frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 February 2023
DOI 10.3389/fgene.2023.1122909



OPEN ACCESS

EDITED BY
Pu-Feng Du,
Tianjin University, China

REVIEWED BY
Li Peng,
Hunan University of Science and Technology, China
Chen Qingfeng,
Guangxi University, China

*CORRESPONDENCE
Chang-Qing Yu,
xaycq@163.com
Li-Ping Li,
cs2bioinformatics@gmail.com

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 13 December 2022
ACCEPTED 30 January 2023
PUBLISHED 10 February 2023

CITATION
Wei M-M, Yu C-Q, Li L-P, You Z-H, Ren Z-H, Guan Y-J, Wang X-F and Li Y-C (2023), LPIH2V: LncRNA-protein interactions prediction using HIN2Vec based on heterogeneous networks model. *Front. Genet.* 14:1122909. doi: 10.3389/fgene.2023.1122909



# LPIH2V: LncRNA-protein interactions prediction using HIN2Vec based on heterogeneous networks model

Meng-Meng Wei[1], Chang-Qing Yu[1]*, Li-Ping Li[1,2]*, Zhu-Hong You[3], Zhong-Hao Ren[1], Yong-Jian Guan[1], Xin-Fei Wang[4] and Yue-Chao Li[4]

[1]School of Information Engineering, Xijing University, Xi'an, China, [2]College of Grassland and Environment Sciences, Xinjiang Agricultural University, Urumqi, China, [3]School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [4]Xijing University, Xi'an, China

LncRNA-protein interaction plays an important role in the development and treatment of many human diseases. As the experimental approaches to determine lncRNA–protein interactions are expensive and time-consuming, considering that there are few calculation methods, therefore, it is urgent to develop efficient and accurate methods to predict lncRNA-protein interactions. In this work, a model for heterogeneous network embedding based on meta-path, namely LPIH2V, is proposed. The heterogeneous network is composed of lncRNA similarity networks, protein similarity networks, and known lncRNA-protein interaction networks. The behavioral features are extracted in a heterogeneous network using the HIN2Vec method of network embedding. The results showed that LPIH2V obtains an AUC of 0.97 and ACC of 0.95 in the 5-fold cross-validation test. The model successfully showed superiority and good generalization ability. Compared to other models, LPIH2V not only extracts attribute characteristics by similarity, but also acquires behavior properties by meta-path wandering in heterogeneous networks. LPIH2V would be beneficial in forecasting interactions between lncRNA and protein.

KEYWORDS

**lncRNA-protein interaction, heterogeneous information network, network embedding, HIN2Vec, behavioral features**

## 1 Introduction

LncRNAs are a group of RNA molecules that are transcribed and do not have the competence to code for proteins. LncRNAs are normally defined as RNAs over 200 nucleotides long with no potential for symbolic coding, often playing administrative roles in the regulation of gene expression (Prensner and Chinnaiyan, 2011; Volders et al., 2013). LncRNAs can act as gene controllers and are involved in many of the biological processes as other epigenetic mechanisms (Esteller, 2011). LncRNA is poorly identified in the genome and is known as the "dark matter" of the genome. The function of most ncRNAs remains unclear, and a lot of lncRNAs may have no apparent function. In recent decades, a growing number of studies have uncovered that lncRNA plays an influential role in many biological processes (Wang et al., 2018). With ongoing developments in deep RNA sequencing and advanced epigenomics technologies, the rate of discovery of new lncRNA genes is quickly outstripping the rate at which they can be described. When lncRNAs are out of order, they may induce a variety

of diseases, such as Alzheimer (Ng et al., 2013), autism (Cook and Scherer, 2008), cardiovascular diseases (Congrains et al., 2012), and cancer (Chen et al., 2017; Rathinasamy and Velmurugan, 2018). For instance, genetic deletion of the lncRNA locus on human chromosome 2 leads to serious congenital limb deformities (Allou et al., 2021). LncRNAs play regulatory roles in immune response in prostate cancer (Hu et al., 2021). PlncRNA-1 significantly increased in prostate cancer cells (Cui et al., 2013). ANRIL is obviously associated with coronary heart disease, type 2 diabetes and many types of cancer (Pasmant et al., 2011). A new lncRNA, FASRL, was recently discovered to boost the proliferation of hepatocellular carcinoma (HCC) cells *in vitro* and *in vivo* (Peng et al., 2022).

In recent years, probing the interactions between lncRNAs and proteins has been one of the primary ways to deduce the functions of lncRNAs and to do further research on lncRNAs. LncRNA-protein interactions (LPIs) throughout their lifetimes, regulating not only maturation, nuclear export, stability, and eventually translation, but also the functions of RNAs. Given the critical role of lncRNA in various biological processes and complex diseases, there is an urgent need to uncover potential lncRNA-protein associations. Hence, to effectively predict LPIs, a variety of methods have recently been proposed, classified into two broad categories, including experimental and computational methods (Zhang and Fan, 2017; Li et al., 2022). Given the number and variety of lncRNA and protein, an exhaustive experimental validation of each lncRNA and protein would be impractical (Alipanahi et al., 2015). Therefore the need for the use of computational methods to screen for potential LPIs in these high-throughput assays and then validate them experimentally. Current computational methods fall into two broad categories: network-based and machine-learning methods.

