PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# iPiDA-LTR: Identifying piwi-interacting RNA-disease associations based on Learning to Rank

**Wenxiang Zhang**[1], **Jialu Hou**[1], **Bin Liu**[1,2]*

**1** School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, **2** Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing, China

* bliu@bliulab.net



OPEN ACCESS

**Citation:** Zhang W, Hou J, Liu B (2022) iPiDA-LTR: Identifying piwi-interacting RNA-disease associations based on Learning to Rank. PLoS Comput Biol 18(8): e1010404. https://doi.org/10.1371/journal.pcbi.1010404

**Editor:** Quan Zou, University of Electronic Science and Technology, CHINA

**Received:** March 27, 2022

**Accepted:** July 18, 2022

**Published:** August 15, 2022



**Data Availability Statement:** A user-friendly web server of iPiDA-LTR predictor is freely available at http://bliulab.net/iPiDA-LTR/.

**Funding:** This work was supported by the National Key R&D Program of China (No. 2018AAA0100100 to BL) and the Beijing Natural Science Foundation (No. JQ19019 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

Piwi-interacting RNAs (piRNAs) are regarded as drug targets and biomarkers for the diagnosis and therapy of diseases. However, biological experiments cost substantial time and resources, and the existing computational methods only focus on identifying missing associations between known piRNAs and diseases. With the fast development of biological experiments, more and more piRNAs are detected. Therefore, the identification of piRNA-disease associations of newly detected piRNAs has significant theoretical value and practical significance on pathogenesis of diseases. In this study, the iPiDA-LTR predictor is proposed to identify associations between piRNAs and diseases based on Learning to Rank. The iPiDA-LTR predictor not only identifies the missing associations between known piRNAs and diseases, but also detects diseases associated with newly detected piRNAs. Experimental results demonstrate that iPiDA-LTR effectively predicts piRNA-disease associations outperforming the other related methods.

## Author summary

Accumulating evidences have indicated that dysfunction and abnormal expression of piRNAs are closely associated with the emergence and development of diseases. Currently, identifying piRNA-disease associations mainly focuses on biological experimental methods and computational methods. However, biological experimental methods take substantial time and resources. Computational methods mainly focused on identifying diseases associated known piRNAs. With the development of biological technology, more and more newly detected piRNAs were detected. Therefore, identifying diseases associated with newly detected piRNAs is more important compared with identifying diseases associated with known piRNAs. Information retrieval (IR)'s goal is to rank documents based on the relevance to certain topics. This task is particularly similar with identification of piRNA-disease associations. Specifically, ranking documents related to previous topics corresponds to identify diseases associated with known piRNAs, and ranking documents related to novel topics is similar to identify diseases associated with newly detected piRNAs. Therefore, we propose a new predictor called iPiDA-LTR to predict associations

between piRNAs and diseases based on information retrieval technology. Experimental results indicated that iPiDA-LTR is promising in identifying diseases associated with known piRNAs and newly detected piRNAs.

This is a *PLOS Computational Biology* Methods paper.

## Introduction

Piwi-interacting RNA (piRNA) with 24–31 nucleotides in length is a class of small RNAs interacting with Piwi-subfamily Argonaute proteins [1–3]. Early studies found that piRNAs mainly located in germ stem cells on drosophila and mouse, and regulated germ stem cell proliferation [4–6]. With the fast development of biotechnology and computing techniques [7, 8], more and more piRNAs were discovered, and the corresponding functions were also detected, including stem cell proliferation, gene expression, and heterochromatin formation, etc [9–12].

As more and more piRNA functions were detected, many evidences indicated that dysfunction and abnormal expression of piRNAs are closely associated with the emergence and development of diseases [13–17]. Therefore, the identification of associations between piRNAs and diseases is important for diagnosis and treatment of diseases [18, 19]. Currently, it mainly focused on biological experimental methods and computational methods. For biological experiments methods, Cabral et al. indicated that piRNAs play a role in the process of translational research of gastric cancer as potential biomarkers [20]. Krishnan et al. identified eight non-redundant piRNAs as breast cancer markers [21]. Roy et al. studied the reciprocal expression between piRNAs and the corresponding targets, and provided a novel insight into the role of piRNAs in Alzheimer's disease [22]. Although biological experimental methods are highly reliable, it takes substantial time and resources. Some computational methods have been proposed for identifying the associations between non-coding RNAs and diseases, such as miRNA-disease associations [23], circRNA-disease associations [24], etc. In this regard, computational methods are proposed to predict piRNA-disease associations, which can serve as powerful auxiliary tools to save time and cost compared with biological experiments. For example, Wei et al. proposed the first computational predictor for identifying piRNA-disease associations based on the positive unlabelled learning algorithm, and established the first web server [25]. A convolutional neural network was utilized to extract association features between piRNAs and diseases, and then the Support Vector Machine was employed to construct the predictor [26]. Although computational methods have been proposed, they mainly aim at the application scenario of identifying missing associations between known piRNAs and diseases. However, more and more newly detected piRNAs were detected [27–29]. Therefore, the application scenario of identifying piRNA-disease associations of newly detected piRNAs is very important to investigate piRNA functions and disease pathogenesis.

In recent years, information retrieval (IR) becomes a widely used technology, whose ultimate goal is to rank documents based on the relevance to certain topics [30, 31]. As an successful algorithm in information retrieval, Learning to Rank (LTR) [32, 33] has been successfully applied to web page retrieval employed by Google [34], Yahoo [35], Microsoft [36], etc. Compared with traditional IR methods, the advantage of LTR is that it integrates component methods so as to automatically rank documents associated with query from

multiple perspectives [30]. LTR has been applied in identifying circRNA-disease associations [24], detecting protein remote homology [37], predicting protein-phenotype associations [38], drug–target binding affinity prediction [39], etc. The core concept of LTR is to calculate the relevance score $f(q, d)$ between query $q$ and document $d$. Therefore, this task is particularly similar with identification of piRNA-disease associations (see **Fig 1**). PiRNAs and diseases can be treated as queries and documents, respectively. Learning to Rank not only identifies associations between known piRNAs and diseases, but also ranks diseases associated with newly detected piRNAs.

In this study, we propose a new predictor, named iPiDA-LTR, to predict associations between piRNAs and diseases, which has the following advantages. iPiDA-LTR predictor combines component methods and Learning to Rank, which cannot only identify missing associations between known piRNAs and diseases, but also can identify diseases associated with newly detected piRNAs. Experimental results indicated that iPiDA-LTR is promising to identify piRNA-disease associations. A web server of iPiDA-LTR is constructed to identify diseases associated with query piRNAs, which can be accessed at http://bliulab.net/iPiDA-LTR.

## Materials and methods

### Materials

To imitate two application scenarios, we construct two types of datasets based on piRDisease v1.0 database [40] collecting 7939 piRNA-disease associations with 4796 piRNAs and 28 diseases. Firstly, a standard dataset $\mathbb{S}_{all}$ is constructed following [25], which can be represented as:

$$\begin{cases} \mathbb{S}_{all} = \mathbb{A}_{all} \cup \mathbb{P}_{all} \cup \mathbb{D} \\ \mathbb{A}_{all} = \mathbb{A}^{+}_{all} \cup \mathbb{A}^{-}_{all} \end{cases} \tag{1}$$

where $\mathbb{A}_{all}$ represents 5002 piRNA-disease associations from [25]. $\mathbb{P}_{all}$ and $\mathbb{D}$ contain 4350 piRNAs and 21 diseases from $\mathbb{A}_{all}$, respectively. $\mathbb{A}^{+}_{all}$ and $\mathbb{A}^{-}_{all}$ contain known piRNA-disease associations and unknown piRNA-disease associations, respectively. Specifically, piRNA-disease associations contained in $\mathbb{A}^{+}_{all}$ are labelled as 1, otherwise 0. To avoid overfitting problem, $\mathbb{S}_{all}$ is further divided into a benchmark dataset and an independent dataset. The benchmark dataset is used to adjust parameters and train model via cross-validation, and the independent dataset is employed to evaluate the performance of different methods.

### For the first application scenario: predicting associations between known piRNAs and known diseases

Benchmark dataset and independent dataset are constructed as:

$$\begin{cases} \mathbb{S}^{a}_{ben} = \mathbb{S}^{a+}_{ben} \cup \mathbb{S}^{a-}_{ben} \\ \mathbb{S}^{a}_{ind} = \mathbb{S}^{a+}_{ind} \cup \mathbb{S}^{a-}_{ind} \\ \mathbb{S}^{+}_{all} = \mathbb{S}^{a+}_{ben} \cup \mathbb{S}^{a+}_{ind} \\ \mathbb{S}^{-}_{all} = \mathbb{S}^{a-}_{ben} \cup \mathbb{S}^{a-}_{ind} \end{cases} \tag{2}$$

where we randomly select 20% associations from $\mathbb{A}^{+}_{all}$ and $\mathbb{A}^{-}_{all}$ to construct $\mathbb{S}^{a+}_{ind}$ and $\mathbb{S}^{a-}_{ind}$, respectively, and then the remaining associations in $\mathbb{A}^{+}_{all}$ and $\mathbb{A}^{-}_{all}$ are used to construct $\mathbb{S}^{a+}_{ben}$ and $\mathbb{S}^{a-}_{ben}$, respectively. Obviously, $\mathbb{S}^{a}_{ben}$ represents benchmark dataset, which is used to optimize parameters and train models, and then trained models are used to identify unknown associations in $\mathbb{S}^{a}_{ind}$.

