PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

iPiDA-GCN: Identification of piRNA-disease associations based on Graph Convolutional Network

Jialu Hou[1], **Hang Wei**[2], **Bin Liu**[1,3]*

1 School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, **2** School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China, **3** Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing, China

* bliu@bliulab.net

Citation: Hou J, Wei H, Liu B (2022) iPiDA-GCN: Identification of piRNA-disease associations based on Graph Convolutional Network. PLoS Comput Biol 18(10): e1010671. https://doi.org/10.1371/journal.pcbi.1010671

Editor: Serdar Bozdag, University of North Texas, UNITED STATES

Received: July 7, 2022

Accepted: October 20, 2022

Published: October 27, 2022

Data Availability Statement: The data and source code are available at http://bliulab.net/iPiDA-GCN/.

Funding: This work was supported by the National Key R&D Program of China (No. 2018AAA0100100) and the National Natural Science Foundation of China (No. 62271049) to (BL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Motivation

Piwi-interacting RNAs (piRNAs) play a critical role in the progression of various diseases. Accurately identifying the associations between piRNAs and diseases is important for diagnosing and prognosticating diseases. Although some computational methods have been proposed to detect piRNA-disease associations, it is challenging for these methods to effectively capture nonlinear and complex relationships between piRNAs and diseases because of the limited training data and insufficient association representation.

Results

With the growth of piRNA-disease association data, it is possible to design a more complex machine learning method to solve this problem. In this study, we propose a computational method called iPiDA-GCN for piRNA-disease association identification based on graph convolutional networks (GCNs). The iPiDA-GCN predictor constructs the graphs based on piRNA sequence information, disease semantic information and known piRNA-disease associations. Two GCNs (Asso-GCN and Sim-GCN) are used to extract the features of both piRNAs and diseases by capturing the association patterns from piRNA-disease interaction network and two similarity networks. GCNs can capture complex network structure information from these networks, and learn discriminative features. Finally, the full connection networks and inner production are utilized as the output module to predict piRNA-disease association scores. Experimental results demonstrate that iPiDA-GCN achieves better performance than the other state-of-the-art methods, benefitted from the discriminative features extracted by Asso-GCN and Sim-GCN. The iPiDA-GCN predictor is able to detect new piRNA-disease associations to reveal the potential pathogenesis at the RNA level. The data and source code are available at http://bliulab.net/iPiDA-GCN/.

Author summary

PiRNAs play critical roles in various biological processes and the abnormal expression of piRNAs may lead to diseases. Meanwhile, several biological experiments show that piRNAs have the potential to be biomarkers or therapeutic targets to diagnose and prognosticate diseases. Some computational methods have been proposed to detect piRNA-disease associations, and provide promising results. However, with the increasing discovery of piRNA-disease associations, the existing methods fail to capture nonlinear and complex association patterns because of the limited training data and insufficient association representation. To overcome above questions, a novel computational method named iPiDA-GCN is proposed for piRNA-disease association identification based on graph convolutional networks. iPiDA-GCN constructs heterogeneous biological networks, and designs Asso-GCN and Sim-GCN modules for learning hidden association patterns in different biological networks. The experimental results show that iPiDA-GCN is able to detect new piRNA-disease associations, and outperforms the other state-of-the-art methods.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Piwi-interacting RNAs (piRNAs) are a kind of novel small non-coding RNAs (ncRNAs) with 24–35 nucleotides [1,2], often binding to Piwi-subfamily Argonaute proteins [3]. Recently, it is indicated that piRNAs play critical roles in various biological processes by emerging evidences, such as slicing transposable elements in animal's germline [4], genome defence [5], histone modification [6].

More and more studies have revealed that piRNA abnormal expression leads to many diseases, including cancers, neurodegenerative diseases, geriatric diseases, etc [7]. Several biological experiments show that piRNAs are able to be potential biomarkers or therapeutic targets to diagnose and prognosticate diseases [7,8]. Therefore, it is essential to identify piRNA-disease associations to uncover the pathogenesis of diseases by developing computational methods.

Some databases for piRNA-disease interaction have been constructed. For example, piRDisease v1.0 [9] collects 7939 manually curated associations between 4796 piRNAs and 28 diseases. NcRPheno [10] is a comprehensive ncRNA-disease database containing 1282 experimentally validated piRNA-disease associations. MNDR v3.0 [11] has been proposed to integrate different kinds of ncRNA-disease associations supported by biological literatures, where 13128 piRNA-disease associations between 13365 human piRNAs and 21 diseases are collected. The newly constructed databases store more and more piRNA-disease association information. As a result, the interactions between piRNAs and diseases become sparser and more complex. Therefore, advanced machine learning techniques which are able to fully make use of the available data are the keys to enhance the predictive performance of piRNA-disease association identification.

To reveal the complex interactions between ncRNAs and diseases, several computational methods for ncRNAs-diseases association detection have been proposed. There are mainly three categories of methods for ncRNAs-diseases association detection, including methods based on similarity measure [12,13], methods based on machine learning [14,15] and methods based on network [16]. The research on piRNA-disease association detection is still needed, because the performance of existing computational predictors is still relatively low. Recently,

several methods based on machine learning have been proposed for predicting piRNA-disease associations. For example, Wei *et.al.* [17] proposed the first piRNA-disease association predictor iPiDA-PUL employing positive unlabeled learning to select negative samples from all unlabeled piRNA-disease associations. With the development of deep learning and its efficiency in processing non-linear data, the deep learning methods are used to extract the features for piRNA-disease predictor [18–20]. For example, the iPiDA-sHN [18] predictor extracted features by Convolutional Neural Network (CNN), and then trained Support Vector Machines (SVMs) with selected high quality negative samples and positive samples. DFL-PiDA [19] based on extreme learning machine model employed a convolutional denoising auto-encoder to extract hidden features. The iPiDA-GBNN predictor [20] based on GrowNet [21] stacked auto-encoder to extract piRNA features.