Nowadays, several computational methods for predicting LPIs have been proposed. Muppirala et al. (2011) proposed RPISeq, a family of classifiers for predicting RNA-protein interactions, using only sequence information. This model extracts features from lncRNA and protein sequences using two classifiers, support vector machine (SVM) and random forest (RF), respectively, and the results show that the SVM classifier predicts more accurately. Wang et al. (2013) proposed the model, which is based on a Bayesian classifier that first collects a set of known RNA-protein interactions as positive criteria and extracts sequence-based features to represent each RNA-protein pair, selecting valid features by reducing the likelihood ratio score. These valid features are used to build an extended Bayesian classifier for training RNA-protein interaction prediction. Lu et al. (2013) transformed the sequences of amino acids into numerical feature vectors. This model which is named lncPro transformed the sequence information of lncRNAs and proteins into feature vectors then reduced the dimensionality of the lncRNA and protein feature vectors using Fourier series, and finally integrated the information of both using matrix multiplication to score each lncRNA-protein pair. A new sequence distributed representation learning-based method for potential LPI prediction, named LPI-Pred, was developed by Yi et al. (2020a) which regarded lncRNA and protein sequences as "words" in natural language processing and trained the RNA2vec and Pro2vec models using word2vec. However, the above approach only considers information about the properties of lncRNAs and proteins and not their behavior, which has now been demonstrated in many articles in the field of bioinformatics to be powerful in improving the prediction of LPIs (Guan et al., 2022; Guo et al., 2022; Ren et al., 2022). Ge et al. (2016) which proposed and tested a new computational approach, LPBNI, which constructs a bipartite network of lncRNA-proteins, using information about LPIs to link lncRNAs and proteins if they are known to interact with each other. Yang et al. (2016) proposed a model which represented and analyzed the interactions between lncRNAs and proteins as a heterogeneous network. The model used a correlation-search algorithm called HeteSim to predict lncRNA interactions with proteins. Deng et al. (2018) proposed a model, PLIPCOM, which constructed a network from known sequence similarities and LPIs information. First, the model used the random walk algorithm to extract diffusion features and HeteSim features, then reduced the dimensionality by the SVD algorithm, respectively. While the above approach takes into account the acquisition of behavioral information through network embedding, it does not take into account the role of meta-paths in heterogeneous networks.

In this article, a novel model for predicting LPIs based on heterogeneous network embedding, LPIH2V, was proposed. Potential vectors of nodes in heterogeneous networks are learned and represented using a neural network-based HIN2Vec (Fu et al., 2017). The extracted feature vectors are then used to predict LPIs using an SVM classifier. After the 5-fold cross-validation test, the results show that LPIH2V has high accuracy and stability. At the same time, the LPIH2V model achieved the best prediction accuracy compared to known models for the same dataset.

# 2 Results and discussion

## 2.1 Evaluation criteria

In this experiment, we evaluated the performance of LPIH2V using commonly adopted metrics and the 5-fold cross-validation method (Yu et al., 2022). The dataset is divided into five equal subgroups, with data from each subset used for testing in turn and data from the remaining four subsets used for training data. We repeated this process to ensure that each part served as the test set. The mean of the five predictions was ultimately used as the final evaluation result. Six metrics were used to evaluate LPIH2V performance: accuracy, precision, recall, F1 score, SPE, and MCC. For the evaluation criteria above, higher values represent better performance and better performance.

A common description of the ACC is the systematic error, which indicates the discrepancy between the predicted result and the true value. PRE refers to the proportion of true positives in statistical and diagnostic testing. REC measures the proportion of correct positive identifications, also refers to as sensitivity. F1 scores represent the summed mean of accuracy and sensitivity. SPEC is the probability of a negative test result. MCC is the correlation coefficient between true and predicted values. TP and TN are the numbers of correctly identified positive and negative samples, respectively. FN and FP are the numbers of incorrectly identified positive and negative samples, respectively. These metrics can be defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$PRE = \frac{TP}{TP + FP}$$

$$REC = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}$$

$$SPEC = \frac{TN}{FP + TN}$$

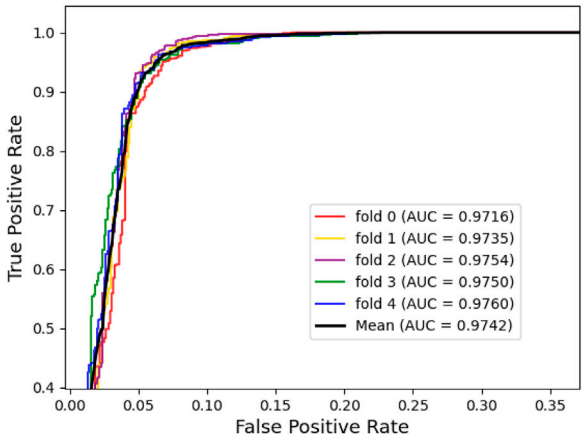


FIGURE 1
ROC curves plot for LPIH2V.

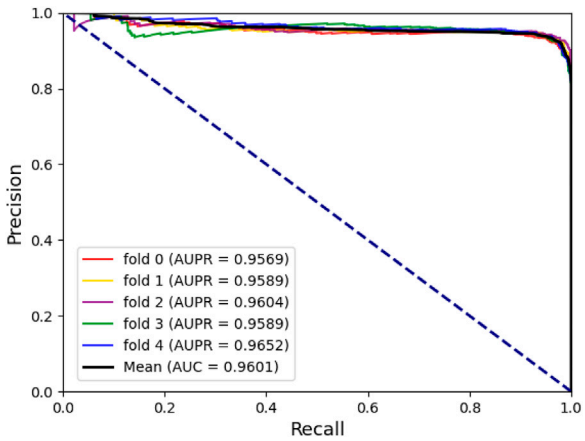


FIGURE 2
PRE curves plot for LPIH2V.

**TABLE 1** 5-fold cross-validation results for LPIH2V.

| Test set | PRE(%) | REC (%) | SPE (%) | ACC(%) | MCC(%) | F1 (%) | AUC(%) |
|---|---|---|---|---|---|---|---|
| 1 | 94.64 | 98.47 | 94.43 | 96.45 | 92.97 | 96.52 | 97.90 |
| 2 | 91.94 | 98.47 | 91.37 | 94.92 | 90.06 | 95.09 | 97.40 |
| 3 | 93.33 | 97.92 | 93.01 | 95.46 | 91.04 | 95.57 | 97.10 |
| 4 | 93.13 | 97.81 | 92.79 | 95.30 | 90.72 | 95.42 | 97.90 |
| 5 | 92.75 | 97.81 | 92.35 | 95.08 | 90.30 | 95.21 | 97.50 |
| Average | 93.16 ± 0.88 | 98.10 ± 0.31 | 92.79 ± 0.99 | 95.44 ± 0.54 | 91.02 ± 1.03 | 95.56 ± 0.51 | 97.56 ± 0.31 |

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In addition to the above metrics, we also used AUC, the area under the ROC curve, to evaluate our model. We used the average of the five results to ensure the accuracy of the prediction results.