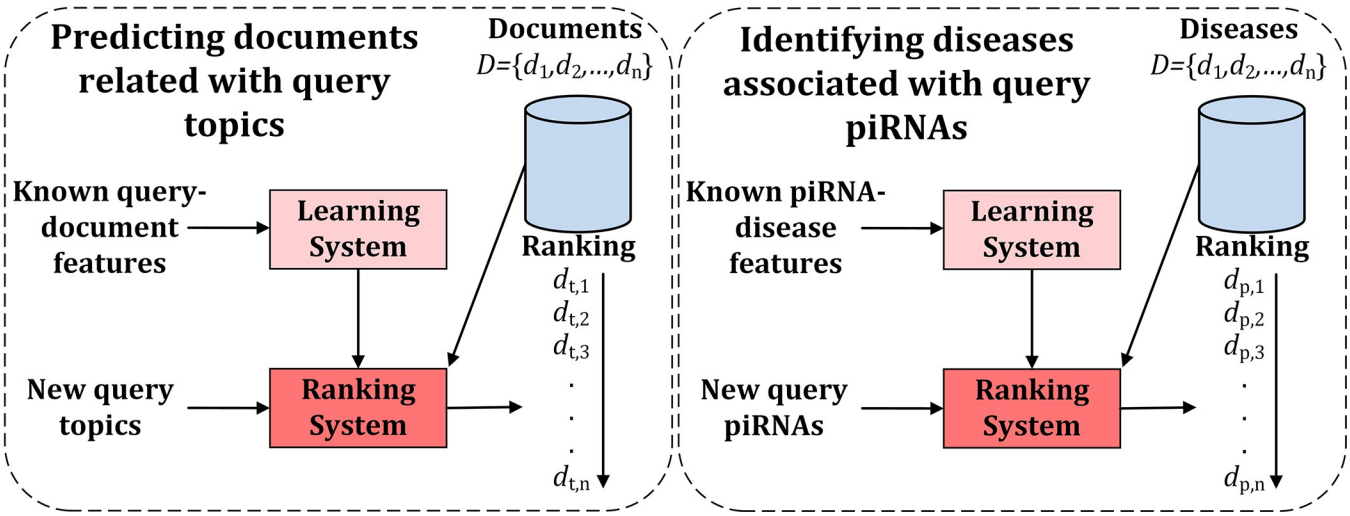


**Fig 1. The similarities between the prediction of documents related with query topics and the identification of diseases associated with query piRNAs, where piRNA and disease can be treated as query and document, respectively.**

https://doi.org/10.1371/journal.pcbi.1010404.g001

## For the second application scenario: predicting the associations between newly detected piRNAs and known diseases

To imitate the second application scenario, we randomly select 80% and 20% piRNAs from $\mathbb{P}_{all}$ as known piRNA set $\mathbb{P}_{all}^{knwon}$ and newly detected piRNA set $\mathbb{P}_{all}^{unknown}$, respectively, based on which benchmark dataset and independent dataset are constructed as:

$$\begin{cases} \mathbb{S}_{ben}^{p} = \{\mathbb{S}_{ben}^{p+} \cup \mathbb{S}_{ben}^{p-} | \text{piRNAs} \in \mathbb{P}_{all}^{knwon}\} \\ \mathbb{S}_{ind}^{p} = \{\mathbb{S}_{ind}^{p+} \cup \mathbb{S}_{ind}^{p-} | \text{piRNAs} \in \mathbb{P}_{all}^{unknown}\} \\ \mathbb{S}_{all}^{+} = \mathbb{S}_{ben}^{p+} \cup \mathbb{S}_{ind}^{p+} \\ \mathbb{S}_{all}^{-} = \mathbb{S}_{ben}^{p-} \cup \mathbb{S}_{ind}^{p-} \end{cases} \quad (3)$$

where $\mathbb{S}_{ben}^{p}$ and $\mathbb{S}_{ind}^{p}$ represent benchmark dataset and independent dataset, respectively. PiRNAs contained in $\mathbb{S}_{ben}^{p}$ and $\mathbb{S}_{ind}^{p}$ belong to $\mathbb{P}_{all}^{known}$ and $\mathbb{P}_{all}^{unknown}$, respectively. Detailed information of $\mathbb{S}_{ben}^{a}$, $\mathbb{S}_{ind}^{a}$, $\mathbb{S}_{ben}^{p}$ and $\mathbb{S}_{ind}^{p}$ is shown in **Table 1**. The datasets can be obtained at http://bliulab.net/iPiDA-LTR/dataset/.

## Method overview

In this study, a novel ranking framework, named iPiDA-LTR, is proposed to solve two application scenarios. The workflow of iPiDA-LTR is shown in **Fig 2** with three steps: (a) Association

**Table 1. The detailed statistical information of $\mathbb{S}_{ben}^{a}$, $\mathbb{S}_{ind}^{a}$, $\mathbb{S}_{ben}^{p}$ and $\mathbb{S}_{ind}^{p}$.**

| Datasets | PiRNAs | Diseases | Known* | Unknown# |
|---|---|---|---|---|
| $\mathbb{S}_{ben}^{a}$ | 4350 | 21 | 4002 | 69079 |
| $\mathbb{S}_{ind}^{a}$ | 4311 | 21 | 1000 | 17269 |
| $\mathbb{S}_{ben}^{p}$ | 3480 | 21 | 3999 | 69081 |
| $\mathbb{S}_{ind}^{p}$ | 870 | 21 | 1003 | 17267 |

* The associations between piRNAs and diseases have been validated by experiments