The existing predictors provide promising results for detecting novel piRNA-disease associations. However, there are still three main problems: (i) With the rapid growth of data, methods based on machine learning showed lower generalization ability and failed to capture more complex and nonlinear relationships between piRNAs and diseases in larger and sparer datasets. (ii) The existing predictors all represent piRNA-disease associations by concatenating piRNA and disease attribute features, ignoring the structure semantic information of biological networks. (iii) The existing deep-learning-based methods treat piRNA-disease association data as Euclidean or grid-like structure data. In fact, the piRNA-disease associations are organized as networks, where the piRNAs or diseases are modelled as vertices, and the associations are viewed as edges. As a result, the existing methods fail to capture complex interactions among piRNA and disease entities, and learn the hidden association patterns in the graph-structured data [22]. To process the complex graph structure data efficiently with deep learning methods, graph convolutional networks (GCNs) [23,24] are proposed to generalize CNN from grid-structured data to graph-structured data, and learn node representations by capturing complex graph structure information and aggregating neighbour node information in the graph. Due to GCN's powerful ability of capturing complex structure information and potential association patterns, it has also been successfully applied to various tasks in bioinformatics, such as disease-gene association detection [25], drug-target interaction prediction [26,27] and drug repositioning [28].

Inspired by the effectiveness of GCN to capture nonlinear association patterns from complex networks, a novel computational method named iPiDA-GCN is proposed to identify piRNA-disease associations. In particular, two GCN modules are designed to capture the rich semantic information of different biological networks. Asso-GCN module is applied to learn node representations from the piRNA-disease association network, where piRNA node features are learned from associated disease nodes, and disease node features are learned from associated piRNA nodes. Sim-GCN modules are used to further learn the node representations from two homogeneous similarity networks, where piRNA node representations are obtained based on the piRNA neighbour information, and disease node representations are obtained in the same way. Finally, we treat this problem as a link prediction task, and predict the piRNA-disease association scores based on learned features. The experimental results show that iPiDA-GCN outperforms the other state-of-the-art methods, and the visualization of the prediction results further illuminates the advantages of iPiDA-GCN.

Materials and methods

Datasets

The comprehensive ncRNA–disease database MNDR v3.0 [11] (http://www.rnadisease.org/) contains the latest and largest piRNA-disease dataset among all the existing piRNA-disease

databases. The human piRNAs with sequence information are extracted from piRBase v3.0 (http://bigdata.ibp.ac.cn/piRBase/) [29]. After removing duplicate associations, 11981 experimentally verified piRNA-disease associations containing 10149 piRNAs and 19 diseases are collected. The datasets are represented as:

$$\begin{cases} \mathbb{S}_{\text{all}} = \mathbb{S}_{\text{independent}} + \mathbb{S}_{\text{benchmark}} \\ \mathbb{S}_{\text{all}} = \mathbb{S}^{+}_{\text{all}} \cup \mathbb{S}^{-}_{\text{all}} \\ \mathbb{S}_{\text{benchmark}} = \mathbb{S}_{\text{train}} + \mathbb{S}_{\text{validation}} \end{cases} \tag{1}$$

where the dataset $\mathbb{S}_{\text{all}}$ is divided into a benchmark set $\mathbb{S}_{\text{benchmark}}$ and an independent test set $\mathbb{S}_{\text{independent}}$. $\mathbb{S}^{+}_{\text{all}}$ represents the positive set containing 11981 positive associations and $\mathbb{S}^{-}_{\text{all}}$ represents the negative set containing 180850 negative associations. The benchmark set $\mathbb{S}_{\text{benchmark}}$ is randomly divided into five subsets, where four subsets are considered as the training set $\mathbb{S}_{\text{train}}$, and the remaining one is used as the validation set $\mathbb{S}_{\text{validation}}$. The hyperparameters of our method are optimized on the validation set via five-fold cross validation. The influence of hyperparameters on the performance of iPiDA-GCN is shown in S1 **Supplementary Material**

Finally, the model is evaluated on the independent test set $\mathbb{S}_{\text{independent}}$ to compare with the other related methods.

Method overview

In this section, we propose a predictor iPiDA-GCN based on GCN to predict piRNA-disease associations. The framework of iPiDA-GCN is shown in **Fig 1**, it mainly contains three steps: heterogeneous network construction (**Fig 1A**), GCN-based node feature extraction (**Fig 1B**) and association prediction for piRNAs and diseases (**Fig 1C**).

Network construction

Edge representation. There are three types of edges in the constructed piRNA-disease network. One type of the edges is the original interactions between m piRNAs (m = 10149) and n diseases (n = 19). The piRNA-disease association adjacency matrix is represented as $\mathbf{A}_{\text{PD}}$ in **Eq 2**, where $a_{i,j}$ = 1 if the i-th piRNA is associated with the j-th disease with experimental verification, otherwise $a_{i,j}$ = 0.

$$\mathbf{A}_{\text{PD}} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \tag{2}$$

The other two types of edges named 'similarity edge' are contained in the similarity subnetworks, and calculated based on the biological entities' information. Specifically, piRNA-piRNA similarity $\mathbf{S}_p$ is obtained based on piRNA sequence information, downloaded from piRBase v3.0 [29]. The sequence information contains the attribute information of non-coding RNAs, and the Smith-Waterman alignment algorithm [30] can effectively capture the functional similarities among RNAs. In this study, the piRNA sequence similarity $\mathbf{S}_p(p_i, p_k)$ between i-th piRNA p_i and k-th piRNA p_k is computed as:

$$\mathbf{S}_p(p_i, p_k) = \frac{\text{SW}(p_i, p_k)}{\sqrt{\text{SW}(p_i, p_i) \times \text{SW}(p_k, p_k)}} \tag{3}$$