## 2.2 Evaluation of predictive capability

To assess the predictive power of the model, we used the 5-fold cross-validation and plotted the ROC and PRE curves for the 5 trials, as shown in Figure 1 and Figure 2. The mean predictive AUC for LPIH2V was 0.974. In addition, Table 1

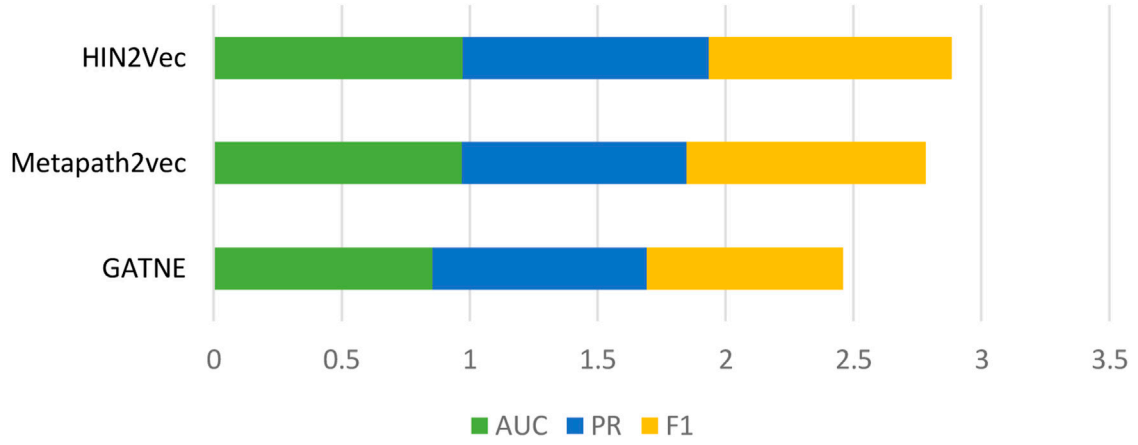


FIGURE 3
Graph embedding methods comparison.

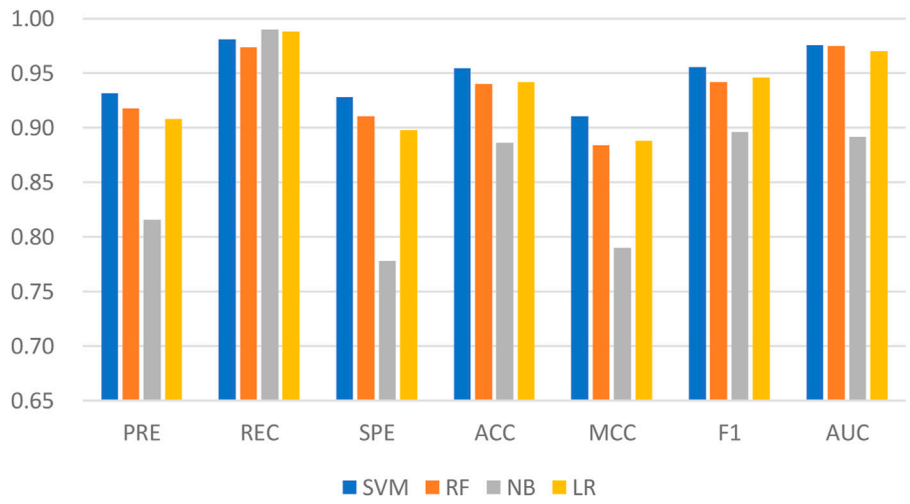


FIGURE 4
Comparison of the performance of LPIH2V and other classifiers.

**TABLE 2 Performance comparison of LPIH2V with other tools for predicting lncRNA-protein interactions.**

| Method | PRE | REC | SPE | ACC | MCC | F1 | AUC |
|---|---|---|---|---|---|---|---|
| CF | 0.583 | 0.894 | 0.361 | 0.627 | 0.301 | 0.706 | 0.761 |
| RWR | 0.739 | 0.798 | 0.717 | 0.757 | 0.517 | 0.767 | 0.830 |
| LPBNI | 0.740 | 0.840 | 0.698 | 0.769 | 0.548 | 0.785 | 0.859 |
| SFPEL-LPI | 0.769 | 0.920 | 0.724 | 0.822 | 0.657 | 0.838 | 0.916 |
| LPIHN | 0.807 | 0.966 | 0.769 | 0.867 | 0.750 | 0.879 | 0.938 |
| LPLNP | 0.832 | 0.943 | 0.810 | 0.876 | 0.761 | 0.884 | 0.944 |
| RPI-SE | 0.877 | 0.974 | 0.863 | 0.919 | 0.843 | 0.923 | 0.959 |
| IPMiner | 0.886 | 0.970 | 0.875 | 0.922 | 0.849 | 0.926 | 0.961 |
| LncPNet | 0.908 | 0.957 | 0.903 | 0.930 | 0.860 | 0.932 | 0.971 |
| LPIH2V | 0.932 | 0.981 | 0.928 | 0.954 | 0.9102 | 0.956 | 0.976 |