# The associations between piRNAs and diseases without experiment validations.

https://doi.org/10.1371/journal.pcbi.1010404.t001

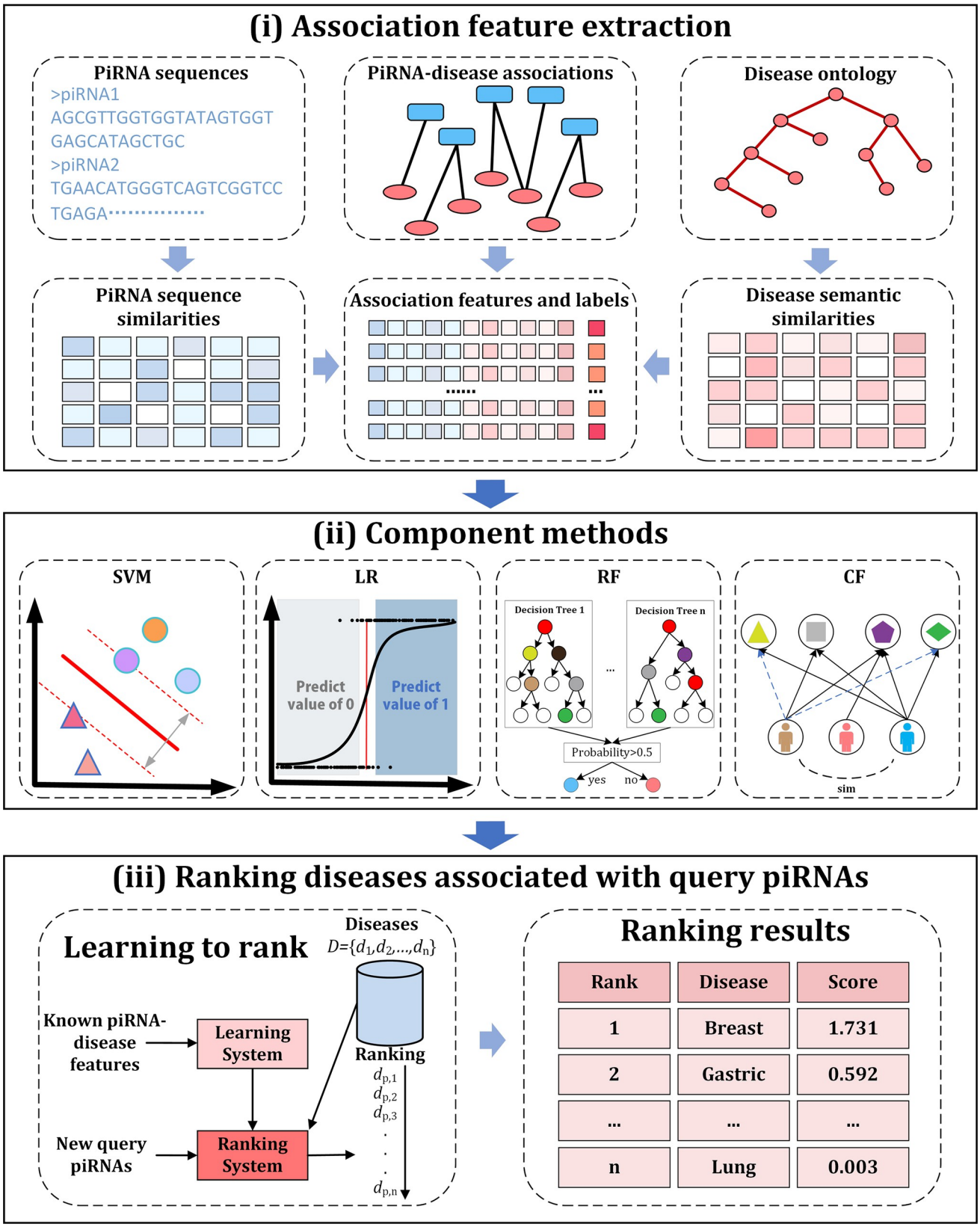


**Fig 2. The workflow of iPiDA-LTR predictor.** (i) Association feature extraction: piRNA sequences and disease ontology are used to calculate piRNA sequence similarities and disease semantic similarities by combining them and piRNA-disease associations to construct association features and labels; (ii) Component methods: four methods are used to train models with benchmark dataset, and then trained models are utilized to calculate association scores

of query piRNAs; (iii) Ranking diseases associated with query piRNAs: association scores of samples in the benchmark dataset are used to train LambdaMART model, and then trained LambdaMART model is employed to rank diseases associated with query piRNAs.

https://doi.org/10.1371/journal.pcbi.1010404.g002

feature extraction; (b) Component methods; (c) Ranking diseases associated with query piRNAs.

## Association feature extraction

**PiRNA sequence similarities.** The piRNA similarities play a vital role in RNA-disease association identification [24–26], and piRNA sequence similarities have been applied to piRNA-disease association identification [25, 26]. Many methods have been proposed to calculate sequence similarities [41–43]. For example, Smith-Waterman algorithm has been successfully applied to multiple sequence analysis tasks, including RNA sequence similarity analysis [25, 26, 44], protein sequence analysis [45, 46], etc. In this study, we employ Smith-Waterman algorithm [41, 44] to calculate piRNA sequence similarities:

$$\mathbf{S_P}\left(p_i, p_j\right) = \frac{\mathrm{SW}(p_i, p_j)}{\sqrt{\mathrm{SW}(p_i, p_i) \times \mathrm{SW}(p_j, p_j)}} \tag{4}$$

where $\mathbf{S_P}(p_i, p_j)$ is similarity between piRNA $p_i$ and piRNA $p_j$. $\mathrm{SW}(p_i, p_j)$ represents local alignment score between piRNA $p_i$ and piRNA $p_j$ based on Smith-Waterman algorithm.

**Disease semantic similarities.** The disease semantic similarity calculation is a key component in RNA-disease association identification. The disease ontology [47] has been applied to RNA-disease association identification so as to calculate disease semantic similarities [48–53]. Disease ontology organized by the directed acyclic graph (DAG) provides a hierarchical structure of the complex disease parent node [47]. Similar diseases share similar hierarchical structure in DAG of disease ontology. Therefore, DAG of disease ontology helps to measure similarity between two diseases. In this study, we use DAG of disease ontology to calculate disease semantic similarities [54, 55]:

$$\mathbf{S_D}(m, n) = \frac{\sum_{i \in \mathrm{T}_m \cap \mathrm{T}_n} (\mathrm{S}_m(i) + \mathrm{S}_n(i))}{\sum_{j \in \mathrm{T}_m} \mathrm{S}_m(j) + \sum_{j \in \mathrm{T}_n} \mathrm{S}_n(j)} \tag{5}$$

$$\begin{cases} \mathrm{S}_n(i) = max\{0.5 * \mathrm{S}_n(j) | j \in children\ of\ i\} & \text{if } i \neq n \\ \mathrm{S}_n(i) = 1 & \text{if } i = n \end{cases} \tag{6}$$

where $\mathbf{S_D}(m, n)$ is similarity between disease $m$ and disease $n$. $\mathrm{T}_k$ represents the node set containing the ancestor nodes of $k$ and itself. $\mathrm{S}_n(i)$ is the semantic value of node $i$ to node $n$.

**Association features and labels.** The association feature between disease $d$ and the query piRNA $p$ is:

$$\mathbb{F}(p, d) = \{\mathbf{S_P}(p, :), \mathbf{S_D}(d, :)\} \tag{7}$$

where $\mathbb{F}(p, d)$ is the association features of piRNA $p$ and disease $d$. $\mathbf{S_P}(p,:)$ and $\mathbf{S_D}(d,:)$ represent $p$th row and $d$th row in the $\mathbf{S_P}$ and $\mathbf{S_D}$, respectively. If piRNA $p$ is associated with disease $d$, the label of $\mathbb{F}(p, d)$ is equal to 1, otherwise 0.

## Component methods

In this study, we select two types of component methods to calculate association scores, including machine learning methods and collaborative filtering (CF). For machine learning methods,

Random Forest (RF) [56–60], Logistic Regression method (LR) [61], and Support Vector Machine (SVM) [62–64] are employed, treating the identification of piRNA-disease association as a classification problem. CF is a recommendation algorithm [65, 66], which utilizes guilt-by-association assumption to identify piRNA-disease association focusing on local information. In this study, association features of benchmark dataset (see **Eq 7**) are used to train machine learning models, and then used to calculate association scores for $\mathbb{S}_{all}$ dataset. Finally, association features between piRNA $p$ and disease $d$ can be represented as:

$$\mathbb{Q}(p,d) = \{\mathbb{V}_{CF}(p,d), \mathbb{V}_{LR}(p,d), \mathbb{V}_{RF}(p,d), \mathbb{V}_{SVM}(p,d)\} \tag{8}$$

where $\mathbb{Q}(p,d)$ represents association features of piRNA $p$ and disease $d$. $\mathbb{V}_{CF}(p,d)$, $\mathbb{V}_{LR}(p,d)$, $\mathbb{V}_{RF}(p,d)$ and $\mathbb{V}_{SVM}(p,d)$ are association scores between piRNA $p$ and disease $d$ calculated by CF, LR, RF and SVM, respectively.

### Ranking diseases associated with query piRNAs

In this study, we employ Learning to Rank (LTR) to solve the problem of identifying potential piRNA-disease associations motivated by information retrieval [24, 37, 38, 67]. LTR is generally classified into three categories, including ListWise, PairWise and PointWise [68]. In this study, a ListWise method LambdaMART [32] is selected to obtain high quality of top-ranked diseases, which has been applied in identifying circRNA-disease associations [24], detecting protein remote homology [37], predicting protein-phenotype associations [38] and drug–target binding affinity prediction [39]. The number of trees, the truncation level $k$, shrinkage and the number of leaves are the four main parameters. The truncation level of $k$ influences the quality of top-ranked results by Normalized Discounted Cumulative Gain (NDCG), which can be formulated as [32]:

$$\begin{cases} \mathrm{DCG@}k = \sum_{i=1}^{k} \dfrac{2^{rel_i} - 1}{\log_2(i+1)} \\ \mathrm{NDCG@}k = \dfrac{\mathrm{DCG@}k}{\mathrm{IDCG@}k} \end{cases} \tag{9}$$

where $k$ represents the truncation level. IDCG@$k$ is the value of DCG@$k$ in the best optimal ranking results. If a query piRNA is associated with disease located in position $i$, $rel_i$ is equal to 1, otherwise 0. To obtain the final ranking results, association scores calculated by **Eq 8** for training set are used to train LambdaMART model, and the trained LambdaMART model is employed to rank diseases associated with query piRNAs based on association scores of query piRNAs.

## Results and discussion

### Evaluation criteria

In this study, the benchmark dataset is employed to optimize the parameters of the models, and the independent dataset is used to evaluate the performance of predictors. How to evaluate the ranking quality and prediction performance is crucial for identifying piRNA-disease associations. Because iPiDA-LTR predictor treats the identification of piRNA-disease associations as an information retrieval ranking task, we employ three important ranking criteria to evaluate the rank quality of different predictors: Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP) and ROC$k$. Besides, Area Under the ROC Curve (AUC) and Area Under the Precision-Recall Curve (AUPR) are also used to measure

comprehensive performance [69–72]. The average values of these criteria for all query piRNAs are calculated to evaluate performance of predictors.