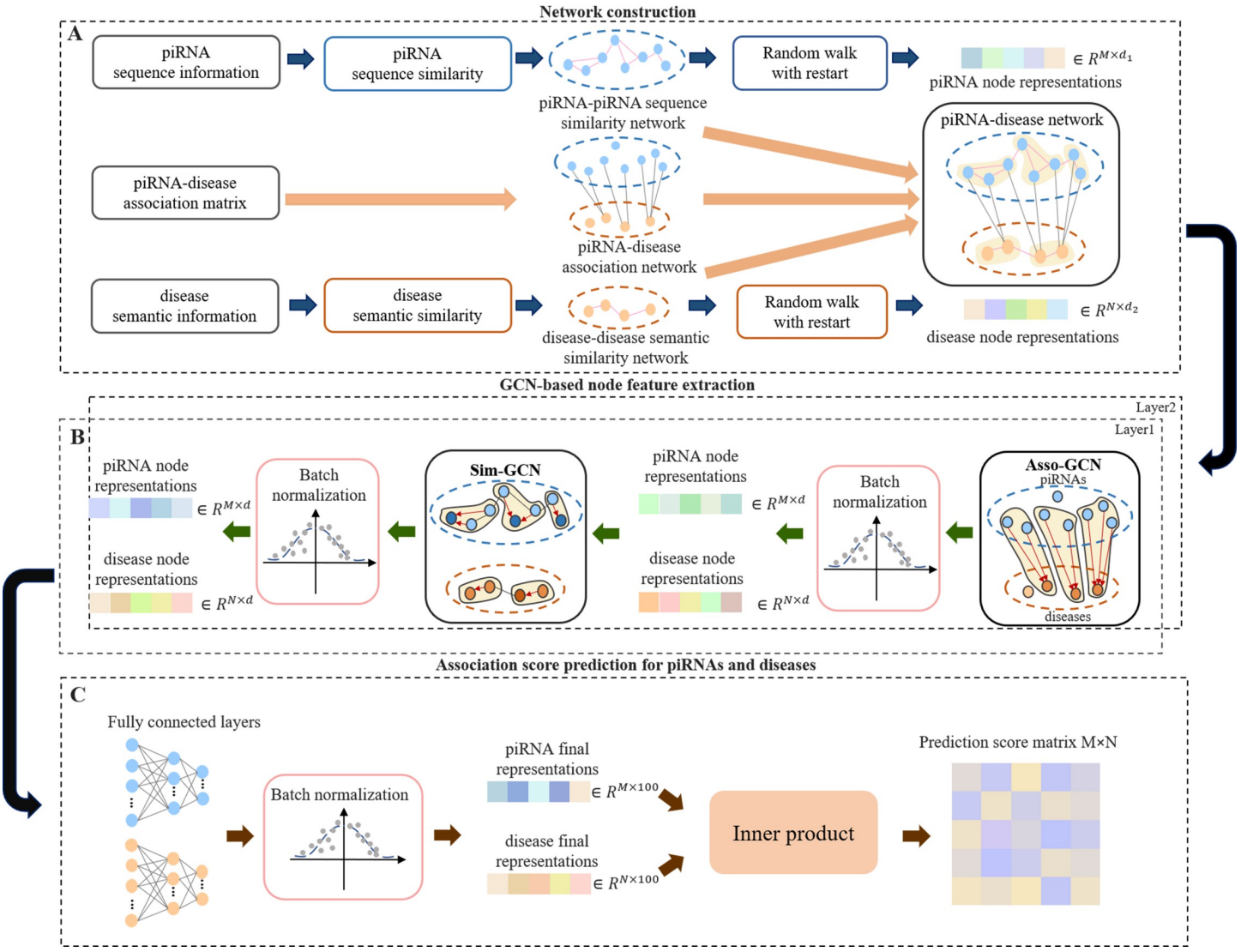

Fig 1. The flowchart of iPiDA-GCN. iPiDA-GCN mainly contains three modules: (i) Heterogeneous network construction (**Fig 1A**). Three kinds of edges are collected in the heterogeneous piRNA-disease association network, including piRNA-piRNA similarities, disease-disease similarities and piRNA-disease interactions. (ii) GCN-based node feature extraction (**Fig 1B**). Asso-GCN and Sim-GCN modules are designed to continuously learn node features from different subnetworks of piRNA-disease association network. Specifically, Asso-GCN captures hidden associated features of heterogeneous nodes from piRNA-disease interaction subnetwork, while Sim-GCN captures hidden associated features of homogeneous nodes from two similarity subnetworks. (iii) Association prediction for piRNAs and diseases (**Fig 1C**). Three fully connected layers are employed to learn the low-dimensional representations of piRNAs and diseases. Finally, association scores between piRNAs and diseases are computed through inner product operation.

https://doi.org/10.1371/journal.pcbi.1010671.g001

where the SW(p_i, p_k) is the sequence alignment value between the i-th piRNA and k-th piRNA calculated by the Smith-Waterman alignment algorithm [30].

Disease-disease similarity is computed based on disease ontology (DO) [31], which is used as a standard representation of human disease in biomedical ontologies [32]. DO is capable of translating molecular findings from high-throughput data to clinical relevance. DOSE [33] provides different semantic similarity algorithms based on DO terms. The algorithm based on Directed Acyclic Graph (DAG) [34] has been widely used in ncRNA-disease association detection [35–37]. It cannot only provide consistent semantic similarities, but also can detect potential relations between complex diseases. Therefore, the disease semantic similarity between

disease d_k and disease d_j is be calculated as [34]:

$$\mathbf{S}_d\left(d_k, d_j\right) = \frac{\sum_{t \in \mathrm{T}_k \cap \mathrm{T}_j}\left(\mathbf{S}_{d_k}(t) + \mathbf{S}_{d_j}(t)\right)}{\sum_{t \in \mathrm{T}_k}\mathbf{S}_{d_k}(t) + \sum_{t \in \mathrm{T}_j}\mathbf{S}_{d_j}(t)} \tag{4}$$

where T_k is the set containing all diseases in the DAG of disease d_k, and $\mathbf{S}_{d_k}(t)$ denotes the semantic contribution of disease $t \in \mathrm{T}_k$ to the k-th disease calculated by [34]:

$$\begin{cases} \mathbf{S}_{d_k}(t) = \max\{\alpha * \mathbf{S}_{d_k}(t\prime) | t' \in \text{children of}(t)\} & \text{if } d_k \neq d_j \\ \mathbf{S}_{d_k}(t) = 1 & \text{otherwise} \end{cases} \tag{5}$$

where α is the semantic contribution factor set as 0.5 following [34]. The farther the distance between disease t and its ancestor is, the lower the semantic contribution of disease t to disease d_k is.

