

RESEARCH ARTICLE

iPiDA-SWGCN: Identification of piRNA-disease associations based on Supplementarily Weighted Graph Convolutional Network

Jialu Hou¹, Hang Wei², 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@biiulab.net

Abstract

Accurately identifying potential piRNA-disease associations is of great importance in uncovering the pathogenesis of diseases. Recently, several machine-learning-based methods have been proposed for piRNA-disease association detection. However, they are suffering from the high sparsity of piRNA-disease association network and the Boolean representation of piRNA-disease associations ignoring the confidence coefficients. In this study, we propose a supplementarily weighted strategy to solve these disadvantages. Combined with Graph Convolutional Networks (GCNs), a novel predictor called iPiDA-SWGCN is proposed for piRNA-disease association prediction. There are three main contributions of iPiDA-SWGCN: (i) Potential piRNA-disease associations are preliminarily supplemented in the sparse piRNA-disease network by integrating various basic predictors to enrich network structure information. (ii) The original Boolean piRNA-disease associations are assigned with different relevance confidence to learn node representations from neighbour nodes in varying degrees. (iii) The experimental results show that iPiDA-SWGCN achieves the best performance compared with the other state-of-the-art methods, and can predict new piRNA-disease associations.

OPEN ACCESS

Citation: Hou J, Wei H, Liu B (2023) iPiDA-SWGCN: Identification of piRNA-disease associations based on Supplementarily Weighted Graph Convolutional Network. PLoS Comput Biol 19(6): e1011242. <https://doi.org/10.1371/journal.pcbi.1011242>

Editor: Ilya Ioshikhes, CANADA

Received: November 23, 2022

Accepted: June 5, 2023

Published: June 20, 2023

Copyright: © 2023 Hou et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data and source code are available at <http://biiulab.net/iPiDA-SWGCN/>.

Funding: This work was supported by the National Natural Science Foundation of China (No. U22A2039, 62271049, 62250028 to BL) and Natural Science Basic Research Program of Shaanxi (No. 2023-JC-QN-0636 to HW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author summary

PIWI-interacting RNAs (piRNAs) are a kind of small non-coding RNAs with approximately 23–36 nucleotides in length, showing critical impact on various biological processes. It is crucial to develop computational methods for identifying piRNA-disease associations. Although several computational methods have contributed to the piRNA-disease association detection, there are still two main limitations for the further improvement of piRNA-disease association prediction: (i) The high sparsity of piRNA-disease associations prevent the predictors to accurately infer potential associations. (ii) Most existing methods focus on the Boolean connectivity of piRNA-disease associations, ignoring the correlation information with different confidence. To solve aforementioned limitations, we propose a supplementarily weighted strategy in this study. Combined with Graph Convolutional Networks (GCNs), a novel predictor named iPiDA-SWGCN is

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

proposed for piRNA-disease association prediction. The supplementarily weighted strategy for piRNA-disease association complement cannot only enrich network structure information, but also provide the association confidence information for GCN to learn node representations. As a result, the evaluation results show that iPiDA-SWGCN can effectively detect missing piRNA-disease associations, and outperforms the other state-of-the-art methods.

Introduction

PIWI-interacting RNAs (piRNAs) belong to a category of small non-coding RNAs composed of approximately 23–36 nucleotides [1,2], showing a critical impact on various biological processes by regulating gene expression at epigenetic and post-transcriptional levels [3,4]. For example, piRNAs bind to piwi proteins for regulating transposon silencing, genome rearrangement, spermiogenesis, germ stem-cell maintenance, etc [5,6].

Emerging evidence indicate that piRNAs participate in many disease genesis and prognosis [7–9]. For example, piR-36712 is downregulated in breast cancer by suppressing cell proliferation, invasion and migration through the combination with SEPW1P RNA [10]. The piR-651 shows upregulated expression in gastric cancer, and tends to be associated with TNM stages [11]. Several studies highlight that piRNAs can be viewed as potential biomarkers of disease diagnosis and prognosis for effective therapeutic project design [12,13]. Therefore, developing computational methods has great significance for identifying piRNA-disease associations.

Several computational methods have been proposed to predict the associations between non-coding RNA s and diseases. These methods usually rely on network link [14], recommendation system [15], matrix completion [16], classical machine learning [17] and deep learning [18]. However, research on the detection of piRNA-disease associations is still in its preliminary stages. To unravel the complex interactions between piRNAs and diseases, several computational methods for predicting piRNA-disease associations have been proposed [19–23]. For example, iPiDi-PUL [19] and iPiDA-sHN [20] were proposed to predict piRNA-disease association based on positive-unlabeled learning. APDA [21] and iPiDA-GBNN [22] employed a stacked auto-encoder to extract piRNA-disease pair features, and then it was trained with random forests and Gradient Boosting Neural Network respectively to predict new piRNA-disease associations. DFL-PiDA [23] combined convolutional de-noising auto-encoder and extreme learning machine to identify potential associations. iPiDA-LTR [24] employed a ranking framework to integrate several component methods for piRNA-disease association detection. To further improve the representation ability of association, iPiDA-GCN [25] designed Asso-GCN and Sim-GCN models for iteratively extracting features for piRNAs and diseases.

Although the aforementioned methods have contributed to the piRNA-disease association detection, there are two main limitations for the further improvement of piRNA-disease association prediction: (i) The high sparsity of piRNA-disease associations. For example, there are only about 5% piRNA-disease associations with experimental validation in piRDisease v1.0 [26] and MNDR v3.0 [27]. The lack of association information will prevent the predictors to accurately infer the piRNA-disease associations. (ii) The Boolean associations between piRNAs and diseases. Most of the existing methods for piRNA-disease association detection utilize Boolean values to denote whether a piRNA is related with a disease or not during the training process. However, piRNAs are related with diseases with different probabilities. Boolean associations only focusing on the connectivity will ignore the confidence information.