## The effect of parameters for identifying piRNA-disease associations

iPiDA-LTR predictor mainly contains the following four parameters: the number of trees, the truncation level *k*, shrinkage and the number of leaves. Due to the large number of combinations of the four parameters, we fix three parameters in turns, and then find the local optimal values of the remaining parameters according to AUPR. The influences of different combinations of parameters for iPiDA-LTR on $\mathbb{S}^{a}_{ben}$ dataset and $\mathbb{S}^{p}_{ben}$ dataset are shown in **Figs** 3 and 4, respectively, from which we can see that the final optimized combinations of four parameters on iPiDA-LTR predictor on $\mathbb{S}^{a}_{ben}$ dataset and $\mathbb{S}^{p}_{ben}$ dataset are (120, 14, 0.22, 3) and (30, 15, 0.10, 29), respectively.

## Complementary analysis for component methods

In this study, iPiDA-LTR incorporates two types of component methods, including machine learning methods (LR, RF and SVM) and collaborative filtering (CF). LR, RF and SVM are obtained by python package Scikit-learn [73]. For LR's parameters, max_iter and solver are assigned as 300 and liblinear, respectively. For RF's parameters, n_estimators, max_leaf_nodes, n_jobs and max_features are assigned as 80, 10, -1 and 0.2, respectively. For SVM's parameters, kernel and probability are assigned as linear and True, respectively. We analyze the impact of different types of component methods to identify associations between piRNAs and diseases, and the results are shown in **Tables** 2 and 3, from which we can see the followings: (i) iPiDA-LTR predictor outperforms iPiDA-LTR-ML predictor on $\mathbb{S}^{a}_{ben}$ dataset and $\mathbb{S}^{p}_{ben}$ dataset; (ii) The iPiDA-LTR obviously outperforms iPiDA-LTR-ML in terms of ranking criteria (NDCG@5 and ROC1), especially for the second application scenario (see **Table 3**). Machine learning methods based on classification algorithms focus on global predictive performance, and collaborative filtering can identify special piRNA-related diseases focusing on local predictive performance. Therefore, machine learning methods and collaborative filtering are complementary. It is not surprising that iPiDA-LTR predictor obtains the best performance compared with iPiDA-LTR-ML, because iPiDA-LTR shares the advantages of these two types of methods.

The usage frequencies of component methods measure the contribution of component methods for iPiDA-LTR. **Fig 5** shows the usage frequencies of component methods on iPiDA-LTR, from which we can see that each component method is frequently used, indicating that they are important for iPiDA-LTR. **Tables** 2 and 3 and **Fig 5** show that component methods are complementary, and iPiDA-LTR combines them leading to better performance for identifying piRNA-disease associations.

## Comparison with related methods

In this section, the two state-of-the-art predictors including iPiDi-PUL predictor [25] and iPiDA-sHN predictor [26] are compared with iPiDA-LTR predictor, and the results are shown in **Tables** 4 and 5, from which we can see that iPiDA-LTR is better than the other methods, indicating that iPiDA-LTR is more suitable for identifying piRNA-disease associations. Researchers tend to focus on the top ranked predicted associations in practical application scenarios. Therefore, we analyze the quality of the predicted results (see **Fig 6**), from which we can see that iPiDA-LTR outperforms the other predictors in terms of ROC1-ROC10. It is not surprising because the loss function of LambdaMART NDCG mainly focuses on the top-ranked predictive known associations (see **Eq 9**).

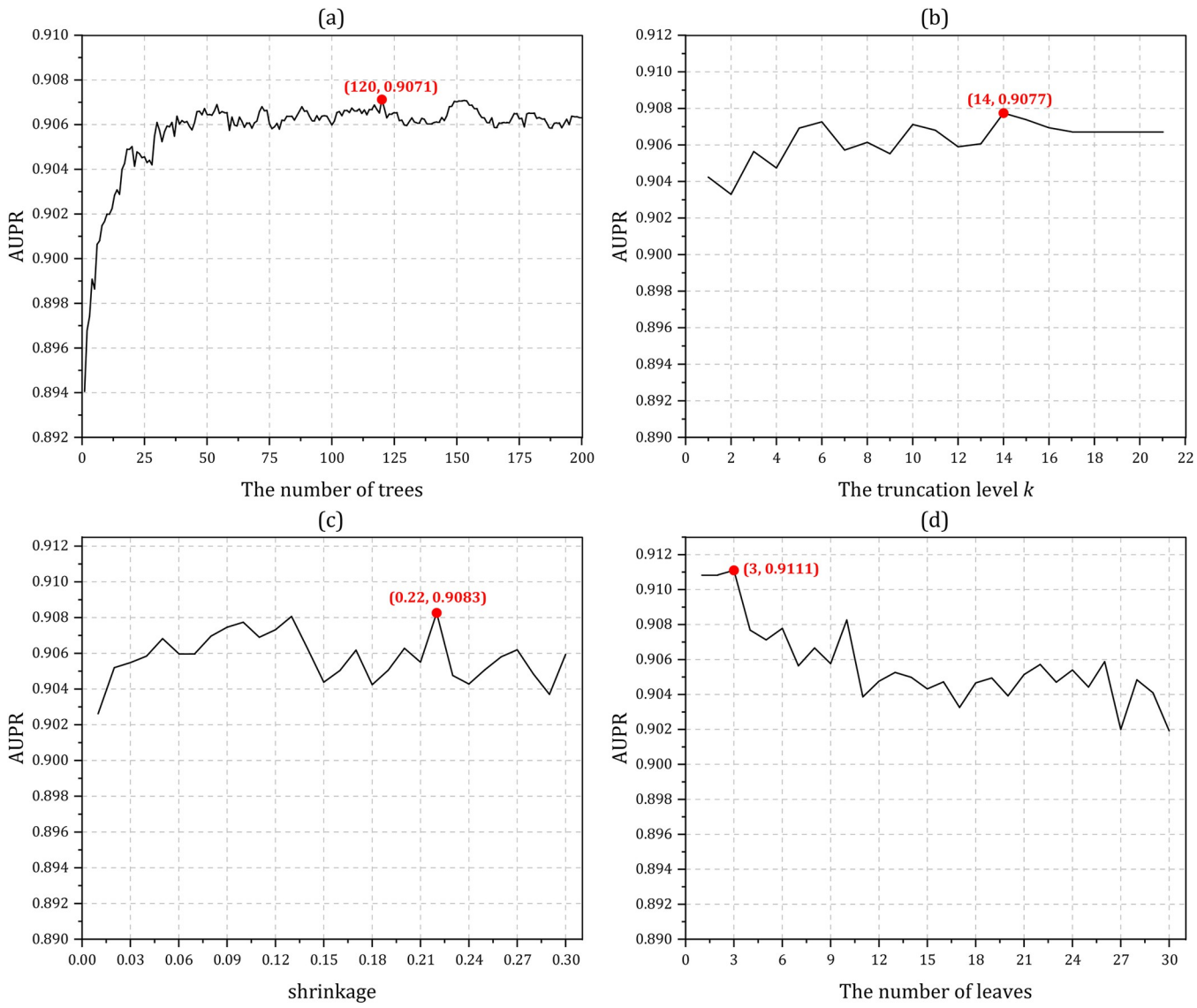


**Fig 3. The predictive results of iPiDA-LTR predictor on $\mathbb{S}^{a}_{ben}$ dataset via five-fold cross-validation.** (a) The truncation level *k*, shrinkage and the number of leaves are assigned as 10, 0.10, and 10 respectively, which are RankLib's default values (https://sourceforge.net/p/lemur/wiki/RankLib/), and the optimal value of the number of trees is 120; (b) The number of trees, shrinkage and the number of leaves are fixed as 120, 0.10 and 10 respectively, and the truncation level *k* is optimized as 14; (c) The number of trees, the truncation level *k* and the number of leaves are fixed as 120, 14 and 10 respectively, and the shrinkage is optimized as 0.22; (d) The number of trees, the truncation level *k* and shrinkage are 120, 14 and 0.22 respectively, and the number of leaves is set as 3.

https://doi.org/10.1371/journal.pcbi.1010404.g003

## Case study

To illustrate the predictive performance of iPiDA-LTR predictor for the identification of associations between new piRNAs and diseases, two query piRNAs, including piR-hsa-23210 and piR-hsa-15023, are selected as query piRNAs from $\mathbb{S}_{all}$ dataset, respectively. The remaining piRNAs in $\mathbb{S}_{all}$ are used to train iPiDA-LTR model, and then the trained iPiDA-LTR model is employed to predict diseases associated with piR-hsa-15023 and piR-hsa-23210.

The predicted results of piR-hsa-23210 and piR-hsa-15023 are shown in **Tables 6** and **7**, respectively, from which we can see the followings: (i) The evidences for the top five predicted piR-hsa-23210-associated diseases are supported by PubMed (https://pubmed.ncbi.nlm.nih.gov/).

