*, **231** (2023), 120699. https://doi.org/10.1016/j.eswa.2023.120699
18. S. Forouzandeh, A. R. Aghdam, S. Forouzandeh, S. Xu, Addressing the cold-start problem using data mining techniques and improving recommender systems by cuckoo algorithm: A case study of Facebook, *Comput. Sci. Eng.*, **22** (2018), 62–73. https://doi.org/10.1109/MCSE.2018.2875321
19. S. Forouzandeh, A. Sheikhahmadi, A. R. Aghdam, S. Xu, New centrality measure for nodes based on user social status and behavior on Facebook, *Int. J. Web Inf. Syst.*, **14** (2018), 158–176. https://doi.org/10.1108/IJWIS-07-2017-0053
20. G. Li, J. Luo, D. Wang, C. Liang, Q. Xiao, P. Ding, et al., Potential circRNA-disease association prediction using DeepWalk and network consistency projection, *J. Biomed. Inf.*, **112** (2020), 103624. https://doi.org/10.1016/j.jbi.2020.103624
21. C. Fan, X. Lei, F. X Wu, Prediction of circRNA-disease associations using KATZ model based on heterogeneous networks, *Int. J. Biol. Sci.*, **14** (2018), 1950–1959. https://doi.org/10.7150/ijbs.28260
22. L. Deng, W. Zhang, Y. Shi, Y. Tang, Fusion of multiple heterogeneous networks for predicting circRNA-disease associations, *Sci. Rep.*, **9** (2019), 9605. https://doi.org/10.1038/s41598-019-45954-x
23. X. Lei, Z. Fang, L. Chen, F. X. Wu, PWCDA: Path weighted method for predicting circRNA-disease associations, *Int. J. Mol. Sci.*, **19** (2018), 3410. https://doi.org/10.3390/ijms19113410

24. K. Zheng, Z. You, J. Li, L. Wang, Z. H. Guo, Y. Huang, iCDA-CGR: Identification of circRNA-disease associations based on Chaos game representation, *PLoS. Comput. Biol.*, **16** (2020), e1007872. https://doi.org/10.1371/journal.pcbi.1007872
25. M. Kouhsar, E. Kashaninia, B. Mardani, H. R. Rabiee, CircWalk: A novel approach to predict CircRNA-disease association based on heterogeneous network representation learning, *BMC Bioinf.*, **23** (2022), 331. https://doi.org/10.1186/s12859-022-04883-9
26. L. Wang, Z. H. You, Y. M. Li, K. Zheng, Y. A. Huang, GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm, *PLoS Comput. Biol.*, **16** (2020), e1007568. https://doi.org/10.1371/journal.pcbi.1007568
27. C. Lu, M. Zeng, F. X. Wu, M. Li, J. Wang, Improving circRNA-disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks, *Bioinformatics*, **36** (2021), 5656–5664. https://doi.org/10.1093/bioinformatics/btaa1077
28. L. Deng, D. Liu, Y. Li, R. Wang, J. Liu, J. Zhang, et al., MSPCD: Predicting circRNA-disease associations via integrating multi-source data and hierarchical neural network, *BMC Bioinf.*, **23** (2022), 427. https://doi.org/10.1186/s12859-022-04976-5
29. C. Lu, M. Zeng, F. Zhang, F. X. Wu, M. Li, J. Wang, Deep matrix factorization improves prediction of human circRNA-disease associations, *IEEE J. Biomed. Health. Inf.*, **25** (2021), 891–899. https://doi.org/10.1109/JBHI.2020.2999638
30. H. Wei, B. Liu, iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization, *Briefings Bioinf.*, **21** (2020), 1356–1367. https://doi.org/10.1093/bib/bbz057
31. M. Li, M. Liu, Y. Bin, J. Xia, Prediction of circRNA-disease associations based on inductive matrix completion, *BMC Med. Genomics*, **13** (2020), 42. https://doi.org/10.1186/s12920-020-0679-0
32. H. Cho, B. Berger, J. Peng, Compact integration of multi-network topology for functional analysis of genes, *Cell Syst.*, **3** (2016), 540–548. https://doi.org/ 10.1016/j.cels.2016.10.017
33. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
34. A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers, et al., Comparative Toxicogenomics Database (CTD): Update 2021, *Nucleic Acids Res.*, **49** (2021), D1138–D1143. https://doi.org/10.1093/nar/gkaa891
35. W. Wu, P. Ji, F. Zhao, CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes, *Genome Biol.*, **21** (2020), 101. https://doi.org/10.1186/s13059-020-02018-y