Node representation. There are two types of nodes representing piRNAs and diseases in the constructed heterogeneous piRNA-disease association network. In this study, random walk with restart (RWR) [38] is employed to optimize the connectivity relationships among the same biological entities, especially for the non-neighbouring and higher-order nodes, and then the optimized similarity matrices are used as the initial feature matrices. The piRNA sequence similarity matrix and disease semantic similarity matrix are used as the input of RWR. The initial node features can be obtained by considering the global topology information of each network. The piRNA node representation generated by RWR is calculated as [38]:

$$\mathbf{P}_{i,j}^{k+1}(i) = (1-\alpha)e_{i,j} + \alpha\mathbf{P}_{i,j}^{k}(i)\mathbf{S}_p(p_i, p_j) \tag{6}$$

$$\mathbf{P}(i) = [\mathbf{P}_{i,1}^{\infty}(i), \mathbf{P}_{i,2}^{\infty}(i), \ldots, \mathbf{P}_{i,j}^{\infty}(i), \ldots, \mathbf{P}_{i,m}^{\infty}(i)] \tag{7}$$

where $\mathbf{P}_{i,j}^{k}(i)$ denotes the probability of walking from piRNA node p_i to node p_j after k steps. $e_{i,j}$ denotes the initial probability of walking from piRNA node p_i to node p_j, which is the element of an identity matrix. $\mathbf{S}_p(p_i, p_j)$ denotes the transition probability obtained from similarity matrix $\mathbf{S}_p$, α is the restart probabilities. The probability of p_i associated with all the other piRNA nodes are concatenated to generate the node representation $\mathbf{P}(i)$ for piRNA p_i. Similarly, the disease node representation $\mathbf{D}(i)$ can be represented as [38]:

$$\mathbf{D}_{i,j}^{k+1}(i) = (1-\alpha)e_{i,j} + \alpha\mathbf{D}_{i,j}^{k}(i)\mathbf{S}_{d(d_i,d_j)} \tag{8}$$

$$\mathbf{D}(i) = [\mathbf{D}_{i,1}^{\infty}(i), \mathbf{D}_{i,2}^{\infty}(i), \ldots, \mathbf{D}_{i,j}^{\infty}(i), \ldots, \mathbf{D}_{i,n}^{\infty}(i)] \tag{9}$$

To overcome the problem of insufficient representation, the feature dimension of each disease node is further extended from 19 to 1000 with polynomial features derived from the original input features, which can better reflect the interactions of different features in different dimensions. Specifically, polynomial features denote the polynomial combinations of the features with degree less than or equal to the specified degree. For example, given a disease node represented by the semantic similarities with three diseases a, b and c, it can be extended by degree-2 polynomial features as $[a, b, c, a^2, a \times b, a \times c, b^2, b \times c, c^2]$.

GCN-based node feature extraction

The key step of identifying piRNA-disease associations is node representation based on Graph Convolution Network (GCN). GCN can aggregate neighbour node information, and capture the hidden network structures to powerfully extract discriminative node features. Therefore,

we employ GCN to learn the features of piRNA and disease nodes from the heterogeneous piRNA-disease association network.

Let $\mathbf{H}^l \in \mathrm{R}^d$ denotes the node embedding of l-th GCN layer, the node embedding $\mathbf{H}^{l+1} \in \mathrm{R}^d$ is computed by (l+1)-th GCN layer according to [24] as:

$$\mathbf{H}^{l+1} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{S}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^l\mathbf{W}^l\right) \tag{10}$$

$$\tilde{\mathbf{S}} = \mathbf{I} + \mathbf{S} \tag{11}$$

$$\tilde{\mathbf{D}}(i,i) = \sum\nolimits_j \tilde{\mathbf{S}}(i,j) \tag{12}$$

where $\mathbf{S}$ is the adjacency matrix denoting the relationships among all nodes in the network, and $\mathbf{I}$ is an identity matrix. $\tilde{\mathbf{D}}$ represents the degree matrix of $\tilde{\mathbf{S}}$, $\mathbf{W}^l$ denotes the trainable parameter matrix of GCN model, $\sigma(\cdot)$ is a nonlinear activation function.

Two main modules Asso-GCN and Sim-GCN are designed to extract node representation. As shown in **Fig 1B**, Asso-GCN is adopted to aggregate node information from the piRNA-disease interaction network, $\mathrm{G}_{asso} = \{\mathrm{V}_p, \mathrm{V}_d, \mathrm{E}_{p-d}\}$ where V_p and V_d represent piRNA and disease nodes respectively, while E_{p-d} represents the interactions between piRNA and disease nodes. The piRNA node features are captured from neighbor disease node information and vice versa. Secondly, Sim-GCN is further used to capture semantic information from two different homogeneous similarity networks. The node representations obtained by Asso-GCN are viewed as initial node features in Sim-GCN module. The constructed piRNA-piRNA similarity network $\mathrm{G}_p = \{\mathrm{V}_p, \mathrm{E}_{p-p}\}$ and disease-disease similarity network $G_d = \{\mathrm{V}_d, \mathrm{E}_{d-d}\}$ are two main inputs for Sim-GCN. PiRNA node representations are generated by capturing neighbor piRNA information and disease node representations are learned from neighbor disease information. It is worth noting that batch normalization [39] is conducted following each deep learning module so as to reduce internal covariate shift and increase stability.

The reasons why we designed Asso-GCN and Sim-GCN modules to learn node representations in turn are as followings: In Asso-GCN module, GCN captures node information from the bi-partite piRNA-disease graph, where node representations are only learned from their heterogeneous neighbor nodes. However, piRNA-disease associations are too sparse to provide enough information for GCN to capture discriminative representations. Therefore, we introduce side information, including disease semantic similarity and piRNA sequence similarity. Then, we performed Sim-GCN on the similarity networks to fine-tune the node representations by their homogeneous neighbor nodes.

Association prediction for piRNAs and diseases

To further eliminate redundancy and noise, three consecutive fully connected layers are designed to extract high-level node features. There are 400, 200 and 100 neurons in each layer. Given a piRNA node representation $\mathbf{h}_{p_i}$ and a disease node representation $\mathbf{h}_{d_j}$ extracted from GCN modules, the final piRNA and disease node representation $\mathbf{h}'_{p_i}$ and $\mathbf{h}'_{d_j}$ can be obtained through the dense operation. Then the association score between piRNA p_i and disease d_j is calculated as:

$$\mathbf{U}_{i,j} = \mathbf{h}'_{p_i}\mathbf{h}'^{\mathrm{T}}_{d_j} \tag{13}$$

where $\mathbf{U}$ is the final prediction score matrix. The higher the element $\mathbf{U}_{i,j}$ is, the more likely piRNA p_i is associated with disease d_j.