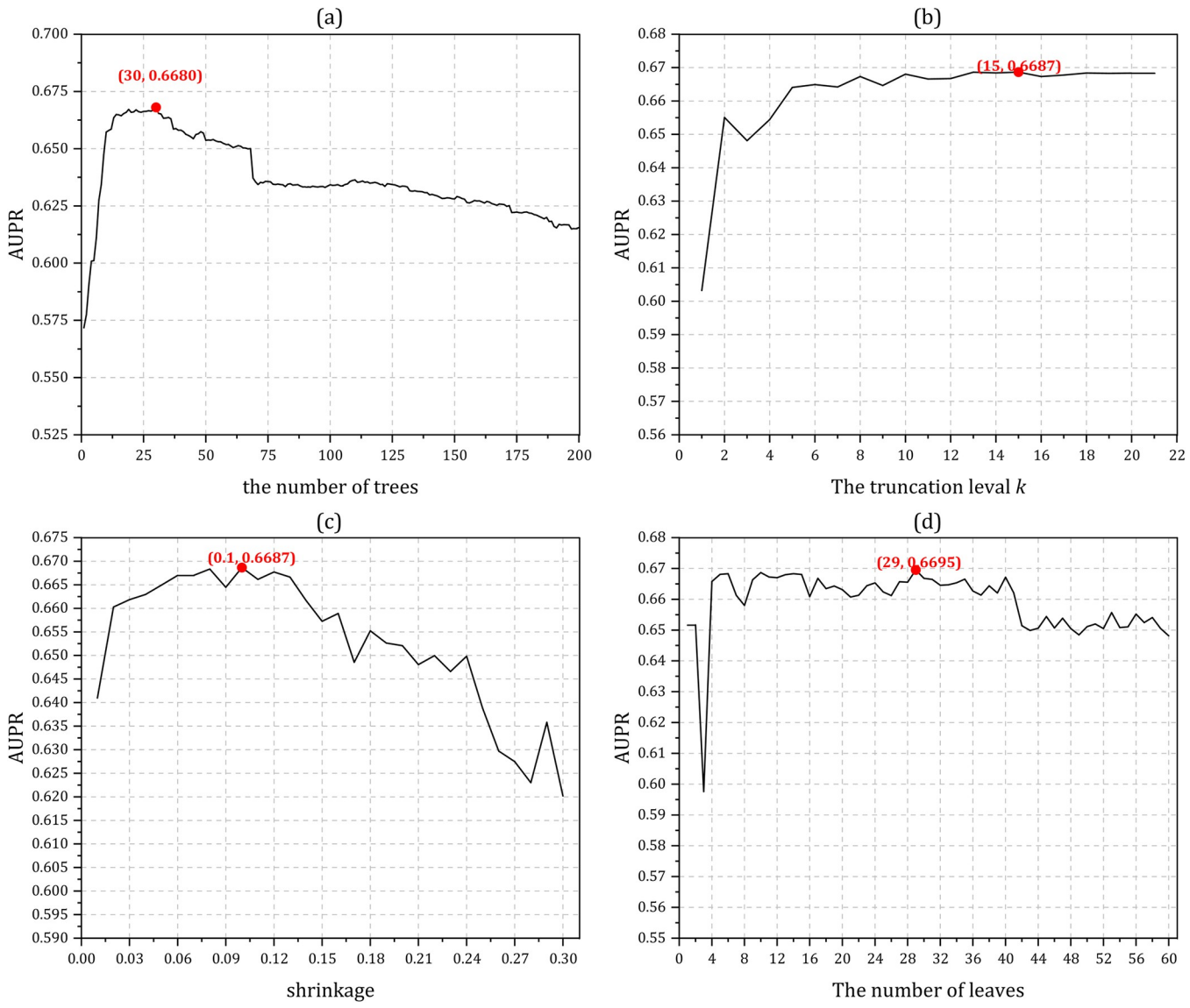


**Fig 4. The predictive results of iPiDA-LTR predictor on $\mathbb{S}^{p}_{ben}$ dataset via five-fold cross-validation.** (a) The truncation level *k*, shrinkage and the number of leaves are assigned as 10, 0.10, and 10 respectively, which are the RankLib's default values (https://sourceforge.net/p/lemur/wiki/RankLib/), and the optimal value of the number of trees is 30; (b) The number of trees, shrinkage and the number of leaves are fixed as 30, 0.10 and 10 respectively, and the truncation level *k* is optimized as 15; (c) The number of trees, the truncation level *k* and the number of leaves are fixed as 30, 15 and 10 respectively, and the shrinkage is optimized as 0.10; (d) The number of trees, the truncation level *k* and shrinkage are 30, 15 and 0.1 respectively, and the number of leaves is optimized as 29.

https://doi.org/10.1371/journal.pcbi.1010404.g004

**Table 2. The comparison results of predictors based on Learning to Rank integrating different component methods via five-fold cross-validation on $\mathbb{S}^{a}_{ben}$ dataset.**

| | AUC | AUPR | NDCG@5 | MAP | ROC1 | ROC3 | ROC5 |
|---|---|---|---|---|---|---|---|
| **iPiDA-LTR-ML**[a] | 0.9511 | 0.9003 | 0.9492 | 0.9305 | 0.8678 | 0.9407 | 0.9503 |
| **iPiDA-LTR**[b] | 0.9543 | 0.9111 | 0.9545 | 0.9379 | 0.8822 | 0.9457 | 0.9538 |

[a] The component methods include RF, LR and SVM

[b] The component methods include RF, LR, SVM and CF.

https://doi.org/10.1371/journal.pcbi.1010404.t002

**Table 3. The comparison results of predictors based on Learning to Rank integrating different component methods via five-fold cross-validation on and $\mathbb{S}^{p}_{ben}$ dataset.**

| | AUC | AUPR | NDCG@5 | MAP | ROC1 | ROC3 | ROC5 |
|---|---|---|---|---|---|---|---|
| **iPiDA-LTR-ML**[a] | 0.9544 | 0.6243 | 0.7722 | 0.7299 | 0.5103 | 0.7779 | 0.8478 |
| **iPiDA-LTR**[b] | 0.9558 | 0.6695 | 0.7884 | 0.7581 | 0.5725 | 0.7918 | 0.8515 |

[a] The component methods include RF, LR and SVM

[b] The component methods include RF, LR, SVM and CF.