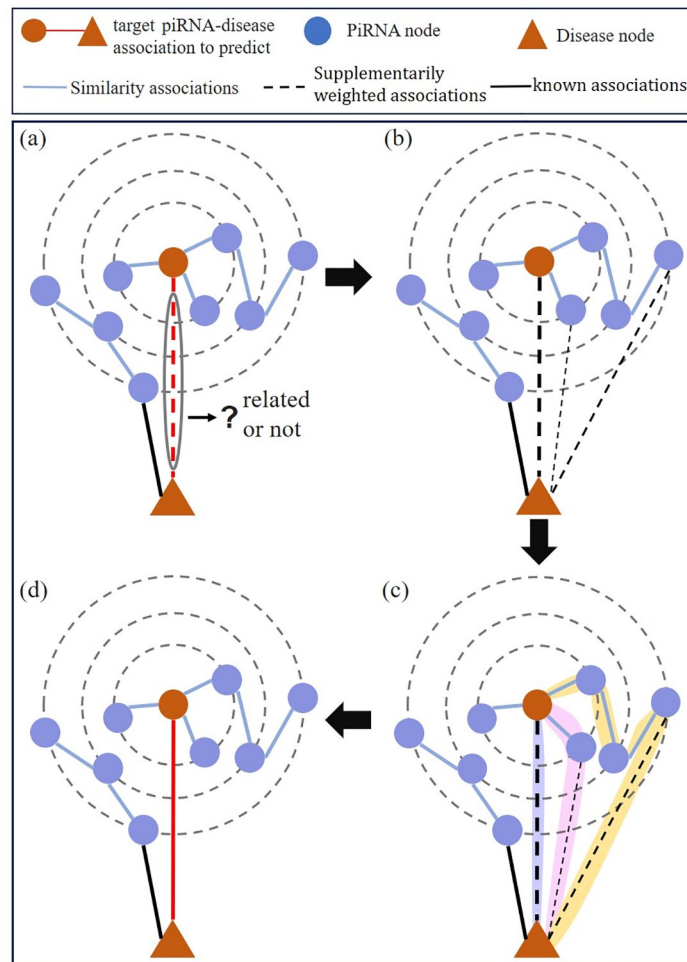


Fig 1. The effectiveness of the supplementarily weighted strategy. (a) The original piRNA-disease network is constructed based on similarity information and piRNA-disease associations. (b) PiRNA-disease network with supplementarily weighted associations. (c) Three feasible paths for piRNA-disease association detection, highlighted in purple, pink and yellow shadows. (d) The final prediction results for the target association. Integrating the inferences from three feasible paths, the final prediction result is obtained and denoted as the red line.

<https://doi.org/10.1371/journal.pcbi.1011242.g001>

To solve the above limitations, we propose a supplementarily weighted strategy to enrich the topology structure information of piRNA-disease network so as to provide more information for piRNA-disease association detection. As shown in Fig 1, our goal is to infer whether the target piRNA is associated with the target disease or not. In the original network, as the lack of link information, it is difficult to detect whether the target association exists or not (Fig 1a). Then three weighted piRNA-disease associations are preliminarily supplemented (Fig 1b). Therefore, three feasible paths can be generated to infer the association probability (Fig 1c). Finally, the prediction results are obtained by integrating different inference results based on their feasible paths (Fig 1d). As a result, supplementarily weighted associations can enrich the connectivity information, and yield more possible paths to comprehensively predict the relationship between piRNAs and diseases.

With the development of representation learning, Graph Convolutional Network (GCN) [28] is proposed to extend the CNN for graph-structured data, and achieves powerful ability of representation learning by capturing rich structural information [29]. In this paper, we combine the supplementarily weighted strategy and GCN, and propose a novel predictor named

iPiDA-SWGCN for piRNA-disease association detection with three major contributions: (i) The supplementarily weighted matrix with high-quality computed by machine learning methods can provide more information for the original sparse piRNA-disease association network, based on which GCN can aggregate more neighbour node information for expressive feature learning. (ii) The Boolean associations are replaced by weighted associations with the computed relevance confidence. With the piRNA-disease associations assigned with different weights, GCN can learn node representations from neighbour nodes in varying degrees. (iii) The evaluation results indicate that iPiDA-SWGCN has the ability to effectively detect missing piRNA-disease associations, and shows superior performance than the other state-of-the-art methods.

Materials and methods

Datasets

A recently constructed database MNDR v3.0 (<http://www.rnadisease.org/>) [27] records different categories of ncRNA-disease associations. After removing duplicated associations and extracting human-related piRNA-disease associations, 11981 experimentally validated piRNA-disease associations are selected to construct the dataset with 10149 piRNAs and 19 diseases.

In this study, machine learning predictors are trained with two phases: (i) Training several basic classifiers to compute the weights of unknown piRNA-disease associations. (ii) Training GCN to capture piRNA and disease features. In the first phase, the dataset D_{all} is split into a benchmark dataset D_{ben}^1 and an independent dataset D_{ind}^1 . We train several basic predictors on the benchmark dataset, and then utilize the trained predictors to score the associations in the independent dataset. The datasets can be formulated as:

$$\begin{cases} D_{all} = D_{ben}^1 + D_{ind}^1 \\ D_{all} = D_{all}^+ + D_{all}^- \\ D_{ben}^1 = D_{ben}^{1+} + D_{ben}^{1-} \\ D_{ind}^1 = D_{ind}^{1+} + D_{all}^- \end{cases} \quad (1)$$

where D_{all}^+ is the positive set with 11981 known piRNA-disease associations. D_{all}^- contains 180850 unknown piRNA-disease associations. D_{ben}^{1+} and D_{ben}^{1-} are randomly selected from D_{all} with the equivalent number of piRNA-disease associations, representing the positive and negative subset of D_{ben}^1 , respectively. In order to assign weights for all unknown piRNA-disease associations, the negative independent set D_{all}^- contains all unknown associations. The positive independent set D_{ind}^{1+} are constructed by randomly selecting 20% positive associations in D_{all}^+ .

In the second training phase, we divide the dataset following the previous studies [30,31]:

$$\begin{cases} D_{all} = D_{ben}^2 + D_{ind}^2 \\ D_{ben}^2 = D_{ben}^{2+} + D_{ben}^{2-} \\ D_{ind}^2 = D_{ind}^{2+} + D_{ind}^{2-} \end{cases} \quad (2)$$

where the positive benchmark dataset D_{ben}^{2+} contains 80% of positive piRNA-disease associations randomly selected from D_{all}^+ , and the remaining positive associations constitute the positive independent dataset D_{ind}^{2+} . The negative independent dataset D_{ind}^{2-} are randomly selected from D_{all}^- with the equal size of D_{ind}^{2+} , and the rest of negative samples constitute the sub-dataset D_{ben}^{2-} .

To prevent overestimating the performance of the proposed method, the piRNA-disease associations in the independent dataset for model evaluation are removed from the training phases:

$$\begin{cases} D_{ben}^{1+} \cap D_{ind}^{1+} = \emptyset \\ D_{ben}^2 \cap D_{ind}^2 = \emptyset \\ D_{ben}^1 \cap D_{ind}^2 = \emptyset \end{cases} \quad (3)$$

Method overview

In this study, we propose a novel method named iPiDA-SWGCN for piRNA-disease association detection. The overall process of iPiDA-SWGCN is shown in Fig 2 with four parts: (i) Network construction (Fig 2a). A heterogeneous piRNA-disease network is constructed by integrating piRNA and disease information; (ii) Supplementedly weighted piRNA-disease network generation (Fig 2b). Different supplementary weights are assigned to unknown piRNA-disease pairs based on the scores computed by several predictors; (iii) GCN-based feature extraction (Fig 2c). In this section, GCN is performed on the supplementedly weighted piRNA-disease network to capture the structural information, and extract feature representations of piRNAs and diseases. (iv) Association prediction (Fig 2d). Finally, we use the fully connected layers for dimension reduction and inner product operation so as to calculate the association scores between piRNAs and diseases.