The mean square error is adopted as the loss function to minimize the Frobenius norm of the difference between predicted score matrix **U** and label matrix $\mathbf{A}_{\mathrm{PD}}$. However, the number of negative associations is much more than that of positive associations. In order to alleviate the imbalance of training samples, α-enhanced loss function [25,40] focusing on positive sample learning is used, and can be formulated as:

$$\mathrm{Loss} = \|\tilde{\mathbf{A}}_{\mathrm{PD}} - \mathbf{U}\|_F^2 + \mu\|\mathbf{W}\|_2^2 \tag{14}$$

where

$$\tilde{\mathbf{A}}_{\mathrm{PD}} = \begin{cases} 0 \text{ if } \mathbf{A}_{\mathrm{PD}}(i,j) = 0 \text{ or } \mathbf{A}_{\mathrm{PD}}(i,j) \in \mathbb{S}_{\mathrm{independent}} \\ \alpha \text{ otherwise} \end{cases} \tag{15}$$

$\tilde{\mathbf{A}}_{\mathrm{PD}}$ is the enhanced association matrix generated based on the original adjacency matrix $\mathbf{A}_{\mathrm{PD}}$. α is a hyper parameter controlling the margin between true labels and predicted scores. μ is a decay factor regulating all trainable model parameters **W**. **U** is the predicted score matrix predicted by iPiDA-GCN.

Performance evaluation

PiRNA-disease association identification can be viewed as a link prediction task. Two widely used evaluation metrics, including AUC (area under the receiver operating characteristics curve) and AUPR (area under the precision recall curve) [41,42] are used to measure the performance of different methods. The higher AUC and AUPR are, the better the performance of the method is [43].

Results and discussion

The effect of GCN layers

GCN is the key module of iPiDA-GCN, which can aggregate information from neighbor nodes and obtain representations of piRNAs and diseases. The number of GCN layers has an important impact on the predictive performance. The influence of the different number of GCN layers is shown in **Table 1**, from which we can see the followings: (i) iPiDA-GCN turns to approximately randomly guess without using GCN module (layers = 0), where the input features are directly processed by fully connected layers and inner production. It achieves better performance when using GCN to capture the potential network structures. The reason is that limited GCN layers cannot capture enough structural information, while stacking more GCN layers can expand the receptive field with aggregating high-order connected node information to obtain expressive representations. (ii) When more layers are added, the performance of iPiDA-GCN gradually increases in terms of both AUC and AUPR, but its performance decreases when more than two layers are added. The reason is that more layers may introduce more noise and irrelevant information into node representation learning, leading to over-smoothing

Table 1. The impact of GCN layers on the predictive performance of iPiDA-GCN on $\mathbb{S}_{\mathrm{benchmark}}$.

Number of Layers	AUC	AUPR
0	0.5373	0.5328
1	0.6461	0.6320
2	**0.6822**	**0.6620**
3	0.6785	0.6566
4	0.6782	0.6606

https://doi.org/10.1371/journal.pcbi.1010671.t001

and performance decrement [44,45]. We conclude that GCN with two layers cannot only capture the complex interaction patterns, but also incorporate the node attribute features for representation learning so as to enhance the predictive ability.

Impact of the different components on the performance of iPiDA-GCN

There are three main components in iPiDA-GCN for extracting node features, including a fully connected network, Asso-GCN and Sim-GCN. To analyze the contributions of different components in iPiDA-GCN, three comparative baseline predictors (iPiDA-FN, iPiDA-AssoGCN and iPiDA-SimGCN) are constructed, where iPiDA-FN is constructed only by a fully connected network, while iPiDA-AssoGCN and iPiDA-SimGCN are constructed by combining different GCN modules and fully connected networks. Their performance along with iPiDA-GCN is shown in **Table 2**, from which we can see the followings: (i) Compared with iPiDA-FN, two predictors based on GCN modules achieve much better performance, indicating that GCN contributes to node representations; (ii) Sim-GCN plays a more important role than Asso-GCN for capturing semantic information from two similarity networks; (iii) iPiDA-GCN is superior to all the other baseline predictors, indicating that different components are complementary and contribute to extracting high-level node features, leading to performance improvement.

Performance comparison among different methods

To demonstrate the effectiveness of iPiDA-GCN, three state-of-the-art predictors are compared, including iPiDA-PUL [17], iPiDA-sHN [18] and piRDA [46]. All these three predictors have released the source codes or constructed the web servers, facilitating fair performance comparison. Furthermore, in order to evaluate the impact of different learning node representations on the performance of piRNA-disease association prediction, a predictor iPiDA-DW based on the node representation algorithm DeepWalk [47] is also compared with our method, which performs on the heterogeneous piRNA-disease network ignoring node attribute information, and computes association scores followed by full connection networks and inner production. The results of various methods on $\mathbb{S}_{\text{independent}}$ are listed in **Table 3**, from which we can see the followings: (i) iPiDA-GCN achieves the best performance; (ii) Compared with the methods based on node attribute iPiDA-PUL [17], iPiDA-sHN [18] and piRDA [46], iPiDA-GCN is able to capture hidden structural features, leading to better performance; (iii) Compared with iPiDA-DW which is a method based on network embedding, iPiDA-GCN cannot only incorporate hidden structural and attribute features, but also can learn discriminative node representations through two-level GCNs.

Visualization of predicted associations by iPiDA-GCN

In order to explore why iPiDA-GCN is able to accurately predict the potential associations between piRNAs and diseases, the prediction results of three piRNA-disease associations in the test set (<piR-has-1002, Parkinson's disease>, <piR-has-10009, Parkinson's disease> and

Table 2. The performance of iPiDA-FN, iPiDA-AssoGCN, iPiDA-SimGCN and iPiDA-GCN on $\mathbb{S}_{\text{independent}}$.

Method	AUC	AUPR
iPiDA-FN	0.5291	0.5107
iPiDA-AssoGCN	0.5603	0.5767
iPiDA-SimGCN	0.6765	0.6559
iPiDA-GCN	**0.7149**	**0.7036**

https://doi.org/10.1371/journal.pcbi.1010671.t002

Table 3. Performance comparison among different methods on $\mathbb{S}_{independent}$.