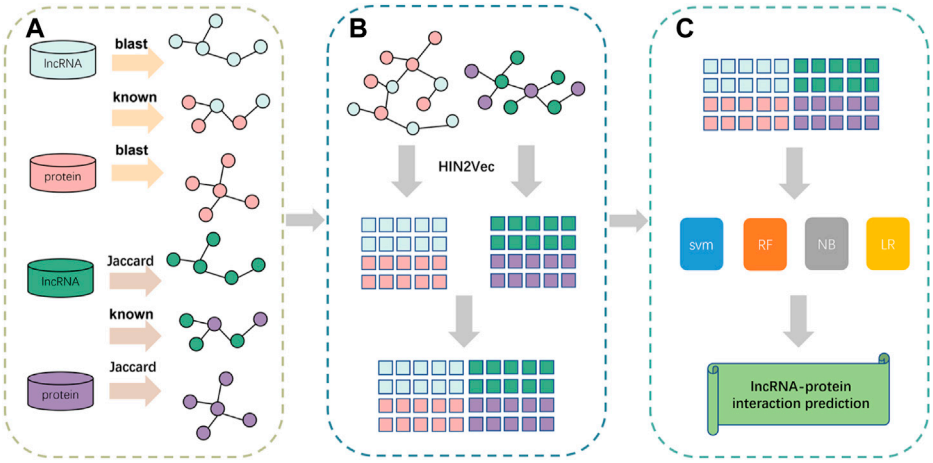


FIGURE 5
Pipeline of the framework of LPIH2V. **(A)** Calculation of similarity between lncRNA-lncRNA and protein-protein. **(B)** Construction of heterogeneous networks using similarity information obtained from **(A)** and extraction of node features using HIN2Vec. **(C)** Use of multiple classifiers to predict LPIs.

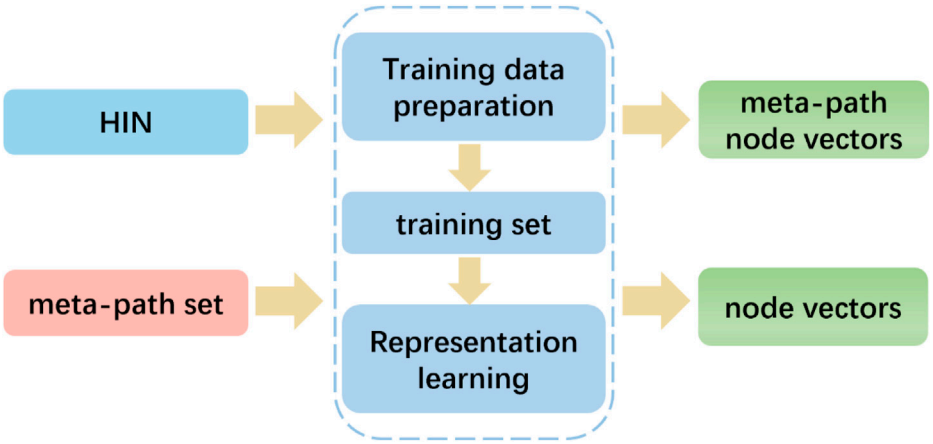


FIGURE 6
HIN2Vec flowchart: Extraction of node eigenvectors using HINs and meta-paths.

summarizes the results of the model under 5-fold cross-validation.

## 2.3 Comparison with graph embedding methods

Behavioral attributes are a very important feature. To demonstrate the predictive power of our model, we compared three heterogeneous network graph embedding methods based on meta-path theory (Yi et al., 2022). GATNE uses base embedding and edge embedding to extract hidden properties of different types of edges between nodes. Metapath2vec travels through the network through custom meta-paths to capture potential behavioral features. HIN2Vec computes meta-paths in heterogeneous networks, then travels across the calculated meta-paths to capture properties of each node across the board. In the same dataset, as shown in Figure 3, the results show that the HIN2Vec method used in our model is effective in predicting LPIs.

## 2.4 Comparing with different base classifiers

Machine learning has been applied successfully to predict LPIs. In the LPIH2V, we used the SVM to classify the integrated features. SVM is one of the well-known algorithms based on statistical learning theory. In order to fully demonstrate the superiority of fusing information from attribute features and behavioral features, four classical machine learning algorithms were tested on LPIH2V, including SVM, RF, Gaussian NB (NB), and Logistic Regression (LR). As we can see in Figure 4, the proposed model achieved the best results according to precision, recall, SPE, accuracy, MCC, F1, and

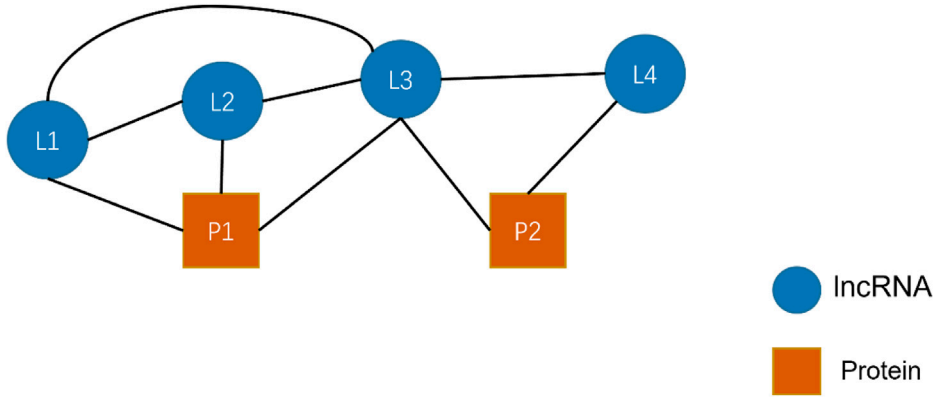


FIGURE 7
An example of a simple heterogeneous network of lncRNA-protein interactions.