https://doi.org/10.1371/journal.pcbi.1010404.t003

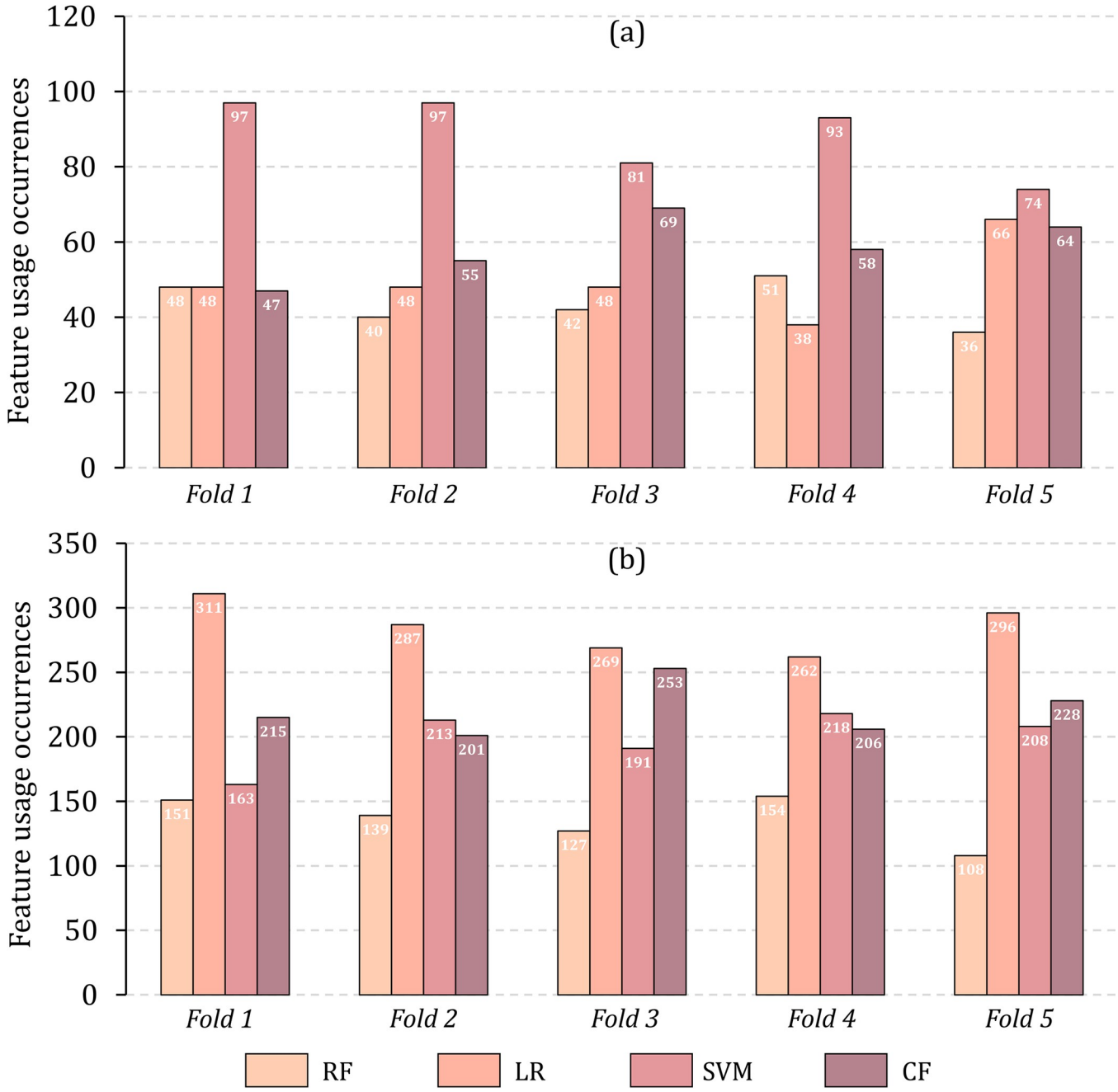


**Fig 5. The usage frequencies of component methods, calculated by Apache commons-math3 library and RankLib library.** (a) shows the usage frequencies of component methods on $\mathbb{S}^{a}_{ben}$ dataset via five-fold cross-validation; (b) shows the usage frequencies of component methods on $\mathbb{S}^{p}_{ben}$ dataset via five cross-validation.

https://doi.org/10.1371/journal.pcbi.1010404.g005

**Table 4. The comparison results between iPiDA-LTR and two state-of-the-art predictors on $\mathbb{S}^{a}_{ind}$ dataset.**

| | AUC | AUPR | NDCG@5 | MAP |
|---|---|---|---|---|
| **iPiDi-PUL** | 0.9153 | 0.8511 | 0.9190 | 0.8847 |
| **iPiDA-sHN** | 0.8042 | 0.7023 | 0.8198 | 0.7705 |
| **iPiDA-LTR** | 0.9521 | 0.8987 | 0.9472 | 0.9283 |

Note: iPiDi-PUL and iPiDA-sHN are reproduced, and their parameters are set as the optimized values reported in [25] and [26], respectively.

https://doi.org/10.1371/journal.pcbi.1010404.t004

**Table 5. The comparison results between iPiDA-LTR and two state-of-the-art $\mathbb{S}^{p}_{ind}$ dataset.**

| | AUC | AUPR | NDCG@5 | MAP |
|---|---|---|---|---|
| **iPiDi-PUL** | 0.9413 | 0.6154 | 0.7736 | 0.7110 |
| **iPiDA-sHN** | 0.8015 | 0.3702 | 0.4875 | 0.4583 |
| **iPiDA-LTR** | 0.9623 | 0.6780 | 0.8067 | 0.7697 |

Note: iPiDi-PUL and iPiDA-sHN are reproduced, and their parameters are set as the optimized values reported in [25] and [26], respectively.

https://doi.org/10.1371/journal.pcbi.1010404.t005

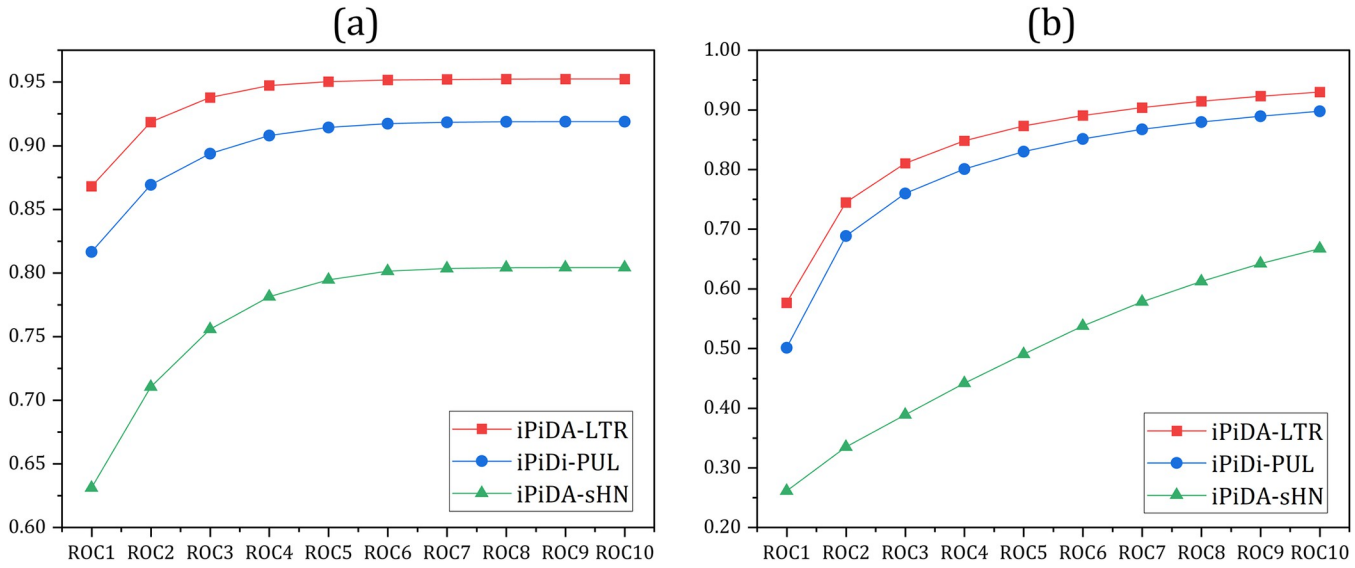


**Fig 6. The comparison results of different methods.** (a) and (b) are based on $\mathbb{S}^{a}_{ind}$ dataset and $\mathbb{S}^{p}_{ind}$ dataset, respectively.

https://doi.org/10.1371/journal.pcbi.1010404.g006

**Table 6. The top five piR-hsa-23210 associated diseases and relevant evidences.**

| Rank | disease name | Evidence |
|---|---|---|
| 1 | Cardiovascular diseases (CDC, CF, CCS) cardicregeneration | PMID: 28289238 |
| 2 | Renal Cell Carcinoma | PMID: 25998508 |
| 3 | Alzheimer Disease | PMID: 28127595 |
| 4 | Male Infertility | PMID: 24855106 |
| 5 | Gastric Cancer | PMID: 25779424 |

Note: the evidences can be found at https://pubmed.ncbi.nlm.nih.gov/.

https://doi.org/10.1371/journal.pcbi.1010404.t006

**Table 7. The top five piR-hsa-15023 associated diseases and relevant evidences.**

| Rank | disease name | Evidence |
|---|---|---|
| 1 | Cardiovascular diseases (CDC, CF, CCS) cardicregeneration | PMID: 28289238 |
| 2 | Renal Cell Carcinoma | PMID: 26071182 |
| 3 | Alzheimer Disease | PMID: 28127595 |
| 4 | Rheumatoid Arthritis | None |
| 5 | Gastric Cancer | PMID: 25779424 |

Note: the evidences can be found at https://pubmed.ncbi.nlm.nih.gov/.

https://doi.org/10.1371/journal.pcbi.1010404.t007

For example, the target gene of piR-hsa-23210 is SMC5, which plays crucial roles in the process of human spermatogenesis, such as on the synaptonemal complex between synapsed chromosomes, and in the development of spermatogonial cells [74]. Roy et al. found that piR-33044 (piR-hsa-23210) is significantly abnormal expression in Alzheimer Disease [22]. (ii) Four diseases in **Table 7** have been proved to be associated with piR-hsa-15023. For example, Busch et al. found that piR-hsa-15023 is down-regulated in renal cell carcinoma [75]. piR-hsa-15023 showed a significantly differentially expression in gastric adenocarcinoma and non-malignant stomach tissue [76]. Therefore, these results demonstrated that iPiDA-LTR predictor is an effective approach to identify associated diseases for newly detected query piRNAs.

## Conclusion

In this study, we treat the task of piRNA-disease associations as a search task based on Learning to Rank [32, 68], where piRNA and disease are regarded as query and document, respectively. The following conclusions can be drawn: (i) iPiDA-LTR can effectively handle with two types of application scenarios compared with the other state-of-the-art methods, especially for the identification of diseases associated with newly detected piRNAs, which is important for studying the pathogenesis of disease and the function of piRNAs; (ii) iPiDA-LTR incorporates component methods into Learning to Rank so as to improve the predictive performance; (iii) The corresponding web server of iPiDA-LTR is freely accessed at http://bliulab.net/iPiDA-LTR/. Although iPiDA-LTR effectively predicts piRNA-disease associations, it only integrates basic machine learning methods and collaborative filtering. In future studies, we will integrate the other state-of-the-art methods and features to improve piRNA-disease associations. The LTR-based framework discussed in this study is a general framework, which would have many other applications in bioinformatics, such as protein function prediction, remote homology detection, etc.