Network construction

In this section, we construct a heterogeneous piRNA-disease network denoted as $G = \{E_{piR-disease}, E_{piR-piR}, E_{disease-disease}, V_{piR}, V_{disease}\}$. In detail, three types of edges and two

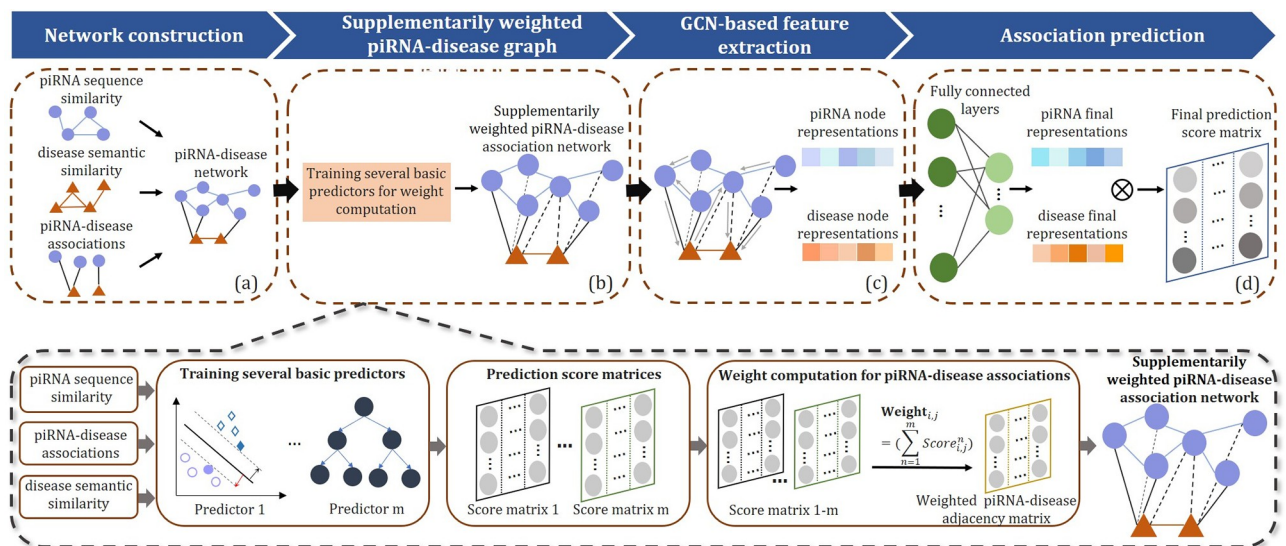


Fig 2. The flowchart of iPiDA-SWGCN. (a) piRNA-disease network is constructed by integrating piRNA sequence similarity, disease semantic similarity and piRNA-disease associations. (b) The supplementedly weighted piRNA-disease network is generated based on the proposed supplementedly weighted strategy and the weights are computed via several basic predictors. (c) piRNA and disease node features are extracted by performing GCN on the supplementedly weighted piRNA-disease network. (d) Fully connected layers and inner production are employed to predict final association scores.

<https://doi.org/10.1371/journal.pcbi.1011242.g002>

types of nodes are included in the piRNA-disease network. The three kinds of edges are piRNA-disease associations, piRNA-piRNA edges and disease-disease edges, represented as $E_{piR-disease}$, $E_{piR-piR}$ and $E_{disease-disease}$, respectively. $E_{piR-disease}$ are the original piRNA-disease associations derived from the database MNDP v3.0. The other two kinds of edges are obtained based on the similarity between homogeneous biological entities. V_{piR} and $V_{disease}$ represent the nodes of piRNAs and diseases. The specific calculation of above edge and node representations will be introduced in the following sections.

PiRNA-disease associations. The adjacency matrix A_{PD} represents the associations for each pair of piRNA-disease, denoted as:

$$A_{PD} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & & a_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \quad (4)$$

where m and n are the number of piRNAs and diseases, respectively. The element a_{ij} is 1, if the association between i -th piRNA and the j -th disease is confirmed with experimental verification, otherwise $a_{ij} = 0$.

PiRNA-piRNA sequence similarity. For each pair of piRNAs, the sequence similarity is calculated, where the sequence information is downloaded from piRBase v3.0 (<http://bigdata.ibp.ac.cn/piRBase/>) [32]. PiRNA-piRNA sequence similarity are calculated via Smith-Waterman local alignment algorithm [33], which is highly sensitive and can detect subtle similarities between sequences with low identity in a robust and accurate manner [34,35]. We compute the sequence similarity score for given pair of piRNAs as [33]:

$$P_{seq}(p_i, p_j) = \frac{SW(p_i, p_j)}{\sqrt{SW(p_i, p_i) \times SW(p_j, p_j)}} \quad (5)$$

where $SW(p_i, p_j)$ denotes the sequence alignment score between p_i and p_j calculated by Smith-Waterman alignment algorithm.

Disease-disease semantic similarity. Disease semantic similarity has been extensively used in ncRNA-disease association detection [36–39]. Directed Acyclic Graph (DAG) describes the relationship among different disease terms. Disease semantic similarity can be calculated based on the Disease Ontology (DO) descriptors in DAG [40–42]. The Directed Acyclic Graph (DAG) based algorithm not only uses a consistent standard to make the calculated disease similarity uniform and comparable but also is capable of fully capturing complex relationships between diseases and better represent the disease space and the interconnection of diseases. In this study, we adopt one of the most effective semantic similarity measurements [41,43,44] to compute the disease semantic similarity. For disease d_i and disease d_j , the disease semantic similarity $D_{sem}(d_i, d_j)$ is calculated by [44]:

$$D_{sem}(d_i, d_j) = \frac{\sum_{t \in T_i \cap T_j} (S_{d_i}(t) + S_{d_j}(t))}{\sum_{t \in T_i} S_{d_i}(t) + \sum_{t \in T_j} S_{d_j}(t)} \quad (6)$$

$$\begin{cases} S_{d_i}(t) = \max \left\{ \theta * S_{d_i}(t') \mid t' \in \text{children of } (t) \right\} & \text{if } d_i \neq t \\ S_{d_i}(t) = 1 & \text{otherwise} \end{cases} \quad (7)$$

where T_i is composed of all subterms in the DAG of disease d_i , $S_{d_i}(t)$ indicates the semantic contribution of disease $t \in T_i$ to the i -th disease according to [44]. θ is the semantic contribution factor which is set as 0.5 following [44].

Node feature extraction. The row vector of similarity matrix $\mathbf{P}_{\text{seq}}$ or $\mathbf{D}_{\text{sem}}$ can serve as the feature vector for a piRNA or disease. However, they ignore the connectivity information, especially for non-neighboring and higher-order connected nodes. Therefore, in order to introduce the global topology information of similarity networks, we further apply random walk with restart (RWR) [45] on $\mathbf{P}_{\text{seq}}$ and $\mathbf{D}_{\text{sem}}$ to extract piRNA and disease features. The feature generation for node i based on RWR can be formulated as [45]:

$$\mathbf{F}_i^{k+1} = (1 - \alpha)\mathbf{F}_i^k\mathbf{S} + \alpha y_i \quad (8)$$

where $\mathbf{F}_i^k$ is the row vector of node i , whose elements indicate the probabilities of walking from node i to all the other homogeneous nodes after k steps. $\mathbf{S}$ is the probability transition matrix obtained from similarity matrix ($\mathbf{P}_{\text{seq}}$ or $\mathbf{D}_{\text{sem}}$) with row-wise normalization, y_i is a one-hot vector representing the initial probability vector of node i , and α is the restart probabilities. Finally, we obtain $\mathbf{F}_i^\infty$ as the feature descriptor of node i .