Method	AUC	AUPR
iPiDi-PUL[a]	0.6653	0.6550
iPiDA-sHN[b]	0.5226	0.5203
iPiDA-DW[c]	0.6317	0.6211
piRDA[d]	0.4939	0.5116
iPiDA-GCN[e]	**0.7149**	**0.7036**

[a] Results obtained by reproducing the iPiDi-PUL predictor with the help of its source code with parameters (n_components = 200, n_estimators = 150, max_features = 0.2, number of ensemble learner = 5)

[b] Results obtained by reproducing the iPiDA-sHN predictor with the help of its source code with parameters (C = 1.0, kernel = 'rbf', gamma = 1)

[c] The parameters number-walks = 10, walk-length = 80, window-size = 10

[d] The results are generated with the help of the web server of piRDA (http://nsclbio.jbnu.ac.kr/tools/piRDA/). Because piRDA is constructed based on an outdated dataset, it can only predict the piRNAs associated with 13 diseases in $\mathbb{S}_{independent}$. Therefore, only the prediction results for these associations are evaluated

[e] The parameters epoch = 2000, learning rate = 0.001, weight decay factor = 1.0.

https://doi.org/10.1371/journal.pcbi.1010671.t003

<piR-has-10111, Cardiovascular disease>) are selected, and visualized in Fig 2, from which we can see the followings: (i) iPiDA-PUL [17] and iPiDA-sHN [18] predict that piR-has-1002 and piR-has-10009 are associated with cardiovascular disease without experimental verification. iPiDA-PUL is a discriminative model based on manually constructed features, failing to learn complex association patterns. iPiDA-sHN adopts CNN to extract node features, but

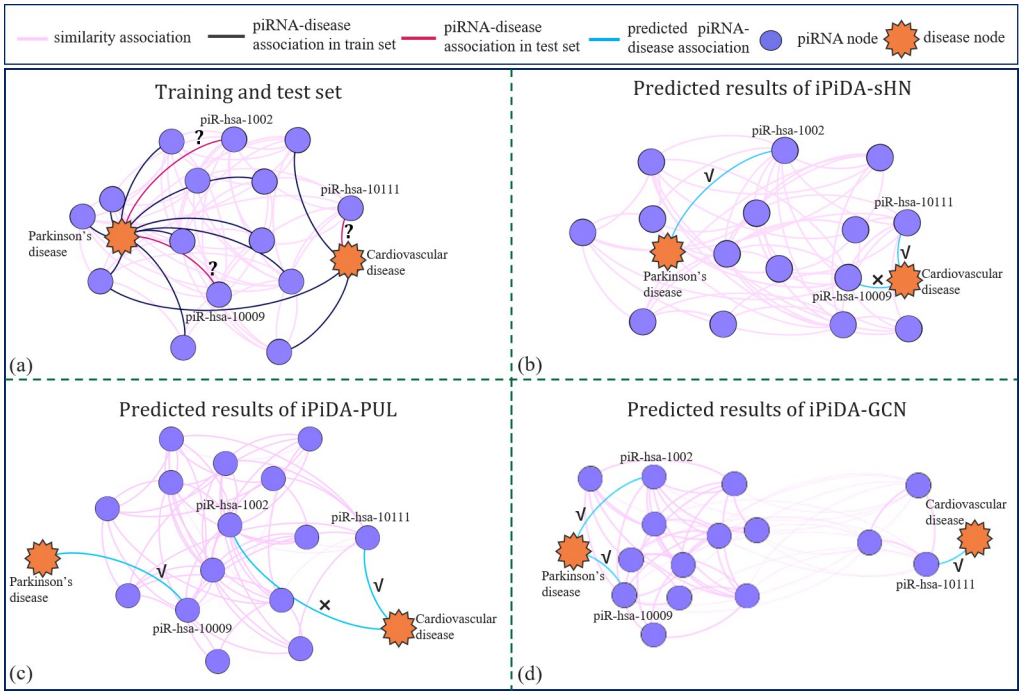

Fig 2. The prediction visualization of iPiDA-GCN and compared methods. These figures are plotted with the help of Gephi [49]. The nodes shown in orange and purple represent diseases and piRNAs, respectively. Pink lines denote the similarity associations between piRNAs, and black and red lines denote piRNA-disease associations in the training set and test set, respectively. The piRNA-disease associations predicted by different models are represented by blue lines.

https://doi.org/10.1371/journal.pcbi.1010671.g002

CNN is not suitable for analyzing the graph-structured piRNA-disease association data [48]. (ii) iPiDA-GCN correctly predicts the piRNA-disease associations in the test set, owing to its informative features learned by aggregating graph structure information from the complex piRNA-disease network. Therefore, iPiDA-GCN outperforms the other existing methods for predicting these three piRNA-disease associations.

Case study

In order to evaluate the performance of iPiDA-GCN for identifying piRNAs associated with known diseases, four important and major diseases ("Cardiovascular disease", "Renal cell carcinoma", "Alzheimer's disease" and "Parkinson's disease") are selected and their associated piRNAs are predicted by using iPiDA-GCN. **Table 4** lists the top 5 predicted piRNAs for each disease. It can be seen from **Table 4** that 19 of the 20 predicted piRNA-disease associations have been verified by the biological literatures. For example, piR-hsa-31280 is down-regulated in cardiovascular disease tissues [50]. piR-hsa-8245 is up-regulated in cardiovascular disease tissue and has a higher expression about 5-fold in cardio sphere (CS) compared with cardiosphere-derived cells (CDC) [50]. piR-hsa-10732 shows down-regulation in renal cell carcinoma tissue [51]. The expression of piR-hsa-28131 is different in Alzheimer's disease-affected brain compared with the normal human brain [3]. In addition, the top five identified piRNAs associated with Parkinson's disease (PD) are differently regulated in cells between control and PD-patients [52]. The prediction results show that iPiDA-GCN can discover new potential piRNA-disease associations, where the unconfirmed associations can be viewed as candidates to provide guidance for biological experiments in the future.

Table 4. The top 5 piRNAs associated with different diseases predicted by iPiDA-GCN.