## Author Contributions

**Conceptualization:** Bin Liu.

**Data curation:** Wenxiang Zhang.

**Formal analysis:** Wenxiang Zhang, Bin Liu.

**Funding acquisition:** Bin Liu.

**Investigation:** Wenxiang Zhang, Bin Liu.

**Methodology:** Wenxiang Zhang, Bin Liu.

**Software:** Wenxiang Zhang.

**Supervision:** Bin Liu.

**Validation:** Wenxiang Zhang, Jialu Hou.

**Writing – original draft:** Wenxiang Zhang.

**Writing – review & editing:** Wenxiang Zhang, Jialu Hou, Bin Liu.

## References


1. Seto AG, Kingston RE, Lau NC. The coming of age for Piwi proteins. Molecular cell. 2007; 26(5):603–9. Epub 2007/06/15. https://doi.org/10.1016/j.molcel.2007.05.021 PMID: 17560367.
2. Kirino Y, Mourelatos Z. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nature structural & molecular biology. 2007; 14(4):347–8. Epub 2007/03/27. https://doi.org/10.1038/nsmb1218 PMID: 17384647.
3. Ohara T, Sakaguchi Y, Suzuki T, Ueda H, Miyauchi K, Suzuki T. The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. Nature structural & molecular biology. 2007; 14(4):349–50. Epub 2007/03/27. https://doi.org/10.1038/nsmb1220 PMID: 17384646.
4. 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. Epub 2006/06/06. https://doi.org/10.1038/nature04916 PMID: 16751777.
5. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA complex from rat testes. Science. 2006; 313(5785):363–7. Epub 2006/06/17. https://doi.org/10.1126/science.1130164 PMID: 16778019.
6. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007; 128(6):1089–103. Epub 2007/03/10. https://doi.org/10.1016/j.cell.2007.01.043 PMID: 17346786.
7. Yu L, Su Y, Liu Y, Zeng X. Review of unsupervised pretraining strategies for molecules representation. Briefings in Functional Genomics. 2021; 20(5):323–32. https://doi.org/10.1093/bfgp/elab036 PMID: 34342611
8. Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Current Opinion in Structural Biology. 2022; 72:114–26. https://doi.org/10.1016/j.sbi.2021.09.003 PMID: 34649044
9. 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. Epub 2006/07/01. https://doi.org/10.1126/science.1129333 PMID: 16809489.
10. 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. Epub 2017/12/07. https://doi.org/10.1038/nature25018 PMID: 29211718; PubMed Central PMCID: PMC5933846.
11. Lim AK, Tao L, Kai T. piRNAs mediate posttranscriptional retroelement silencing and localization to pi-bodies in the Drosophila germline. Journal of cell biology. 2009; 186(3):333–42. Epub 2009/08/05. https://doi.org/10.1083/jcb.200904063 PMID: 19651888; PubMed Central PMCID: PMC2728408.
12. Singh G, Swain AC, Mallick B. Delineating Characteristic Sequence and Structural Features of Precursor and Mature Piwi-interacting RNAs of Epithelial Ovarian Cancer. Current Bioinformatics. 2021; 16 (4):541–52. https://doi.org/10.2174/1574893615999200715164755 WOS:000669437400006.
13. Qiu W, Guo X, Lin X, Yang Q, Zhang W, Zhang Y, et al. Transcriptome-wide piRNA profiling in human brains of Alzheimer's disease. Neurobiology of aging. 2017; 57:170–7. Epub 2017/06/28. https://doi.org/10.1016/j.neurobiolaging.2017.05.020 PMID: 28654860; PubMed Central PMCID: PMC5542056.
14. 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. Epub 2011/05/28. https://doi.org/10.1016/j.cca.2011.05.015 PMID: 21616063.
15. 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):123. Epub 2019/08/11. https://doi.org/10.1186/s12943-019-1052-9 PMID: 31399034; PubMed Central PMCID: PMC6688334.
16. Liu Y, Li A, Xie G, Liu G, Hei X. Computational Methods and Online Resources for Identification of piRNA-Related Molecules. Interdisciplinary Sciences-Computational Life Sciences. 2021; 13(2):176–91. Epub 2021/04/23. https://doi.org/10.1007/s12539-021-00428-5 PMID: 33886096.
17. Ding X, Li Y, Lü J, Zhao Q, Guo Y, Lu Z, et al. piRNA-823 Is Involved in Cancer Stem Cell Regulation Through Altering DNA Methylation in Association With Luminal Breast Cancer. Frontiers in cell and

developmental biology. 2021; 9:641052. Epub 2021/04/02. https://doi.org/10.3389/fcell.2021.641052 PMID: 33791297; PubMed Central PMCID: PMC8005588.

18. Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. Briefings in Bioinformatics. 2021; 22(6):bbab344. https://doi.org/10.1093/bib/bbab344 PMID: 34415297

19. Zeng X, Song X, Ma T, Pan X, Zhou Y, Hou Y, et al. Repurpose open data to discover therapeutics for COVID-19 using deep learning. Journal of proteome research. 2020; 19(11):4624–36. https://doi.org/10.1021/acs.jproteome.0c00316 PMID: 32654489

20. Cabral GF, Pinheiro JADS, Vidal AF, Santos S, Ribeiro-Dos-Santos Â. piRNAs in Gastric Cancer: A New Approach Towards Translational Research. International journal of molecular sciences. 2020; 21 (6):2126. Epub 2020/03/25. https://doi.org/10.3390/ijms21062126 PMID: 32204558; PubMed Central PMCID: PMC7139476.

21. Krishnan P, Ghosh S, Graham K, Mackey JR, Kovalchuk O, Damaraju S. Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. Oncotarget. 2016; 7(25):37944–56. Epub 2016/10/23. https://doi.org/10.18632/oncotarget.9272 PMID: 27177224; PubMed Central PMCID: PMC5122362.

22. 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. Epub 2017/01/28. https://doi.org/10.1039/c6mb00699j PMID: 28127595.

23. Zhang W, Wei H, Liu B. idenMD-NRF: a ranking framework for miRNA-disease association identification. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbac224 PMID: 35679537

24. Wei H, Xu Y, Liu B. iCircDA-LTR: identification of circRNA-disease associations based on Learning to Rank. Bioinformatics. 2021; 37(19):3302–10. Epub 2021/05/09. https://doi.org/10.1093/bioinformatics/btab334 PMID: 33963827.

25. 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

26. 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. Epub 2020/09/12. https://doi.org/10.1016/j.compbiolchem.2020.107361 PMID: 32916452.

27. Zhang P, Si X, Skogerbø G, Wang J, Cui D, Li Y, et al. piRBase: a web resource assisting piRNA functional study. Database. 2014; 2014:bau110. Epub 2014/11/27. https://doi.org/10.1093/database/bau110 PMID: 25425034; PubMed Central PMCID: PMC4243270.

28. Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. Nucleic acids research. 2016; 44(D1):D223–30. Epub 2015/11/20. https://doi.org/10.1093/nar/gkv1265 PMID: 26582915; PubMed Central PMCID: PMC4702893.

29. Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic acids research. 2019; 47(D1):D175–D80. Epub 2018/10/30. https://doi.org/10.1093/nar/gky1043 PMID: 30371818; PubMed Central PMCID: PMC6323959.

30. Hang LJITol Systems. A Short Introduction to Learning to Rank. 2011; 94-D(10):1854–62.

31. Song B, Li F, Liu Y, Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Briefings in Bioinformatics. 2021; 22(6):bbab282. https://doi.org/10.1093/bib/bbab282 PMID: 34308472

32. Burges CJC. From ranknet to lambdarank to lambdamart: An overview. Learning. 2010; 11(23–581):81.

33. He S, Guo F, Zou Q, Ding H. MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. Current Bioinformatics. 2020; 15(10):1213–21. https://doi.org/10.2174/1574893615999200503030350 WOS:000617680800012.

34. Wang X, Li C, Golbandi N, Bendersky M, Najork M, editors. The lambdaloss framework for ranking metric optimization. Proceedings of the 27th ACM International Conference on Information and Knowledge Management; 2018.

35. Figueroa A, Neumann G, editors. Learning to Rank Effective Paraphrases from Query Logs for Community Question Answering. Twenty-seventh Aaai Conference on Artificial Intelligence; 2013.

36. Liu TY, Xu J, Qin T, Xiong W, Li H. LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. Proceedings of the Workshop on Learning to Rank for Information Retrieval2007. p. 137–45.

37. Jin X, Liao Q, Wei H, Zhang J, Liu B. SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection. Bioinformatics. 2021; 37(7):913–20. https://doi.org/10.1093/bioinformatics/btaa772 MEDLINE: PMID: 32898222.