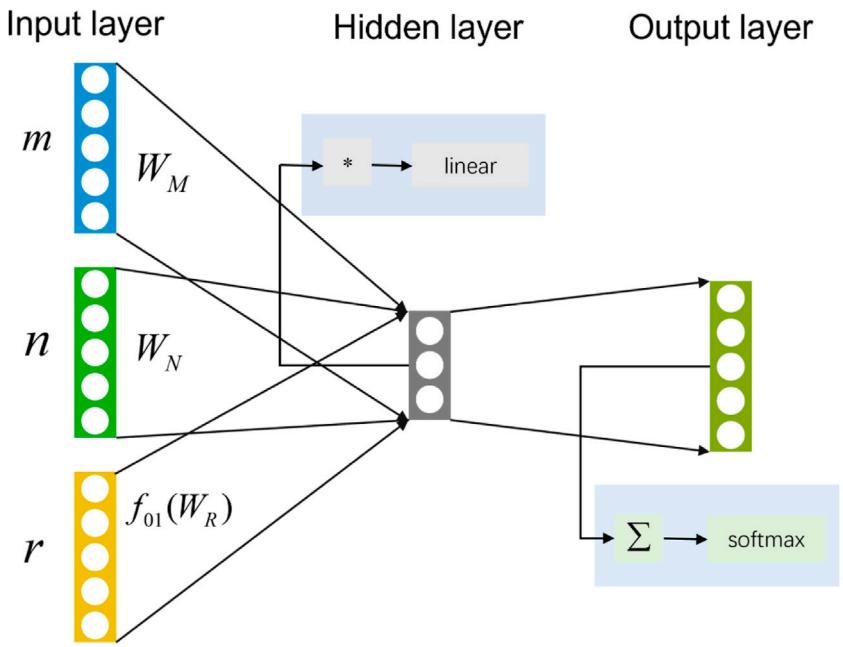


FIGURE 8
HIN2Vec NN model includes input layer, hidden layer and output layer.

AUC. These results indicate that the SVM classifier is a good fit for the proposed model and can predict LPIs effectively.

## 2.5 Comparison with other models

In addition, to verify the validity and stability of LPIH2V, we compared LPIH2V with other computational methods include CF (Sarwar et al., 2001), RWR (Köhler et al., 2008), LPBNI (Ge et al., 2016), SFPEL-LPI (Zhang et al., 2018a), LPIHN (Li et al., 2015), LPLNP (Zhang et al., 2018b), RPI-SE (Yi et al., 2020b), IPMiner (Pan et al., 2016), and LncPNet (Zhao et al., 2021a). The CF model uses cosine similarity and Pearson correlation similarity to calculate the correlation between the two users based on collaborative filtering recommendation algorithms. To define the similarity of nodes in networks, the RWR model uses random walk with restart to calculate the distance between nodes in the network. LPBNI used information on known RNA-protein interactions to build a lncRNA-protein bipartite network. Following this, a propagation method was performed in the bipartite network to score and rank the candidate proteins for each lncRNA. The SFPEL-LPI model extracts lncRNA and protein sequence features using linear neighborhood similarity, then predicts LPIs using the feature projection ensemble learning method. The LPIHN model constructs a heterogeneous network by linking the lncRNA-lncRNA similarity network, protein-protein interaction network, and lncRNA-protein interaction network. LncRNAs and proteins were scored in heterogeneous networks using random walk with restart. LPLNP calculates linear neighborhood similarities in feature spaces and transfers them to interaction space to predict unknown

interactions through the label propagation process. RPI-SE integrates three independent models, including XGBoost, SVM, and ExtraTree, to predict ncRNA-protein interactions using sequence information and fully exploit the potent properties of RNA and protein employing PWM and K-mer sparse matrices. The IPMiner model extracts the 3-mers and 4-mers from protein and RNA sequences, respectively, and then uses a stacked autoencoder to derive high-level features from the retrieved RNA and protein sequence features, respectively. The above stacked autoencoders were fine-tuning using the label information from the training data to change the index of the network. The LncPNet model constructs a heterogeneous network by calculating the similarity between lncRNA-lncRNA and protein-protein, then extracts features using metapath2vec. Table 2 shows that our model achieved the highest AUC value of 0.976. LPIH2V performed the best on other metrics because we used the HIN2Vec method to extract features that capture rich relational semantics and details of the network structure to learn the representation of nodes in HIN. In addition, the model learns the representation of meta-paths for meta-path analysis.

# 3 Materials and methods

## 3.1 Datasets

In this experiment, lncRNA sequence data were obtained from NONCODE (Zhao et al., 2021b), protein sequence data from UniProt (UniProt Consortium, 2021), and known lncRNA-protein interaction data from NPInter (Teng et al., 2020). UniProt provides a complete compilation of all known protein sequence data and links it to a summary of validated experience or computational predictions of the protein's functional information. To avoid data redundancy, the UniProt database consolidated the same protein stored in different databases into one and gives it a unique and specific identifier. The number of entries contained in UniProt had grown to over 65 million records. NPInter incorporated 600,000 new experimentally determined ncRNA interactions, primarily by hand mining the literature and processing high-throughput sequencing data to collect interaction data. Researchers integrated data from different sources and eliminated redundant entries. NONCODE was a comprehensive database of assembled and illustrated non-coding RNAs, typically animal lncRNAs. NONCODE not only provides access to the names and NONCODE IDs of commonly used lncRNAs, but some lncRNAs also support other database names for conducting searches. The number of lncRNAs has rapidly increased to 644,510, of which 173,112 were human lncRNAs. First, the LPIs information extracted from NPInter were limited to "Homo," "lncRNA" and "protein" respectively, and filtered for operationally proven human LPIs information. The resulting lncRNA ID and protein ID were then plotted as NONCODE ID and UniProt IDs, respectively. Finally, invalid lncRNA and protein information was deleted from the sequence information, leaving 4,578 pairs of known LPIs information.

## 3.2 Overview of methods

In this paper, we propose a prediction framework based on fusing attribute features and behavioral features, called LPIH2V. As shown in Figure 5, LPIH2V is divided into three parts. In the first part, we compute Jaccard similarity and BLAST similarity for the lncRNAs and proteins, respectively, given the lncRNA-lncRNA and protein-protein sequence information in the dataset, as well as the known LPIs information. In the second part, we construct two heterogeneous networks of lncRNA-protein and use the HIN2Vec method to extract behavioral features of the nodes in the network based on meta-paths. The features derived from the two heterogeneous networks were integrated. In the third part, we predict the LPIs by multiple classifiers. The feature vectors are fed into the classifiers to produce predictions. It is worth noting that many models ignore the semantic relationships between different node types. We fully learn the structural information of the nodes embedded in the network by constructing a heterogeneous network and HIN2Vec method. The feature information obtained through learning is fused to retain the sequence information of the nodes.