Disease	Rank	piRNA	Evidence[a]
Cardiovascular disease	1	piR-hsa-1191	PMID:27131603
	2	piR-hsa-31280	PMID:27131603
	3	piR-hsa-8245	PMID:27131603
	4	piR-hsa-18089	PMID:27131603
	5	piR-hsa-27115	PMID:27131603
Renal cell carcinoma	1	piR-hsa-10732	PMID:26071182
	2	piR-hsa-29578	PMID:26071182
	3	piR-hsa-9186	PMID:26071182
	4	piR-hsa-19501	PMID:26071182
	5	piR-hsa-3161	PMID:26071182
Alzheimer's disease	1	piR-hsa-28131	PMID:28127595
	2	piR-hsa-2107	PMID:28127595
	3	piR-hsa-1207	PMID:28127595
	4	piR-hsa-12790	Unconfirmed
	5	piR-hsa-2106	PMID:26934981
Parkinson's disease	1	piR-hsa-356	PMID:29986767
	2	piR-hsa-6015	PMID:29986767
	3	piR-hsa-5249	PMID:29986767
	4	piR-hsa-24512	PMID:29986767
	5	piR-hsa-10122	PMID:29986767

[a] The detected piRNA-disease associations are validated by the biological literatures in PubMed. The PMIDs of these literatures are listed.

https://doi.org/10.1371/journal.pcbi.1010671.t004

Conclusion

In this study, we propose a novel computational method called iPiDA-GCN to identify piRNA-disease associations based on graph convolutional networks. Experimental results show that iPiDA-GCN is superior to the other state-of-the-art methods. Three main factors attribute to the superior performance of iPiDA-GCN: (i) Multiple biological data sources are used to construct the heterogonous piRNA-disease association network, covering more informative interactions among biological entities; (ii) Asso-GCN and Sim-GCN modules are designed to reasonably capture the graph structure information and hidden association patterns; (iii) iPiDA-GCN obtains final piRNA and disease features with three fully connected networks, which is able to filter noise, and extract meaningful information.

Besides, although iPiDA-GCN is designed for piRNA-disease association detection, it has the potential to be extend to other biological link prediction tasks, such as protein-protein interaction prediction [53], RNA-gene interaction detection [54,55].

Supporting information

S1 Supplementary Material. The hyper-parameters of GCN in iPiDA-GCN.
(PDF)

Author Contributions

Conceptualization: Jialu Hou, Bin Liu.

Data curation: Jialu Hou.

Formal analysis: Jialu Hou.

Funding acquisition: Bin Liu.

Investigation: Jialu Hou.

Methodology: Jialu Hou.

Project administration: Jialu Hou, Hang Wei, Bin Liu.

Resources: Jialu Hou, Hang Wei, Bin Liu.

Software: Jialu Hou.

Supervision: Jialu Hou, Bin Liu.

Validation: Jialu Hou, Hang Wei, Bin Liu.

Visualization: Jialu Hou.

Writing – original draft: Jialu Hou, Hang Wei.

Writing – review & editing: Jialu Hou, Hang Wei, Bin Liu.

References

1. Liu Y, Dou M, Song X, Dong Y, Liu S, Liu H, et al. The emerging role of the piRNA/piwi complex in cancer. Molecular cancer. 2019; 18(1):1–15.
2. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006; 442(7099):203–7. https://doi.org/10.1038/nature04916 PMID: 16751777
3. Roy J, Sarkar A, Parida S, Ghosh Z, Mallick B. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Molecular BioSystems. 2017; 13 (3):565–76. https://doi.org/10.1039/c6mb00699j PMID: 28127595

4. Teixeira FK, Okuniewska M, Malone CD, Coux R-X, Rio DC, Lehmann R. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. Nature. 2017; 552(7684):268–72. https://doi.org/10.1038/nature25018 PMID: 29211718
5. Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, et al. piRNA-guided genome defense: from biogenesis to silencing. Annual review of genetics. 2018; 52:131–57. https://doi.org/10.1146/annurev-genet-120417-031441 PMID: 30476449
6. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006; 442(7099):199–202. https://doi.org/10.1038/nature04917 PMID: 16751776
7. Wang K, Wang T, Gao XQ, Chen XZ, Wang F, Zhou LY. Emerging functions of piwi-interacting RNAs in diseases. Journal of Cellular and Molecular Medicine. 2021; 25(11):4893–901. https://doi.org/10.1111/jcmm.16466 PMID: 33942984
8. Halajzadeh J, Dana PM, Asemi Z, Mansournia MA, Yousefi B. An insight into the roles of piRNAs and PIWI proteins in the diagnosis and pathogenesis of oral, esophageal, and gastric cancer. Pathology-Research and Practice. 2020; 216(10):153112. https://doi.org/10.1016/j.prp.2020.153112 PMID: 32853949
9. Muhammad A, Waheed R, Khan NA, Jiang H, Song X. piRDisease v1. 0: a manually curated database for piRNA associated diseases. Database. 2019;2019.
10. Zhang W, Yao G, Wang J, Yang M, Wang J, Zhang H, et al. ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA biology. 2020; 17(7):943–55. https://doi.org/10.1080/15476286.2020.1737441 PMID: 32122231
11. Ning L, Cui T, Zheng B, Wang N, Luo J, Yang B, et al. MNDR v3. 0: mammal ncRNA–disease repository with increased coverage and annotation. Nucleic Acids Research. 2021; 49(D1):D160–D4. https://doi.org/10.1093/nar/gkaa707 PMID: 32833025
12. Chen X, Yan CC, Zhang X, You Z-H, Deng L, Liu Y, et al. WBSMDA: within and between score for MiRNA-disease association prediction. Scientific reports. 2016; 6(1):1–9.
13. Che K, Guo M, Wang C, Liu X, Chen X. Predicting MiRNA-disease association by latent feature extraction with positive samples. Genes. 2019; 10(2):80. https://doi.org/10.3390/genes10020080 PMID: 30682853
14. Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. PLoS computational biology. 2019; 15(7):e1007209. https://doi.org/10.1371/journal.pcbi.1007209 PMID: 31329575
15. Lu C, Yang M, Luo F, Wu F-X, Li M, Pan Y, et al. Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics. 2018; 34(19):3357–64. https://doi.org/10.1093/bioinformatics/bty327 PMID: 29718113
16. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics. 2020; 36(8):2538–46. https://doi.org/10.1093/bioinformatics/btz965 PMID: 31904845
17. Wei H, Xu Y, Liu B. iPiDi-PUL: identifying Piwi-interacting RNA-disease associations based on positive unlabeled learning. Briefings in Bioinformatics. 2021; 22(3):bbaa058. https://doi.org/10.1093/bib/bbaa058 PMID: 32393982
18. Wei H, Ding Y, Liu B. iPiDA-sHN: Identification of Piwi-interacting RNA-disease associations by selecting high quality negative samples. Computational Biology and Chemistry. 2020; 88:107361. https://doi.org/10.1016/j.compbiolchem.2020.107361 PMID: 32916452
19. Ji B, Luo J, Pan L, Xie X, Peng S, editors. DFL-PiDA: Prediction of Piwi-interacting RNA-Disease Associations based on Deep Feature Learning. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021: IEEE.
20. Qian Y, He Q, Deng L, editors. iPiDA-GBNN: Identification of Piwi-interacting RNA-disease associations based on gradient boosting neural network. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021: IEEE.
21. Badirli S, Liu X, Xing Z, Bhowmik A, Doan K, Keerthi SS. Gradient boosting neural networks: Grownet. arXiv preprint arXiv:200207971. 2020.
22. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine. 2017; 34(4):18–42.
23. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems. 2016;29.
24. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016.