Supplementarily Weighted Graph Convolutional Network (SWGCN)

Many unknown pairs lead to the high sparsity of piRNA-disease association network. GCN performing on such a network may lead to performance degradation because of limited neighbor information aggregation. To overcome this problem, we propose the Supplementarily Weighted Graph Convolutional Network (SWGCN). Firstly, the supplementarily weighted adjacency matrix is computed to achieve an informative piRNA-disease association network with high quality, and then GCN is adopted to capture hidden structure information for feature learning.

Weights computation for piRNA-disease association. In this section, we define the supplementarily weighted adjacency matrix in SWGCN for piRNA-disease association completion as follows:

Definition 1. Supplementarily weighted adjacency matrix. Given a network $G = \{V, E\}$, we train several basic machine-learning-based predictors to compute the average relevance scores for unknown associations among nodes. Specifically, $w_{i,j}$ represents the average score of edge $e_{i,j}$ between node n_i and node n_j , and the adjacency matrix of the completed network can be denoted as $\mathbf{A}'_{\text{PD}} \in \mathbb{R}^{m \times n}$, where $\mathbf{A}'_{\text{PD}}(i, j) = \{w_{i,j}, 1\}$, i.e., if $\mathbf{A}_{\text{PD}}(i, j) = 0$, then $\mathbf{A}'_{\text{PD}}(i, j) = w_{i,j}$, otherwise $\mathbf{A}'_{\text{PD}}(i, j) = 1$.

To improve the quality of SWGCN, 15 different basic predictors based on Random Forest [46], Support Vector Machine (SVM) [47], Gradient Boosting Decision Tree (GBDT) [48] are trained for weight computation. It should be noted that we train various predictors on different training sets, where the positive training set D_{ben}^{1+} keeps the same while the negative training sets D_{ben}^{1-} are randomly selected from D_{all}^- for five times. For a pair of piRNA p_i and disease d_j , the element $\mathbf{A}'_{\text{PD}}(i, j)$ is formulated as:

$$\mathbf{A}'_{\text{PD}}(i, j) = \begin{cases} 1 & \text{if } (p_i, d_j) \in D_{\text{ben}}^{1+} \\ \frac{\sum_{k=1}^5 (s_{\text{SVM}}^k + s_{\text{RF}}^k + s_{\text{GBDT}}^k)}{15} & \text{otherwise} \end{cases} \quad (9)$$

where s_{SVM}^k denotes the prediction score computed by the k -th SVM classifier. 15 predictors are employed to score the unknown associations, and the average scores are set as the weights for

unknown associations. Finally, the weighted edges are added into the original heterogeneous network to generate the supplementarily weighted piRNA-disease network.

Feature learning based on Graph Convolutional Network. GCN has been widely used for capturing graph structural information, and aggregating neighbour information so as to extract the node features [49,50]. After assigning weights to piRNA-disease associations, GCN is performed on the supplementarily weighted piRNA-disease network to learn the feature representations of different nodes by information aggregation from neighbours. The node embedding $\mathbf{H}^l$ learned by GCN in the l -th layer can be formulated as [28]:

$$\mathbf{H}^l = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}_{all}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{l-1}\mathbf{W}^{l-1}\right) \quad (10)$$

where

$$\mathbf{A}_{all} = \begin{bmatrix} \mathbf{P}_{seq} & \mathbf{A}'_{PD} \\ \mathbf{A}'_{PD}^T & \mathbf{D}_{sem} \end{bmatrix} \quad (11)$$

$$\tilde{\mathbf{A}}_{all} = \mathbf{I} + \mathbf{A}_{all} \quad (12)$$

$$\tilde{\mathbf{D}}(i, i) = \sum_j \tilde{\mathbf{A}}_{all}(i, j) \quad (13)$$

where the adjacency matrix $\mathbf{A}_{all} \in \mathbb{R}^{(m+n) \times (m+n)}$ indicates the association adjacency matrix for the whole network, and $\tilde{\mathbf{A}}_{all}$ is $\mathbf{A}_{all}$ added with self-loop. $\tilde{\mathbf{D}}$ denotes the degree matrix of $\tilde{\mathbf{A}}_{all}$, $\mathbf{W}^l$ represents the trainable parameters of GCN model, $\sigma(\cdot)$ denotes the nonlinear activation function. $\mathbf{H}^l \in \mathbb{R}^{(m+n) \times c}$ is initialized by concatenating the piRNA and disease representations, where c denotes the common dimension of piRNA and disease features obtained by the dense layer. It is denoted that the batch norm layer is added before each dense layer and convolution layer to alleviate the problem of vanishing or exploding gradients and speed up the convergence. The batch norm layer can standardize the mean and variance of the input to each layer based on the statistics computed over a batch of training examples to make the distribution of each layer's input relatively stable, and further reduce the risk of overfitting.

Compared with the original piRNA-disease network, GCN can effectively convolve more useful neighbor information with supplementarily weighted edges. Fig 3 shows the comparison of performing GCN on different networks. Take the disease node a as an example, on the original piRNA-disease network, the first layer of GCN updates the feature representation of node a by aggregating the features of its first-order neighbor node $b1$ and $b2$ (Fig 3a and 3b), failing to capture the indirectly connected node information of $c1, c2, c3, d1, d2, d3$. In contrast, the node representation of a can be learned by aggregating all piRNA node information with different concerns on the supplementarily weighted piRNA-disease association network (Fig 3c and 3d). As a result, SWGCN is able to capture deeper and wider neighbor information for informative feature learning.

Association prediction

After obtaining the feature representation learned by GCN, the full connection networks with three layers are constructed to separately reduce the dimensions of representations for piRNAs and diseases. The probability of piRNA p_i associated with disease d_j is calculated by inner production as:

$$\mathbf{U}_{ij} = \mathbf{f}'_{p_i} \mathbf{f}'_{d_j}^T \quad (14)$$

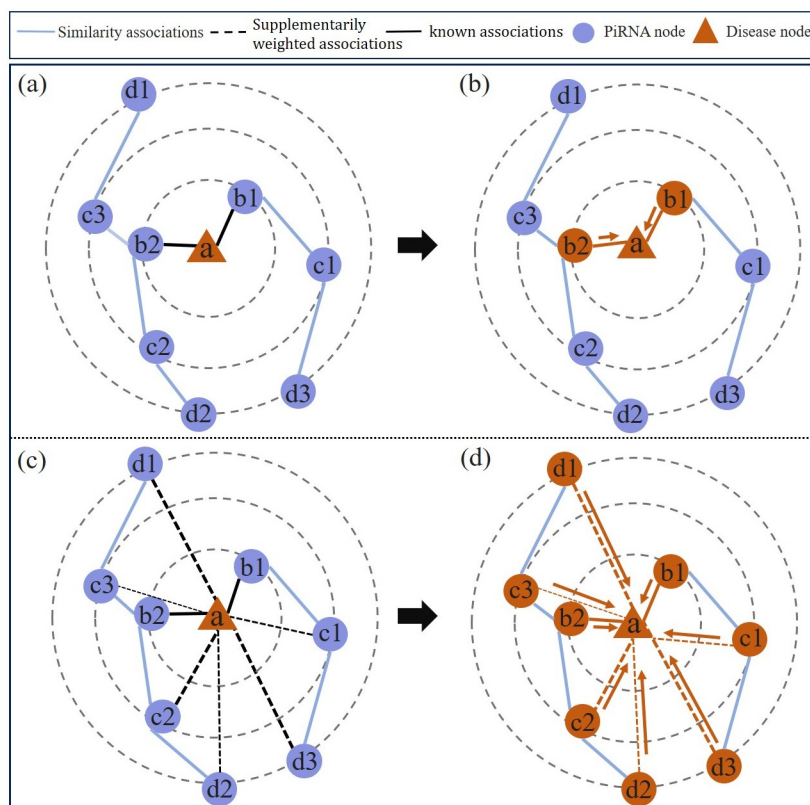


Fig 3. Comparison of performing GCN on different networks to aggregate neighbor information. Fig 3a and 3b show the receptive field of 1-layer of GCN on the original piRNA-disease association network, while Fig 3c and 3d show the receptive field of 1-layer GCN on the supplementarily weighted association network.

<https://doi.org/10.1371/journal.pcbi.1011242.g003>

where f'_{p_i} and f'_{d_j} denote the feature representation of piRNA p_i and disease d_j after dimensionality reduction, and $\mathbf{U}$ is the predicted final score matrix for the associations between piRNAs and diseases. It should be noted that we conduct Batch Normalization (BN) [51] following each convolution layer to mitigate internal covariate shift and increase the stability.

We utilize the mean square error as the loss function which minimizes the Frobenius norm of the difference between prediction score matrix $\mathbf{U}$ and the label matrix $\mathbf{A}_{PD}$. Nevertheless, the high sparsity of piRNA-disease association matrix may cause prediction bias to the unknown associations. To alleviate this problem, we adopt β -enhanced loss function [52] that enlarges the margin between the real label matrix $\mathbf{A}_{PD}$ and predicted score matrix $\mathbf{U}$ with hyper-parameter α . The loss function is formulated as:

$$\text{Loss} = \|\hat{\mathbf{A}}_{PD} - \mathbf{U}\|_F^2 + \mu \|\mathbf{W}\|_2^2 \quad (15)$$

where

$$\hat{\mathbf{A}}_{PD} = \begin{cases} 0 & \text{if } \mathbf{A}_{PD}(i, j) = 0 \text{ or } \mathbf{A}_{PD}(i, j) \in D_{ind}^2 \\ \beta & \text{otherwise} \end{cases} \quad (16)$$

where $\hat{\mathbf{A}}_{\text{pd}}$ indicates the enhanced association matrix. $\mathbf{W}$ denotes the trainable parameters. μ is a decay factor controlling the regularization term of $\mathbf{W}$ to prevent overfitting.

Performance evaluation

Two metrics are employed to comprehensively evaluate the predictor performance, including AUPR (area under the precision recall curve) [53] and AUC (area under the receiver operating characteristics curve) [54]. AUC measures the sensitivity and specificity of the model [55], and AUPR can avoid the impact of imbalanced data sets, and comprehensively reflect the quality of predictions [56].

Results and discussion

Combination of basic methods can improve the quality of SWGCN

In this study, three basic machine learning methods (RF, SVM and GBDT) are used to compute weights for the unknown piRNA-disease pairs. To analyze their contribution to weighting associations in the proposed model, we compare the predictors based on different basic methods and their combinations. Table 1 lists the results, and their performance differences are shown in Fig 4, from which we can draw the following conclusions: (i) Assigning weights to associations indeed contributes to performance improvement. (ii) It is not surprising that integrating several complementary methods for weights computation can effectively improve the performance compared with using a single basic method. (iii) iPiDA-SWGCN outperforms all the other methods because it integrates all the basic methods to compute weights so as to assign weights to piRNA-disease associations with better performance.

Impact of different association adjacency matrices on the performance of iPiDA-SWGCN

Previous studies represent the piRNA-disease associations as Boolean values, where the element in the adjacency matrix is 1 if the association in the network is known, otherwise it is 0. In this study, we compute weights for unknown associations in the network instead of 0. To

Table 1. Performance comparison between iPiDA-SWGCN and predictors based on different basic methods and their combinations for weighing associations on D_{ind}^2 .

Methods ^a	AUC	AUPR
iPiDA-SWGCN (All)	0.8178	0.8151
R-G	0.8039	0.8026
G-S	0.8043	0.7985
R-S	0.8028	0.7851
G	0.7925	0.7978
R	0.7883	0.7914
S	0.7688	0.7397
None	0.6755	0.6654

^a The basic methods RF, SVM and GBDT are denoted as 'R', 'S' and 'G' respectively. 'R-S' indicates that we employ RF and SVM to weight the associations, and then feed the weighted network into GCN, 'G-S' indicates that GBDT and SVM are used to compute weights, and then GCN is employed. Other predictors are notated with the same rule. 'All' denotes that all the basic methods are adopted for weight computation, and 'None' denote no basic method is used, GCN is performed on the original piRNA-disease network.

<https://doi.org/10.1371/journal.pcbi.1011242.t001>

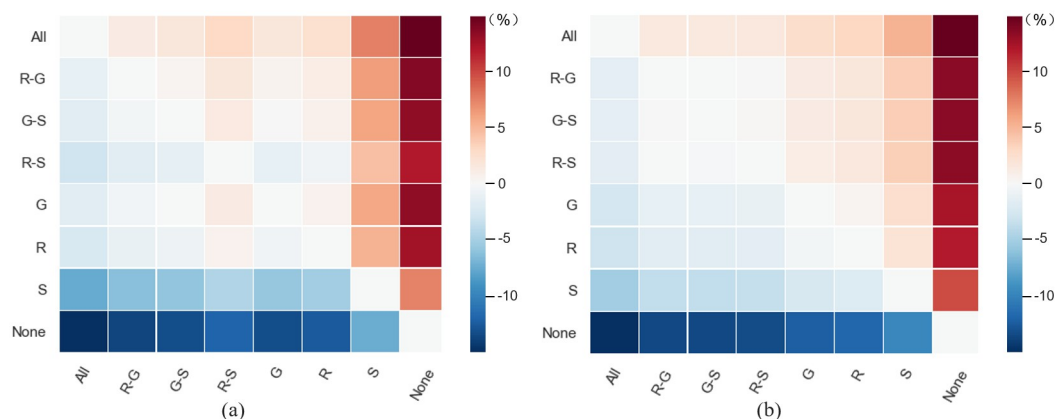


Fig 4. Performance comparison of different basic method combinations. (a) and (b) show the difference values in terms of AUC and AUPR among different predictors, respectively. The difference values are computed by the scores obtained from the method labelled at y-axis minus the scores obtained from the method labelled at x-axis.