36. Y. Yang, L. Chen, Identification of drug–disease associations by using multiple drug and disease networks, *Curr. Bioinf.*, **17** (2022), 48–59. https://doi.org/10.2174/1574893616666210825115406
37. X. Zhao, L. Chen, Z. Guo, T. Liu, Predicting drug side effects with compact integration of heterogeneous networks, *Curr. Bioinf.*, **14** (2019), 709–720. https://doi.org/10.2174/1574893614666190220114644
38. Z. Xian, C. Lei, L. Jing, A similarity-based method for prediction of drug side effects with heterogeneous information, *Math. Biosci.*, **306** (2018), 136–144. https://doi.org/10.1016/j.mbs.2018.09.010

39. H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, et al., Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes, *BMC Syst. Biol.*, **7** (2013), 101. https://doi.org/10.1186/1752-0509-7-101
40. X. Chen, L. Wang, J. Qu, N. Guan, J. Li, Predicting miRNA-disease association based on inductive matrix completion, *Bioinformatics*, **34** (2018), 4256–4265. https://doi.org/10.1093/bioinformatics/bty503
41. L. Zhang, B. Liu, Z. Li, X. Zhu, Z. Liang, J. An, Predicting MiRNA-disease associations by multiple meta-paths fusion graph embedding model, *BMC Bioinf.*, **21** (2020), 470. https://doi.org/10.1186/s12859-020-03765-2
42. G. Li, T. Fang, Y. Zhang, C. Liang, Q. Xiao, J. Luo, Predicting miRNA-disease associations based on graph attention network with multi-source information, *BMC Bioinf.*, **23** (2022), 244. https://doi.org/10.1186/s12859-022-04796-7
43. L. X. Guo, Z. H. You, L. Wang, C. Q. Yu, B. W. Zhao, Z. H. Ren, et al., A novel circRNA-miRNA association prediction model based on structural deep neural network embedding, *Briefings Bioinf.*, **23** (2022), bbac391. https://doi.org/10.1093/bib/bbac391
44. X. F. Wang, C. Q. Yu, L. P. Li, Z. H. You, W. Z. Huang, Y. C. Li, et al., KGDCMI: A new approach for predicting circRNA-miRNA interactions from multi-source information extraction and deep learning, *Front. Genet.*, **13** (2022), 958096. https://doi.org/10.3389/fgene.2022.958096
45. Y. Qian, J. Zheng, Y. Jiang, S. Li, L. Deng, Prediction of circRNA-miRNA association using singular value decomposition and Graph Neural Networks, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **2022** (2022), 1–9. https://doi.org/10.1109/TCBB.2022.3222777
46. Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, et al., HMDD v3.0: A database for experimentally supported human microRNA-disease associations, *Nucleic Acids Res.*, **47** (2019), D1013–D1017. https://doi.org/10.1093/nar/gky1010
47. Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, Xi. Zhang, et al., miR2Disease: A manually curated database for microRNA deregulation in human disease, *Nucleic Acids Res.*, **37** (2009), D98–104. https://doi.org/10.1093/nar/gkn714
48. P. Glažar, P. Papavasileiou, N. Rajewsky, circBase: A database for circular RNAs, *RNA*, **20** (2014), 1666–1670. https://doi.org/10.1261/rna.043687.113
49. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, et al., Biopython: Freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, **25** (2009), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163
50. Y. Yi, Y. Zhao, C. Li, L. Zhang, H. Huang, Y. Li, et al., RAID v2.0: An updated resource of RNA-associated interactions across organisms, *Nucleic Acids Res.*, **45** (2017), D115–D118. https://doi.org/10.1093/nar/gkw1052
51. J. H. Yang, J. H. Li, P. Shao, H. Zhou, Y. Q. Chen, L. H. Qu, starBase: A database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data, *Nucleic Acids Res.*, **39** (2011), D202–209. https://doi.org/10.1093/nar/gkq1056
52. H. Y. Huang, Y. C. D. Lin, J. Li, K. Y. Huang, S. Shrestha, H. C. Hong, et al., miRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database, *Nucleic Acids Res.*, **48** (2020), D148–D154. https://doi.org/10.1093/nar/gkz896