25. Han P, Yang P, Zhao P, Shang S, Liu Y, Zhou J, et al. GCN-MF: Disease-Gene Association Identification By Graph Convolutional Networks and Matrix Factorization. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, AK, USA: Association for Computing Machinery; 2019. p. 705–13.
26. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug–target interactions based on graph convolutional network and deep neural network. Briefings in bioinformatics. 2021; 22(2):2141–50. https://doi.org/10.1093/bib/bbaa044 PMID: 32367110
27. Zhao B-W, You Z-H, Hu L, Guo Z-H, Wang L, Chen Z-H, et al. A novel method to predict drug-target interactions based on large-scale graph representation learning. Cancers. 2021; 13(9):2111. https://doi.org/10.3390/cancers13092111 PMID: 33925568
28. Cai L, Lu C, Xu J, Meng Y, Wang P, Fu X, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Briefings in Bioinformatics. 2021; 22(6):bbab319. https://doi.org/10.1093/bib/bbab319 PMID: 34378011
29. Wang J, Shi Y, Zhou H, Zhang P, Song T, Ying Z, et al. piRBase: integrating piRNA annotation in all aspects. Nucleic acids research. 2022; 50(D1):D265–D72. https://doi.org/10.1093/nar/gkab1012 PMID: 34871445
30. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981; 147(1):195–7. https://doi.org/10.1016/0022-2836(81)90087-5 PMID: 7265238
31. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research. 2015; 43(D1):D1071–D8.
32. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic acids research. 2012; 40(D1):D940–D6. https://doi.org/10.1093/nar/gkr972 PMID: 22080554
33. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015; 31(4):608–9. https://doi.org/10.1093/bioinformatics/btu684 PMID: 25677125
34. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010; 26(13):1644–50. https://doi.org/10.1093/bioinformatics/btq241 PMID: 20439255
35. You Z-H, Huang Z-A, Zhu Z, Yan G-Y, Li Z-W, Wen Z, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS computational biology. 2017; 13(3):e1005455. https://doi.org/10.1371/journal.pcbi.1005455 PMID: 28339468
36. Chen X, You Z-H, Yan G-Y, Gong D-W. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget. 2016; 7(36):57919. https://doi.org/10.18632/oncotarget.11141 PMID: 27517318
37. Wei H, Xu Y, Liu B. iCircDA-LTR: identification of circRNA–disease associations based on Learning to Rank. Bioinformatics. 2021; 37(19):3302–10. https://doi.org/10.1093/bioinformatics/btab334 PMID: 33963827
38. Tong H, Faloutsos C, Pan J-Y, editors. Fast random walk with restart and its applications. Sixth international conference on data mining (ICDM'06); 2006: IEEE.
39. Ioffe S, Szegedy C, editors. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning; 2015: PMLR.
40. Liu L, Mamitsuka H, Zhu S. HPOFiller: identifying missing protein–phenotype associations by graph convolutional network. Bioinformatics. 2021; 37(19):3328–36. https://doi.org/10.1093/bioinformatics/btab224 PMID: 33822886
41. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27(8):861–74.
42. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747
43. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; Pittsburgh, Pennsylvania, USA: Association for Computing Machinery; 2006. p. 233–40.
44. Li G, Muller M, Thabet A, Ghanem B, editors. Deepgcns: Can gcns go as deep as cnns? Proceedings of the IEEE/CVF international conference on computer vision; 2019.
45. Rong Y, Huang W, Xu T, Huang J. Dropedge: Towards deep graph convolutional networks on node classification. arXiv preprint arXiv:190710903. 2019.

46. Ali SD, Tayara H, Chong KT. Identification of piRNA disease associations using deep learning. Computational and Structural Biotechnology Journal. 2022; 20:1208–17. https://doi.org/10.1016/j.csbj.2022.02.026 PMID: 35317234

47. Perozzi B, Al-Rfou R, Skiena S, editors. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014.

48. Estrach JB, Zaremba W, Szlam A, LeCun Y, editors. Spectral networks and deep locally connected networks on graphs. 2nd international conference on learning representations, ICLR; 2014.

49. Bastian M, Heymann S, Jacomy M, editors. Gephi: an open source software for exploring and manipulating networks. Proceedings of the international AAAI conference on web and social media; 2009.

50. Vella S, Gallo A, Nigro AL, Galvagno D, Raffa GM, Pilato M, et al. PIWI-interacting RNA (piRNA) signatures in human cardiac progenitor cells. The international journal of biochemistry & cell biology. 2016; 76:1–11. https://doi.org/10.1016/j.biocel.2016.04.012 PMID: 27131603

51. Busch J, Ralla B, Jung M, Wotschofsky Z, Trujillo-Arribas E, Schwabe P, et al. Piwi-interacting RNAs as novel prognostic markers in clear cell renal cell carcinomas. Journal of experimental & clinical cancer research. 2015; 34(1):1–11. https://doi.org/10.1186/s13046-015-0180-3 PMID: 26071182

52. Schulze M, Sommer A, Plötz S, Farrell M, Winner B, Grosch J, et al. Sporadic Parkinson's disease derived neuronal cells show disease-specific mRNA and small RNA signatures with abundant deregulation of piRNAs. Acta neuropathologica communications. 2018; 6(1):1–18.

53. Rivas JDL, Fontanillo C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. Plos Computational Biology. 2010; 6(6):e1000807. https://doi.org/10.1371/journal.pcbi.1000807 PMID: 20589078

54. Bose B, Bozdag S, editors. miRDriver: A Tool to Infer Copy Number Derived miRNA-Gene Networks in Cancer. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2019.

55. Do D, Bozdag S. Cancerin: A computational pipeline to infer cancer-associated ceRNA interaction networks. PLoS computational biology. 2018; 14(7):e1006318. https://doi.org/10.1371/journal.pcbi.1006318 PMID: 30011266