<https://doi.org/10.1371/journal.pcbi.1011242.g004>

illuminate the effectiveness and necessity of assigning weights instead of Boolean values for unknown edges, we perform experiments on five networks with different edge types. The experimental results are listed in Table 2, from which we can see that the model with the weighted network is superior to those with the network with added Boolean piRNA-disease edges. The reason is that assigning weights to associations cannot only enrich more network structural information, but also can provide association confidence information for GCN so as to learn the expressive node representation. In contrast, Boolean edges only indicate whether the associations exist or not.

Performance comparison among different methods

To illuminate the effectiveness of iPiDA-SWGCN for identifying piRNA-disease associations, we conducted a performance comparison of our method with five state-of-the-art approaches, including iPiDi-PUL [19], iPiDA-sHN [20], iPiDA-LTR [24], iPiDA-GCN [25] and piRDA [57]. The web servers or source codes of all these methods are accessible, enabling unbiased evaluation of their performance. The experimental results are displayed in Table 3, from which

Table 2. The performance comparison of iPiDA-SWGCN based on different adjacency matrices on D_{ind}^2 .

Methods	AUC	AUPR
iPiDA-SWGCN_Bool10 ^a	0.7361	0.7629
iPiDA-SWGCN_Bool20 ^b	0.7357	0.7691
iPiDA-SWGCN_Bool30 ^c	0.7313	0.7569
iPiDA-SWGCN_Bool40 ^d	0.7219	0.7346
iPiDA-SWGCN	0.8178	0.8151

^{a,b,c,d} The unknown piRNA-disease pairs are ranked according to their prediction scores computed by basic predictors. The methods with suffix ‘_Bool10’, ‘_Bool20’, ‘_Bool30’ and ‘_Bool40’ denote that GCN is performed on each supplementarily weighted matrix with bool weights assigned, where the top 10%, 20%, 30% and 40% unknown associations with the highest predicted scores are computed by several base predictors, then these associations are viewed as positive samples and supplemented into the original piRNA-disease networks.

<https://doi.org/10.1371/journal.pcbi.1011242.t002>

Table 3. Performance comparison of different methods on D_{ind}^2 .

Methods ^a	AUC	AUPR
iPiDi-PUL ^b	0.6653	0.6550
iPiDA-sHN ^c	0.5226	0.5203
iPiDA-LTR ^d	0.6147	0.6641
iPiDA-GCN ^e	0.7149	0.7036
piRDA ^f	0.4939	0.5116
iPiDA-SWGCN^g	0.8178	0.8151

^aAll these predictors are trained with D_{ben}^2 . The results of iPiDi-PUL [19], iPiDA-sHN [20] and iPiDA-GCN are obtained from [25]. The result of iPiDA-LTR [24] is obtained by in-house implementation;

^bResults obtained with parameters $n_components = 200$, $n_estimators = 150$, $max_features = 0.2$ and number of ensemble learner = 5;

^cResults obtained with parameters $C = 1.0$, kernel = 'rbf' and gamma = 1;

^dResults obtained with parameters tree = 120, shrinkage = 0.22 and leaves = 3;

^eResults obtained with parameters epoch = 2000, learning rate = 0.001, weight decay factor = 1.0 and learn rate decay = 0.2;

^fThe results were generated using the web server of piRDA (<http://nscbio.jbnu.ac.kr/tools/piRDA/>). Since piRDA was developed based on the piRDisease dataset, it is limited to predict associations related to only 13 diseases in D_{ind}^2 . Therefore, we only evaluated the prediction results for these associations;

^gResults obtained with parameters epoch = 1000, learning rate = 0.001, weight decay factor = 1.0 and learn rate decay = 0.2.

<https://doi.org/10.1371/journal.pcbi.1011242.t003>

we can conclude that iPiDA-SWGCN outperforms all the other methods. iPiDA-SWGCN is superior to iPiDA-GCN by 10.29% and 11.15% in terms of AUC and AUPR, respectively. The performance improvement can be attributed to the fact that iPiDA-SWGCN is able to capture more expressive node representations by aggregating more neighbor information from the supplementarily weighted network.

Visualization of predicted associations by iPiDA-SWGCN

In order to visually illustrate the effectiveness of the supplementarily weighted network used in iPiDA-SWGCN, the prediction results of GCN performing on different networks are compared. We take two piRNA-disease associations for further analysis, including <piR-hsa-22710, Parkinson's disease> and <piR-hsa-28405, renal cell carcinoma>. Parkinson's Disease (PD) is a widely prevalent neurodegenerative disorder. The gross pathological findings reveal significant damage to dopaminergic neurons in the midbrain's substantia nigra (SN), leading to dopamine deficiency in the nerve terminals located in the basal ganglia [58]. PiR-hsa-22710 has length of 30nt and is downregulated in PD-patient tissue samples [59]. Renal cell carcinoma (RCC) is a prevalent cancer, ranking sixth in incidence in men and tenth in women globally [59]. Recent research suggests that the expression levels of piRNA are linked to the histological grade, pathological features of RCC, and patient survival. Multiple studies have demonstrated that piRNAs show differentially expressed in benign versus malignant renal tumor tissues [59,60]. PiR-hsa-28405 is a type of Homo sapiens piRNA with length of 32nt, showing about 4-fold downregulation in renal tumor tissue [60].

The results are shown in Fig 5, from which we can draw the following conclusions: (i) Due to the shortage of verified piRNA-disease associations, the ability of GCN to capture network structural information is limited; (ii) GCN performing on the supplementarily weighted