## 3.3 Sequence similarity calculation

The calculation of similarity is an important part of gene association prediction. Basic Local Alignment Search Tool (BLAST) (Fu et al., 2018) is a dual sequence local alignment algorithm proposed by Altschul et al., in 1990. As a set of analytical tools for similarity comparisons in protein databases or DNA databases, which contains several separate procedures. These programs are defined depending on the query and the database. For example, if the query is for nucleic acid and the query database is a nucleic acid sequence database, then the blastn program should be selected. The basic concept of BLAST is to increase the speed of comparison by generating fewer but better-quality enhancement points. In this paper, we performed BLAST to obtain the similarity between every two lncRNAs and every two proteins. Jaccard similarity (Lu et al., 2018) is a popular approximation measure used to calculate the similarity between two objects. Jaccard similarity can be used to find the similarity between two asymmetric binomial vectors or to find the similarity between two sets. The Jaccard coefficient is often used between sequence-order insensitive texts. The higher the value of the Jaccard coefficient, the higher the similarity of the sample. We calculated lncRNA-lncRNA similarity and protein-protein similarity by using Jaccard similarity principle. The Jaccard coefficient is defined as the size of the intersection of the sample set divided by the size of the merge set. For example, $L_i$ and $L_j$ are datasets of two lncRNAs. The Jaccard similarity between any two sets of lncRNAs is computed as follows:

$$J(L_i, L_j) = \frac{|L_i \cap L_j|}{|L_i \cup L_j|}$$

Jacquard similarity of proteins is calculated in the same way as lncRNA.

## 3.4 Heterogeneous network construction

In recent years, the precise extraction of personalization factors from data has become increasingly difficult due to the rapid growth of Big Data. Heterogeneous information networks (HINs) can represent the attribute information and structure information completely, which not only assumes two nodes are related but also distinguishes between different relationships among nodes and retains more contextual

information by learning the relationship vectors jointly. In detail, the training data is represented using a bipartite graph = $(m, n, r, j/b)$, where $m$ and $n$ are two nodes in the same category, and the $r$ is the link between two nodes. The Jaccard and BLAST values of the two nodes are indicated by $j$ and $b$ respectively. For lncRNA, samples with Jaccard similarity greater than 0.5 and BLAST similarity less than 0.001 were selected. On the protein side, samples with Jaccard similarity greater than 0 and BLAST similarity less than 0.01 were selected. The heterogeneous networks are made up of these two similarity networks and known LPIs.

## 3.5 Heterogenous network embedding

To capture the rich semantics embedded in heterogeneous networks by exploiting the different types of relationships among nodes, we employed the HIN2Vec method. Unlike metapath2vec, which follows a given meta-path pattern, HIN2Vec selects nodes randomly. In particular, as illustrated in Figure 6, the HIN2Vec framework consists of two phases: training data preparation and representation learning. The HIN2Vec matching target relation data is generated based on random wandering and negative sampling to represent the learning for the nodes in the network. To learn the node vectors and relationships between pairs of nodes, we maximized the likelihood of jointly predicting the relationships between lncRNAs and proteins.

Figure 7 shows a simple HIN consisting of four lncRNAs (L) L1, L2, L3, L4, and two proteins (P) P1 and P2. In this figure, edges represent the relationship that exists between nodes. We restricted the meta-path length to be at most 2. Giving R denotes a collection of targets that includes all relations between nodes. Thus, R = {L − L, L − P, P − L, L − L − L, L − L − P, L − P − L, P − L − L, P − L− P}, L1 to P1 as {L − P, L − L − P}, entering training data is {x: $L_1$, y: $P_1$, output: [0, 1, 0, 0, 1, 0, 0, 0]}. From the above example, it is straightforward to see that the HIN2Vec method traverses all the meta-paths in the heterogeneous network and then learns each potential feature vector.

To reduce the time required to traverse all the meta-paths, the HIN2Vec based on a neural network (HIN2Vec NN) will predict the probability of the relationship between two nodes in advance. A three-layer feedforward neural network model adopted by HIN2Vec as a binary classifier. As shown in Figure 8, input the two nodes $m$, $n$ and the relationship $r$ between $m$ and $n$ into the binary classifier. The three parameters at the input layer are mapped into three one-hot feature vectors $\vec{m}, \vec{n}, \vec{r}$, which mapped into potential vectors $W_M^{'}\vec{m}$, $W_N^{'}\vec{n}$ and $W_R^{'}\vec{r}$. In the hidden layer, considering that the semantics of nodes and relations are different, and therefore their representation space should not be the same, which applied a regularization function $f_{01}(.)$ to limit the potential vector for $r$ to be between 0 and 1, then use the Hadamard function and apply Identity function to activate the three vectors. The output layer uses the Summation function as input, summing up three values. To realize logistic classification, the Sigmoid function is used as the activation function.

HIN2Vec uses backpropagation training algorithms and stochastic gradient descent to optimize the model. Assuming a training set $D(m, n, r, L(m, n, r))$, which include $m$, $n$ and the relationship $r$ between $m$ and $n$. We denote by $L$ if there is any interaction between nodes $m$ and $n$, expressed as 0 or 1. A maximization objective function $O$ is used to adjust the weights of each parameter $W_M$ , $W_N$ and $W_R$ , and then maximize $\log O$. Given an objective function $O$ and a derivation of $\log O$ such that:

$$\boldsymbol{O \propto \log O = \Sigma_{m,n,r\in D} \log O_{m,n,r}(m,n,r)}$$

In particular, for an input of training data $(m, n, r, L(m, n, r))$, when $L(m, n, r)$ is 1, $O_{m,n,r}(m, n, r)$ intends to maximize $P(r \mid m, n)$ , Or else, $O_{m,n,r}(m, n, r)$ intends to minimize $P(r \mid m, n)$. Here $B(m, n, r)$, a binary value, indicates whether m and n have r. Therefore, $O_{m,n,r}(m, n, r)$, $\log O_{m,n,r}(m, n, r)$ and $P(r \mid m, n)$ as follows:

$$\boldsymbol{O_{m,n,r}(m,n,r) = \begin{cases} P(r \mid m,n) & \text{if } L(m,n,r) = 1 \\ 1 - P(r \mid m,n) & \text{if } L(m,n,r) = 0 \end{cases}}$$

$$\boldsymbol{\log O_{m,n,r}(m,n,r) = B(m,n,r)\log P(r \mid m,n) + [1 - B(m,n,r)]\log[1 - P(r \mid m,n)]}$$

$$\boldsymbol{P(r \mid m,n) = sigmoid\left(\sum W_M^{'}\vec{m} \odot W_N^{'}\vec{n} \odot f_{01}\left(W_R^{'}\vec{r}\right)\right)}$$