38. Liu L, Huang X, Mamitsuka H, Zhu S. HPOLabeler: improving prediction of human protein-phenotype associations by learning to rank. Bioinformatics. 2020; 36(14):4180–8. Epub 2020/05/08. https://doi.org/10.1093/bioinformatics/btaa284 PMID: 32379868.

39. Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. Bioinformatics. 2015; 31(21):3492–8. https://doi.org/10.1093/bioinformatics/btv413 WOS:000365134400013. PMID: 26163693

40. Muhammad A, Waheed R, Khan NA, Jiang H, Song X. piRDisease v1.0: a manually curated database for piRNA associated diseases. Database. 2019; 2019:baz052. Epub 2019/07/04. https://doi.org/10.1093/database/baz052 PMID: 31267133; PubMed Central PMCID: PMC6606758.

41. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981; 147(1):195–7. Epub 1981/03/25. https://doi.org/10.1016/0022-2836(81)90087-5 PMID: 7265238.

42. Li H-L, Pang Y-H, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic Acids Res. 2021; 49(22):e129. Epub 2021/09/29. https://doi.org/10.1093/nar/gkab829 PMID: 34581805; PubMed Central PMCID: PMC8682797.

43. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019; 47(20):e127. Epub 2019/09/11. https://doi.org/10.1093/nar/gkz740 PMID: 31504851; PubMed Central PMCID: PMC6847461.

44. Wei H, Liao Q, Liu B. iLncRNAdis-FB: identify lncRNA-disease associations by fusing biological feature blocks through deep neural network. IEEE/ACM transactions on computational biology and bioinformatics. 2021; 18(5):1946–57. https://doi.org/10.1109/TCBB.2020.2964221 MEDLINE: PMID: 31905146.

45. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017; 35(11):1026–8. Epub 2017/10/17. https://doi.org/10.1038/nbt.3988 PMID: 29035372.

46. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021; 18(4):366–8. Epub 2021/04/09. https://doi.org/10.1038/s41592-021-01101-x PMID: 33828273; PubMed Central PMCID: PMC8026399.

47. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019; 47(D1):D955–D62. Epub 2018/11/09. https://doi.org/10.1093/nar/gky1032 PMID: 30407550; PubMed Central PMCID: PMC6323977.

48. Chen X, Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. Scientific reports. 2014; 4:5501. Epub 2014/07/01. https://doi.org/10.1038/srep05501 PMID: 24975600; PubMed Central PMCID: PMC4074792.

49. 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. Epub 2018/06/26. https://doi.org/10.1093/bioinformatics/bty503 PMID: 29939227.

50. Wang L, Xuan Z, Zhou S, Kuang L, Pei T. A Novel Model for Predicting LncRNA-disease Associations Based on the LncRNA-MiRNA-disease Interactive Network. Current Bioinformatics. 2019; 14(3):269–78. https://doi.org/10.2174/1574893613666180703105258 WOS:000460522300011.

51. Luo J, Xiao Q. A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. Journal of Biomedical Informatics. 2017; 66:194–203. https://doi.org/10.1016/j.jbi.2017.01.008 WOS:000409293100018. PMID: 28104458

52. Yan C, Wang J, Ni P, Lan W, Wu F-X, Pan Y. DNRLMF-MDA:Predicting microRNA-Disease Associations Based on Similarities of microRNAs and Diseases. IEEE/ACM transactions on computational biology and bioinformatics. 2019; 16(1):233–43. https://doi.org/10.1109/TCBB.2017.2776101 PMID: 29990253

53. Zhu Q, Fan Y, Pan X. Fusing Multiple Biological Networks to Effectively Predict miRNA-disease Associations. Current Bioinformatics. 2021; 16(3):371–84. https://doi.org/10.2174/1574893615999200715165335 WOS:000636235900003.

54. 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. WOS:000247348300013. https://doi.org/10.1093/bioinformatics/btm087 PMID: 17344234

55. Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM transactions on computational biology and bioinformatics. 2016; 14(4):905–15. https://doi.org/10.1109/TCBB.2016.2550432 PMID: 27076459

56. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.

57. Wu C, Lin B, Shi K, Zhang Q, Gao R, Yu Z, et al. PEPRF: Identification of Essential Proteins by Integrating Topological Features of PPI Network and Sequence-Based Features via Random Forest. Current

Bioinformatics. 2021; 16(9):1161–8. https://doi.org/10.2174/1574893616666210617162258 WOS:000711656200006.

58. Ao C, Zou Q, Yu L. NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. Briefings in bioinformatics. 2021; 23(1):bbab480. https://doi.org/10.1093/bib/bbab480 MEDLINE: PMID: 34850821.

59. Zeng X, Zhong Y, Lin W, Zou Q. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Briefings in bioinformatics. 2020; 21(4):1425–36. https://doi.org/10.1093/bib/bbz080 PMID: 31612203

60. Zeng X, Zhu S, Hou Y, Zhang P, Li L, Li J, et al. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. Bioinformatics. 2020; 36(9):2805–12. https://doi.org/10.1093/bioinformatics/btaa010 PMID: 31971579

61. Landwehr N, Hall M, Frank E. Logistic Model Trees. Machine Learning. 2005; 59(1):161–205.

62. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications. 1998; 13(4):18–28.

63. Muflikhah L, Widodo N, Mahmudy WF, Solimun. Detection of Hepatitis B Virus-associated Hepatocellular Carcinoma Disease Using Hybrid Hierarchical k-Means Clustering and SVM Algorithm. Current Bioinformatics. 2021; 16(7):1004–12. https://doi.org/10.2174/1574893615999200626185251 WOS:000726375800013.

64. Basith S, Hasan MM, Lee G, Wei L, Manavalan B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. Brief Bioinform. 2021; 22(6). Epub 2021/07/07. https://doi.org/10.1093/bib/bbab252 PMID: 34226917.

65. Yue W, Wang Z, Zhang J, Liu X. An Overview of Recommendation Techniques and Their Applications in Healthcare. IEEE/CAA Journal of Automatica Sinica. 2021; 8(4):701–17. https://doi.org/10.1109/jas.2021.1003919 WOS:000628913100001.

66. Bayrak T, Ogul H. A New Approach for Predicting the Value of Gene Expression: Two-way Collaborative Filtering. Current Bioinformatics. 2019; 14(6):480–90. https://doi.org/10.2174/1574893614666190126144139 WOS:000475702400002.

67. Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics. 2016; 32(12):i18–i27. Epub 2016/06/17. https://doi.org/10.1093/bioinformatics/btw244 PMID: 27307615; PubMed Central PMCID: PMC4908328.

68. Li H, Systems. A Short Introduction to Learning to Rank. IEICE TRANSACTIONS on Information and Systems. 2011; 94(10):1854–62.

69. Zhu L, Duan G, Yan C, Wang J. Prediction of Microbe-drug Associations Based on Chemical Structures and the KATZ Measure. Current Bioinformatics. 2021; 16(6):807–19. https://doi.org/10.2174/1574893616666210204144721 WOS:000684207300006.

70. Dao F-Y, Lv H, Zhang D, Zhang Z-M, Liu L, Lin H. DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops. Briefings in bioinformatics. 2021; 22(4):bbaa356. https://doi.org/10.1093/bib/bbaa356 PMID: 33279983.

71. Zhang D, Chen H-D, Zulfiqar H, Yuan S-S, Huang Q-L, Zhang Z-Y, et al. iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. Computational and Mathematical Methods in Medicine. 2021; 2021:6664362. https://doi.org/10.1155/2021/6664362 PMID: 33505515

72. Li J, Liu L, Cui Q, Zhou Y. Comparisons of MicroRNA Set Enrichment Analysis Tools on Cancer De-regulated miRNAs from TCGA Expression Datasets. Current Bioinformatics. 2020; 15(10):1104–12. https://doi.org/10.2174/1574893615666200224095041 WOS:000617680800002.

73. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011; 12:2825–30.

74. Verver DE, Langedijk NSM, Jordan PW, Repping S, Hamer G. The SMC5/6 complex is involved in crucial processes during human spermatogenesis. Biology of reproduction. 2014; 91(1):22. Epub 2014/05/24. https://doi.org/10.1095/biolreprod.114.118596 PMID: 24855106; PubMed Central PMCID: PMC6058740.

75. 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):61. Epub 2015/06/14. https://doi.org/10.1186/s13046-015-0180-3 PMID: 26071182; PubMed Central PMCID: PMC4467205.

76. Martinez VD, Enfield KSS, Rowbotham DA, Lam WL. An atlas of gastric PIWI-interacting RNA transcriptomes and their utility for identifying signatures of gastric cancer recurrence. Gastric Cancer. 2016; 19(2):660–5. Epub 2015/03/18. https://doi.org/10.1007/s10120-015-0487-y PMID: 25779424; PubMed Central PMCID: PMC4573768.