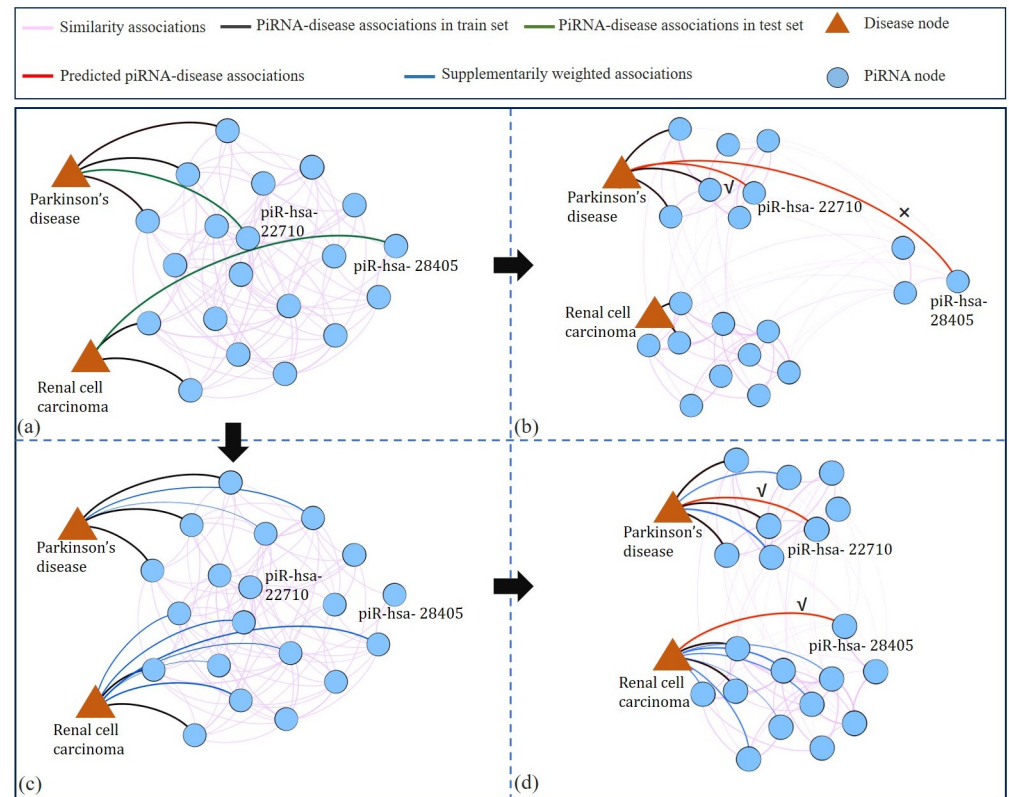


Fig 5. The prediction visualization of iPiDA-SWGCN. The figures are plotted by using Gephi [61]. (a) PiRNA-disease associations for training and test. The black lines denote the associations in the training set, and the green lines denote the associations in test set for prediction. (b) Predicted results of the GCN performed on the original piRNA-disease association network. The red lines show the prediction results on the original network, and Parkinson's disease is incorrectly predicted to be associated with piR-hsa-28405. (c) Supplementarily weighted piRNA-disease association network. The blue lines denote the supplementary weighted associations computed by several basic predictors. (d) Predicted results of the GCN performed on the supplementary weighted piRNA-disease association network. The associations in the test set are correctly predicted by GCN when adding supplementary weighted associations.

<https://doi.org/10.1371/journal.pcbi.1011242.g005>

network can correctly predict the test piRNA-disease associations, because the weighted network cannot only provide more proximity structural information, but also contains the association confidence.

Case study

To illuminate the practicability of iPiDA-SWGCN, we applied iPiDA-SWGCN to detect potential piRNAs associated with three important diseases ('Renal cell carcinoma', 'Parkinson's disease' and 'Cardiovascular disease'). The top five detected piRNAs and their corresponding literature evidence are listed in Table 4. From Table 4 we can see that all the top five detected piRNAs associated with 'Renal cell carcinoma' and 'Parkinson's disease' have been validated by literatures. Four of five detected piRNAs associated with 'cardiovascular disease' are confirmed by the literature. For example, piR-hsa-23184 shows a higher expression of 2.26-fold in metastatic compared to non-metastatic tumor [60]. The piR-hsa-5389 is upregulated in cells and post-mortem tissue samples between control and Parkinson's disease patients [59]. The expression of piR-hsa-25177 in cardio sphere cells is 3.38-fold higher than that in cardio sphere-derived cells [62]. Therefore, iPiDA-SWGCN can effectively detect potential

Table 4. The top 5 piRNAs associated with different diseases detected by iPiDA-SWGCN.

Disease	Rank	piRNA	Evidence ^a
Renal cell carcinoma	1	piR-hsa-753	PMID:26071182
	2	piR-hsa-22558	PMID:26071182
	3	piR-hsa-28427	PMID:26071182
	4	piR-hsa-23184	PMID:26071182
	5	piR-hsa-7714	PMID:26071182
Parkinson's disease	1	piR-hsa-5389	PMID:29986767
	2	piR-hsa-26242	PMID:29986767
	3	piR-hsa-24656	PMID:29986767
	4	piR-hsa-1271	PMID:29986767
	5	piR-hsa-14134	PMID:29986767
Cardiovascular disease	1	piR-hsa-25177	PMID:27131603
	2	piR-hsa-30122	PMID:27131603
	3	piR-hsa-3578	PMID:27131603
	4	piR-hsa-1180	PMID:27131603
	5	piR-hsa-19768	PMID:27131603

^aThe identified piRNA-disease associations are confirmed by the literature. The PMIDs for the evidence literature are given as well.

<https://doi.org/10.1371/journal.pcbi.1011242.t004>

piRNA-disease associations, which is suitable for real world applications. The more specific case study results are shown in Table B in [S1 Text](#).

Conclusion

In this work, we propose a novel predictor named iPiDA-SWGCN for piRNA-disease association prediction by combining the supplementarily weighted strategy and GCN. The iPiDA-SWGCN mainly has following advantages: (i) Potential piRNA-disease associations are supplemented in the piRNA-disease network by integrating various basic predictors to provide an informative network, based on which GCN can capture deep proximity structure, and fully utilize network information for feature learning. (ii) Different confidences are assigned to the piRNA-disease associations instead of Boolean values. Therefore, GCN aggregates node information and accurately learns node representations from neighbor nodes in varying degrees. (iii) Although iPiDA-SWGCN is proposed for predicting piRNA-disease associations, it can be extended to other link prediction tasks.

It is anticipated that the strategy of supplementarily weighted adjacency matrix will be applied to other related fields to solve the problems of limited positive samples, such as lncRNA-disease association detection and drug repositioning. In particular, there are plenty of unknown associations in link prediction field resulting in high sparsity problem. The supplementarily weighted strategy can be implemented to preliminarily enrich the association network so as to provide more useful information and improve the prediction performance.

Besides, the value of piRNA-disease associations without independent experimental validation is worth mentioning. PiRNAs are potential biomarkers that may provide new avenues of investigation into the pathogenesis of diseases. However, it should be noted that these associations predicted by iPiDA-SWGCN are only correlations and require experimental validation to confirm their significance. This is especially relevant for piRNA studies due to their high abundance and poorly understood roles in diseases.

Future studies should aim to validate these associations through a combination of biological experiments and bioinformatics analysis to ensure their reliability and accuracy, promoting to gain a better understanding of the complex mechanisms underlying diseases and develop more effective strategies for prevention and treatment.

Supporting information

S1 Text. Fig A. Parameter analysis of iPiDA-SWGCN. Table A. The performance comparison of basic predictors on D_{ind}^2 . Table B. The top 10 piRNAs associated with different diseases detected by iPiDA-SWGCN.
(DOCX)

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: 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 P, Dong Y, Gu J, Puthiyakunnon S, Wu Y, Chen X-G. Developmental piRNA profiles of the invasive vector mosquito *Aedes albopictus*. *Parasites & vectors*. 2016; 9(1):1–15. <https://doi.org/10.1186/s13071-016-1815-8> PMID: 27686069
2. Seto AG, Kingston RE, Lau NC. The coming of age for Piwi proteins. *Molecular cell*. 2007; 26(5):603–9. <https://doi.org/10.1016/j.molcel.2007.05.021> PMID: 17560367
3. Singh G, Roy J, Rout P, Mallick B. Genome-wide profiling of the PIWI-interacting RNA-mRNA regulatory networks in epithelial ovarian cancers. *PloS one*. 2018; 13(1):e0190485. <https://doi.org/10.1371/journal.pone.0190485> PMID: 29320577
4. Czech B, Hannon GJ. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends in biochemical sciences*. 2016; 41(4):324–37. <https://doi.org/10.1016/j.tibs.2015.12.008> PMID: 26810602
5. Han Y-N, Li Y, Xia S-Q, Zhang Y-Y, Zheng J-H, Li W. PIWI proteins and PIWI-interacting RNA: emerging roles in cancer. *Cellular Physiology and Biochemistry*. 2017; 44(1):1–20. <https://doi.org/10.1159/000484541> PMID: 29130960
6. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*. 2006; 313(5785):320–4. <https://doi.org/10.1126/science.1129333> PMID: 16809489