Finally, for each training data entry, the algorithm reverses weights in $W_M^{'}\vec{m}$ , $W_N^{'}\vec{n}$ and $W_R^{'}\vec{r}$ based on the gradients of $\log O_{m,n,r}(m, n, r)$ differentiating with respect to $W_M^{'}\vec{m}$, $W_N^{'}\vec{n}$ and $W_R^{'}\vec{r}$, respectively, as follows:

$$\boldsymbol{W_M^{'}\vec{m}: = W_M^{'}\vec{m} + \frac{d\log OF_{m,n,r}(m,n,r)}{dW_M^{'}\vec{m}}}$$

$$\boldsymbol{W_N^{'}\vec{n}: = W_N^{'}\vec{n} + \frac{d\log OF_{m,n,r}(m,n,r)}{dW_N^{'}\vec{n}}}$$

$$\boldsymbol{W_R^{'}\vec{r}: = W_R^{'}\vec{r} + \frac{d\log OF_{m,n,r}(m,n,r)}{dW_R^{'}\vec{r}}}$$

# 4 Conclusion

With the rapid development of computer performance, traditional wet experiments have limitations compared to computational methods. The cost of experiments can be greatly reduced and experimental time saved. The computational method not only reduces the interference of other factors in the experiment but also provides specific ideas for biological experiments. Most lncRNAs exert their effects by interacting with proteins, so the prediction of LPIs is critical for the proper functioning of lncRNAs.

The purpose of the study was to propose a model for predicting lncRNA-protein interactions that combines sequence characteristics and behavioral properties to fully exploit node information, and we constructed heterogeneous networks using similarity principles. LPIH2V builds heterogeneous networks through the principle of sequence similarity. HIN2Vec is used to learn the behavioral features of the nodes in the network. This takes into account both the sequence characteristics and the behavioral characteristics of the nodes. We also used a sparse automatic encoder to upgrade the eigenvectors, converting vectors from 64 dimensions to 128 dimensions for comparison, and using different classifiers to predict vectors for elevation. The results show that the recall score of 0.98 using stochastic forest classifiers indicates that the model can identify positive and negative samples more accurately by raising the dimensions. To verify the robustness and reliability of the proposed method, we used multiple classifiers to compare predictions and found that SVM worked best. An additional 5-fold comparison test was performed to test the accuracy of the prediction model. Then used the best SVM classifier to compare with other advanced methods, and

finally concluded that our approach has superiorities in predicting LPIs. Undoubtedly, our proposed model has the potential to provide a useful guide for biomedical research in LPIs prediction. With advances in technology, more efficient feature extraction strategies and incorporation of other information into the model could lead to greater accuracy and improved performance, such as disease and miRNA.

However, LPIH2V has some shortcomings. Learning node features using HIN2V can be time-consuming. The number of datasets also has an impact on the model results. The following will look for ways to optimize the learning time of the model and to further expand the number of datasets.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

## Author contributions

M-MW, C-QY, L-PL, Z-HY, Z-HR, and Y-JG: conceptualization, methodology, software, validation, resources, and data curation. All authors contributed to manuscript revision, read, and approved the submitted version.

## Funding

This work was supported by the NSFC Program, under Grant 62273284, 62072378, and 62002297, and the National Natural Science Foundation of China (No. 62172338).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33 (8), 831–838. doi:10.1038/nbt.3300

Allou, L., Balzano, S., Magg, A., Quinodoz, M., Royer-Bertrand, B., Schöpflin, R., et al. (2021). Non-coding deletions identify Maenli lncRNA as a limb-specific En1 regulator. *Nature* 592 (7852), 93–98. doi:10.1038/s41586-021-03208-9

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2017). Long non-coding RNAs and complex diseases: From experimental results to computational models. *Briefings Bioinforma.* 18 (4), 558–576. doi:10.1093/bib/bbw060

Congrains, A., Kamide, K., Oguro, R., Yasuda, O., Miyata, K., Yamamoto, E., et al. (2012). Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis* 220 (2), 449–455. doi:10.1016/j.atherosclerosis.2011.11.017

Cook, E. H., Jr, and Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455 (7215), 919–923. doi:10.1038/nature07458

Cui, Z., Ren, S., Lu, J., Wang, F., Xu, W., Sun, Y., et al. (2013). The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor. *Urologic Oncol. Seminars Orig. Investigations*, 1117–1123. doi:10.1016/j.urolonc.2011.11.030

Deng, L., Wang, J., Xiao, Y., Wang, Z., and Liu, H. (2018). Accurate prediction of protein-lncRNA interactions by diffusion and HeteSim features across heterogeneous network. *BMC Bioinforma.* 19 (1), 370–411. doi:10.1186/s12859-018-2390-0

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12 (12), 861–874. doi:10.1038/nrg3074

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. *Bioinformatics* 34 (9), 1529–1537. doi:10.1093/bioinformatics/btx794

Fu, T.-y., Lee, W.-C., and Lei, Z. (2017). "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 1797–1806.

Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics, proteomics Bioinforma.* 14 (1), 62–71. doi:10.1016/j.gpb.2016.01.004

Guan, Y.-J., Yu, C.-Q., Li, L.-P., You, Z.-H., Ren, Z.-H., Pan, J., et al. (2022). Bnemdi: A novel MicroRNA–drug interaction prediction model based on multi-source information with a large-scale biological network. *Front. Genet.* 13, 919264. doi:10.3389/fgene.2022.919264

Guo, L.-X., You, Z.-H., Wang, L., Yu, C.-Q., Zhao, B.-W., Ren, Z.-H., et al. (2022). A novel circRNA-miRNA association prediction model based on structural deep neural network embedding. *Briefings Bioinforma.* 23 (5), bbac391. doi:10.1093/bib/bbac391

Hu, W., Wang, Y., Fang, Z., He, W., and Li, S. (2021). Integrated characterization of lncRNA-immune interactions in prostate cancer. *Front. Cell Dev. Biol.* 9, 641891. doi:10.3389/fcell.2021.641891

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82 (4), 949–958. doi:10.1016/j.ajhg.2008.02.013

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting long noncoding RNA and protein interactions using heterogeneous network model. *BioMed Res. Int.* 2015, 671950. doi:10.1155/2015/671950

Li, Y.-C., You, Z.-H., Yu, C.-Q., Wang, L., Wong, L., Hu, L., et al. (2022). Ppaedti: Personalized propagation auto-encoder model for predicting drug-target interactions. *IEEE J. Biomed. Health Inf.* 27, 573–582. doi:10.1109/JBHI.2022.3217433

Lu, C., Yang, M., Luo, F., Wu, F.-X., Li, M., Pan, Y., et al. (2018). Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics* 34 (19), 3357–3364. doi:10.1093/bioinformatics/bty327

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC genomics* 14 (1), 651–710. doi:10.1186/1471-2164-14-651

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinforma.* 12 (1), 489–511. doi:10.1186/1471-2105-12-489

Ng, S.-Y., Lin, L., Soh, B. S., and Stanton, L. W. (2013). Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet.* 29 (8), 461–468. doi:10.1016/j.tig.2013.03.002

Pan, X., Fan, Y.-X., Yan, J., and Shen, H.-B. (2016). IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC genomics* 17 (1), 582–614. doi:10.1186/s12864-016-2931-8

Pasmant, E., Sabbagh, A., Vidaud, M., and Bièche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25 (2), 444–448. doi:10.1096/fj.10-172452

Peng, J. Y., Cai, D. K., Zeng, R. L., Zhang, C. Y., Li, G. C., Chen, S. F., et al. (2022). Upregulation of superenhancer-driven LncRNA FASRL by USF1 promotes de novo fatty acid biosynthesis to exacerbate hepatocellular carcinoma. *Adv. Sci.* 10, 2204711. doi:10.1002/advs.202204711

Prensner, J. R., and Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer Discov.* 1 (5), 391–407. doi:10.1158/2159-8290.CD-11-0209

Rathinasamy, B., and Velmurugan, B. K. (2018). Role of lncRNAs in the cancer development and progression and their regulation by various phytochemicals. *Biomed. Pharmacother.* 102, 242–248. doi:10.1016/j.biopha.2018.03.077

Ren, Z.-H., You, Z.-H., Yu, C.-Q., Li, L.-P., Guan, Y.-J., Guo, L.-X., et al. (2022). A biomedical knowledge graph-based method for drug–drug interactions prediction through combining local and global features with deep neural networks. *Briefings Bioinforma.* 23 (5), bbac363. doi:10.1093/bib/bbac363

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web, 2001, 285–295.

Teng, X., Chen, X., Xue, H., Tang, Y., Zhang, P., Kang, Q., et al. (2020). NPInter v4. 0: An integrated database of ncRNA interactions. *Nucleic acids Res.* 48 (D1), D160–D165. doi:10.1093/nar/gkz969

UniProt Consortium (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucleic acids Res.* 41 (D1), D246–D251. doi:10.1093/nar/gks915

Wang, L., You, Z.-H., Huang, D.-S., and Zhou, F. (2018). Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17 (3), 972–980. doi:10.1109/TCBB.2018.2874267

Wang, Y., Chen, X., Liu, Z.-P., Huang, Q., Wang, Y., Xu, D., et al. (2013). De novo prediction of RNA–protein interactions from sequence information. *Mol. Biosyst.* 9 (1), 133–142. doi:10.1039/c2mb25292a

Yang, J., Li, A., Ge, M., and Wang, M. (2016). Relevance search for predicting lncRNA–protein interactions based on heterogeneous network. *Neurocomputing* 206, 81–88. doi:10.1016/j.neucom.2015.11.109

Yi, H.-C., You, Z.-H., Cheng, L., Zhou, X., Jiang, T.-H., Li, X., et al. (2020). Learning distributed representations of RNA and protein sequences and its application for predicting lncRNA-protein interactions. *Comput. Struct. Biotechnol. J.* 18, 20–26. doi:10.1016/j.csbj.2019.11.004

Yi, H.-C., You, Z.-H., Huang, D.-S., and Kwoh, C. K. (2022). Graph representation learning in bioinformatics: Trends, methods and applications. *Briefings Bioinforma.* 23 (1), bbab340. doi:10.1093/bib/bbab340

Yi, H.-C., You, Z.-H., Wang, M.-N., Guo, Z.-H., Wang, Y.-B., Zhou, J.-R., et al. (2020). A stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. *BMC Bioinforma.* 21 (1), 1–10.

Yu, C.-Q., Wang, X.-F., Li, L.-P., You, Z.-H., Huang, W.-Z., Li, Y.-C., et al. (2022). Sgcncmi: A new model combining multi-modal information to predict circRNA-related miRNAs, diseases and genes. *Biology* 11 (9), 1350. doi:10.3390/biology11091350

Zhang, S.-W., and Fan, X.-N. (2017). Computational methods for predicting ncRNA-protein interactions. *Med. Chem.* 13 (6), 515–525. doi:10.2174/1573406413666170510102405

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi:10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* 14 (12), e1006616. doi:10.1371/journal.pcbi.1006616

Zhao, G., Li, P., Qiao, X., Han, X., and Liu, Z.-P. (2021). Predicting lncRNA–protein interactions by heterogenous network embedding. *Front. Genet.* 12, 814073. doi:10.3389/fgene.2021.814073

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., et al. (2021). NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic acids Res.* 49 (D1), D165–D171. doi:10.1093/nar/gkaa1046