53. C. E. Lipscomb, Medical Subject Headings (MeSH), *Bull. Med. Lib. Assoc.*, **88** (2000), 265–266.

54. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, C. F. Chen, A new method to measure the semantic similarity of GO terms, *Bioinformatics*, **23** (2007), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087
55. Z. Tian, Y. Yu, H. Fang, W. Xie, M. Guo, Predicting microbe-drug associations with structure-enhanced contrastive learning and self-paced negative sampling strategy, *Briefings Bioinf.*, **24** (2023), bbac634. https://doi.org/10.1093/bib/bbac634
56. T. Kawichai, A. Suratanee, K. Plaimas, Meta-path based gene ontology profiles for predicting drug-disease associations, *IEEE Acces*s, **9** (2021), 41809–41820. https://doi.org/10.1109/ACCESS.2021.3065280
57. M. L. Zhang, B. W. Zhao, X. R. Su, Y. Z. He, Y. Yang, L. Hu, RLFDDA: A meta-path based graph representation learning model for drug–disease association prediction, *BMC Bioinf.*, **23** (2022), 516. https://doi.org/10.1186/s12859-022-05069-z
58. B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2014), 701–710. https://doi.org/10.1145/2623330.2623732
59. A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 855–864. https://doi.org/10.48550/arXiv.1607.00653
60. H. Tong, C. Faloutsos, J. Pan, Fast random walk with restart and its applications, in *Sixth International Conference on Data Mining (ICDM'06)*, (2006), 613–622. https://doi.org/10.1109/ICDM.2006.70
61. D. Smedley, S. Köhler, J. C. Czeschik, J. Amberger, C. Bocchini, A. Hamosh, et al., Walking the interactome for prioritization of candidate disease genes, *Am. J. Hum. Genet.*, **82** (2008), 949–958. https://doi.org/10.1016/j.ajhg.2008.02.013
62. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. https://doi.org/10.1007/BF00994018
63. D. R. Cox, The regression analysis of binary sequences, *J. R. Stat. Soc. B*, **20** (1958), 215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
64. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
65. R. E. Schapire, Explaining adaboost, in *Empirical Inference: Festschrift in Honor of Vladimir N Vapnik*, Springer, (2013), 37–52. https://doi.org/10.1007/978-3-642-41136-6_5
66. M. Kubat, Neural networks: A comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7, *Knowl. Eng. Rev.*, **13** (1999), 409–412. https://doi.org/10.1017/S0269888998214044
67. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence-Volume 2*, (1995), 1137–1145.
68. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, **12** (2011), 2825–2830.
69. D. M. W. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation, *arXiv preprint*, (2011), arXiv:2010.16061. https://doi.org/10.48550/arXiv.2010.16061

70. L. Chen, K. Chen, B. Zhou, Inferring drug-disease associations by a deep analysis on drug and disease networks, *Math. Biosci. Eng.*, **20** (2023), 14136–14157. https://doi.org/10.3934/mbe.2023632
71. F. Huang, M. Fu, J. Li, L. Chen, K.Y. Feng, T. Huang, et al., Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores, *Biochim. Biophys. Acta Proteins Proteomics*, **1871** (2023), 140889. https://doi.org/10.1016/j.bbapap.2023.140889
72. F. Huang, Q. Ma, J. Ren, J. Li, F. Wang, T. Huang, et al., Identification of smoking associated transcriptome aberration in blood with machine learning methods, *Biomed Res. Int.*, **2023** (2023), 5333361. https://doi.org/10.1155/2023/5333361
73. J. Ren, Y. Zhang, W. Guo, K. Feng, Y. Yuan, T. Huang, et al., Identification of genes associated with the impairment of olfactory and gustatory functions in COVID-19 via machine-learning Methods, *Life*, **13** (2023), 798. https://doi.org/10.3390/life13030798
74. C. Wu, L. Chen, A model with deep analysis on a large drug network for drug classification, *Math. Biosci. Eng.*, **20** (2023), 383–401. https://doi.org/10.3934/mbe.2023018
75. Y. Li, Z. Guo, K. Wang, X. Gao, G. Wang, End-to-end interpretable disease–gene association prediction, *Briefings Bioinf.*, **24** (2023), bbad118. https://doi.org/10.1093/bib/bbad118