7. Chen C, Liu J, Xu G. Overexpression of PIWI proteins in human stage III epithelial ovarian cancer with lymph node metastasis. *Cancer Biomarkers*. 2013; 13(5):315–21. <https://doi.org/10.3233/CBM-130360> PMID: 24440970
8. 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
9. 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
10. Tan L, Mai D, Zhang B, Jiang X, Zhang J, Bai R, et al. PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. *Molecular cancer*. 2019; 18(1):1–15.
11. Cheng J, Guo J-M, Xiao B-X, Miao Y, Jiang Z, Zhou H, et al. piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clinica chimica acta*. 2011; 412(17–18):1621–5. <https://doi.org/10.1016/j.cca.2011.05.015> PMID: 21616063
12. 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.
13. Ray SK, Mukherjee S. Piwi-interacting RNAs (piRNAs) and Colorectal Carcinoma: Emerging Non-invasive diagnostic Biomarkers with Potential Therapeutic Target Based Clinical Implications. *Current Molecular Medicine*. 2022.
14. Zhang W, Li Z, Guo W, Yang W, Huang F. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM transactions on computational biology and bioinformatics*. 2019; 18(2):405–15.
15. Chen Q, Zhe Z, Lan W, Zhang R, Wang Z, Luo C, et al. Identifying miRNA-disease association based on integrating miRNA topological similarity and functional similarity. *Quantitative Biology*. 2019; 7:202–9.
16. Ha J, Park C, Park C, Park S. Improved prediction of miRNA-disease associations based on matrix completion with network regularization. *Cells*. 2020; 9(4):881. <https://doi.org/10.3390/cells9040881> PMID: 32260218
17. Yao D, Zhan X, Zhan X, Kwok CK, Li P, Wang J. A random forest based computational model for predicting novel lncRNA-disease associations. *BMC bioinformatics*. 2020; 21:1–18.
18. Ji C, Gao Z, Ma X, Wu Q, Ni J, Zheng C. AEMDA: inferring miRNA–disease associations based on deep autoencoder. *Bioinformatics*. 2021; 37(1):66–72. <https://doi.org/10.1093/bioinformatics/btaa670> PMID: 32726399
19. 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
20. 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
21. Zheng K, You Z-H, Wang L, Li H-Y, Ji B-Y, editors. Predicting human disease-associated piRNAs based on multi-source information and random forest. *International Conference on Intelligent Computing*; 2020: Springer.
22. 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.
23. 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.
24. Zhang W, Hou J, Liu B. iPiDA-LTR: Identifying piwi-interacting RNA-disease associations based on Learning to Rank. *PLOS Computational Biology*. 2022; 18(8):e1010404. <https://doi.org/10.1371/journal.pcbi.1010404> PMID: 35969645
25. Hou J, Wei H, Liu B. iPiDA-GCN: Identification of piRNA-disease associations based on Graph Convolutional Network. *PLOS Computational Biology*. 2022; 18(10):e1010671. <https://doi.org/10.1371/journal.pcbi.1010671> PMID: 36301998
26. Muhammad A, Waheed R, Khan NA, Jiang H, Song X. piRDisease v1. 0: a manually curated database for piRNA associated diseases. *Database*. 2019; 2019.

27. 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
28. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907*. 2016.
29. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*. 2016;29.
30. Wang L, You Z-H, Huang Y-A, Huang D-S, Chan KC. An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network. *Bioinformatics*. 2020; 36(13):4038–46. <https://doi.org/10.1093/bioinformatics/btz825> PMID: 31793982
31. 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
32. 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
33. 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](https://doi.org/10.1016/0022-2836(81)90087-5) PMID: 7265238
34. Pearson WR. An introduction to sequence similarity (“homology”) searching. *Current protocols in bioinformatics*. 2013; 42(1):3.1–3.1.8. <https://doi.org/10.1002/0471250953.bi0301s42> PMID: 23749753
35. Petti S, Bhattacharya N, Rao R, Dauparas J, Thomas N, Zhou J, et al. End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman. *Bioinformatics*. 2023; 39(1):btac724. <https://doi.org/10.1093/bioinformatics/btac724> PMID: 36355460
36. 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
37. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, et al. A network based method for analysis of lncRNA–disease associations and prediction of lncRNAs implicated in diseases. *PloS one*. 2014; 9(1):e87797. <https://doi.org/10.1371/journal.pone.0087797> PMID: 24498199
38. Wang L, You Z-H, Li Y-M, Zheng K, Huang Y-A. GCNCDA: a new method for predicting circRNA–disease associations based on graph convolutional network algorithm. *PLOS Computational Biology*. 2020; 16(5):e1007568. <https://doi.org/10.1371/journal.pcbi.1007568> PMID: 32433655
39. Chen X, Wang L, Qu J, Guan N-N, Li J-Q. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics*. 2018; 34(24):4256–65. <https://doi.org/10.1093/bioinformatics/bty503> PMID: 29939227
40. 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
41. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics*. 2007; 23(10):1274–81. <https://doi.org/10.1093/bioinformatics/btm087> PMID: 17344234
42. Kibbe WA, Arze C, Felix V, Mittra 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. <https://doi.org/10.1093/nar/gku1011> PMID: 25348409
43. 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
44. 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
45. 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.
46. Breiman L. Random forests. *Machine learning*. 2001; 45(1):5–32.
47. Noble WS. Support vector machine applications in computational biology. *Kernel methods in computational biology*. 2004; 14:71–92.
48. Friedman JH. Stochastic gradient boosting. *Computational statistics & data analysis*. 2002; 38(4):367–78.

49. 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
50. Rong Y, Huang W, Xu T, Huang J. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*. 2019.
51. Ioffe S, Szegedy C, editors. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*; 2015: PMLR.
52. 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.
53. Flach P, Kull M. Precision-recall-gain curves: PR analysis done right. *Advances in neural information processing systems*. 2015;28.
54. 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
55. Hasanin T, Khoshgoftaar TM, Leevy JL, editors. A comparison of performance metrics with severely imbalanced network security big data. 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI); 2019: IEEE.
56. Yi H-C, You Z-H, Huang D-S, Guo Z-H, Chan KC, Li Y. Learning representations to predict intermolecular interactions on large-scale heterogeneous molecular association network. *Iscience*. 2020; 23(7):101261. <https://doi.org/10.1016/j.isci.2020.101261> PMID: 32580123
57. 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
58. Kakkar AK, Dahiya N. Management of Parkinson's disease: Current and future pharmacotherapy. *European journal of pharmacology*. 2015; 750:74–81.
59. 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.
60. Busch J, Ralla B, Jung M, Wotschovsky 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
61. 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.
62. 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