



OPEN ACCESS

EDITED BY

Leyi Wei,
Shandong University, China

REVIEWED BY

Zhi-Ping Liu,
Shandong University, China
Hongqiang Wang,
Hefei Institutes of Physical Science (CAS),
China

*CORRESPONDENCE

Han Zhang,
✉ zhanghan@nankai.edu.cn

SPECIALTY SECTION

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

RECEIVED 03 January 2023

ACCEPTED 30 January 2023

PUBLISHED 09 February 2023

CITATION

Ma Y, Zhang H, Jin C and Kang C (2023),
Predicting lncRNA-protein interactions
with bipartite graph embedding and deep
graph neural networks.
Front. Genet. 14:1136672.
doi: 10.3389/fgene.2023.1136672

COPYRIGHT

© 2023 Ma, Zhang, Jin and Kang. This is an
open-access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Predicting lncRNA-protein interactions with bipartite graph embedding and deep graph neural networks

Yuzhou Ma¹, Han Zhang^{1*}, Chen Jin² and Chuanze Kang¹

¹College of Artificial Intelligence, Nankai University, Tianjin, China, ²College of Computer Science, Nankai University, Tianjin, China

Background: Long non-coding RNAs (lncRNAs) play crucial roles in numerous biological processes. Investigation of the lncRNA-protein interaction contributes to discovering the undetected molecular functions of lncRNAs. In recent years, increasingly computational approaches have substituted the traditional time-consuming experiments utilized to crack the possible unknown associations. However, significant explorations of the heterogeneity in association prediction between lncRNA and protein are inadequate. It remains challenging to integrate the heterogeneity of lncRNA-protein interactions with graph neural network algorithms.

Methods: In this paper, we constructed a deep architecture based on GNN called BiHo-GNN, which is the first to integrate the properties of homogeneous with heterogeneous networks through bipartite graph embedding. Different from previous research, BiHo-GNN can capture the mechanism of molecular association by the data encoder of heterogeneous networks. Meanwhile, we design the process of mutual optimization between homogeneous and heterogeneous networks, which can promote the robustness of BiHo-GNN.

Results: We collected four datasets for predicting lncRNA-protein interaction and compared the performance of current prediction models on benchmarking dataset. In comparison with the performance of other models, BiHo-GNN outperforms existing bipartite graph-based methods.

Conclusion: Our BiHo-GNN integrates the bipartite graph with homogeneous graph networks. Based on this model structure, the lncRNA-protein interactions and potential associations can be predicted and discovered accurately.

KEYWORDS

lncRNA-protein interaction, graph neural network, bipartite graph embedding, heterogeneous graph, link prediction

1 Introduction

Long non-coding RNAs (lncRNAs) are RNAs with a length of no less than 200 nucleotides that do not encode proteins (Schaukowitch and Kim, 2014). Recently investigators have documented that lncRNAs play a critical role in various pathological and biological processes. Their misimpression can stimulate a series of lesions in humans, such as colon cancer (Pibouin et al., 2002), tumor initiation (Yang et al., 2014), nasopharyngeal carcinoma cell invasion (Wang et al., 2020a), and breast cancer (Wang et al., 2020b). For example, ncRNA miR-106b-5p assists metastasis by suppressing the key gene which links to breast cancer and activating Rho/ROCK1 pathway (Wang et al., 2020b). lncRNA ZNRD1-AS1 promotes the metastasis of

nasopharyngeal carcinoma cells by regulating the miR-335–ROCK1 axis (Wang et al., 2020a). LncRNA LINC00337 promotes tumor angiogenesis, which can lead to colorectal cancer (Xu et al., 2021). Previous studies have shown that only a small fraction of the human genome is protein-coding genes (1.5%). In other words, the function of most human gene sequences that do not encode proteins is anonymous (Chen and Yan, 2013). Accordingly, the lncRNA-protein interactions are essential in analyzing the molecular function of lncRNAs.

Traditional methods can experimentally verify lncRNA-protein interactions. Previous experiments such as PAR-CLIP (Hafner et al., 2010) are time-consuming and cost-effective to predict, then computational approaches have been widely applied to lncRNA-protein interactions, which are based on machine learning and deep learning.

Graph neural network (GNN) is an extension method of a traditional neural network, which transforms the relationship between nodes into structured data and then completes forward propagation in graph domain (Scarselli et al., 2008). With the model iterations in GNN, existing research focuses on convolutions in graph data mining. Graph convolutional network is a variant of convolutional neural networks, which can operate directly on graph-structured data (Kipf and Welling, 2016). GraphSAGE (Hamilton et al., 2017) unifies the information of nodes through its neighbor node feature aggregation.

Link prediction using deep learning methods is commonly prescribed for disease-genes (Chen et al., 2018), miRNA-lncRNA (Huang et al., 2018) and many other fields. Previous study demonstrated that GNN had become the key instrument in link prediction (Zhang and Chen, 2018).

The following approaches achieve relevant results in predicting the interactions on lncRNA-protein. RPISeq method is a classifier for predicting lncRNA-protein interactions with two variants: Support Vector Machine (SVM) and Random Forest (Muppirala et al., 2011). NPI method integrated many advanced deep learning correlation models such as SEAL framework (Zhang and Chen, 2018) to this task (Shen et al., 2021). LPI-deepGBDT utilized gradient boosting decision trees for lncRNA-protein interaction identification (Zhou et al., 2021). LPIGAC implemented autoencoders on two graphs and trained these embedding collaboratively (Jin et al., 2021), however, their work was mainly based on the homogeneous graph, which led to the lack of the capability of heterogeneous features in the framework, including dependencies between heterogeneous nodes. The model results will be limited by some misjudgments, such as connecting homogenous nodes (protein-protein) and time-costing large-scale matrix calculations.

In recent years various methods based on graph embedding have been proposed. Li et al. (2015) proposed a heterogeneous network based on the protein-protein interaction. LPLNP was designed based on linear neighborhood propagation, which transfers the graph similarity into the network embedding (Zhang et al., 2018). LncPNet was proposed based on embedding the heterogeneous network to learn the low-dimensional potential node representations (Zhao et al., 2022).

Major of the biomedical interaction graph is not homogeneous. For instance, lncRNA-protein, disease-genes, hence bipartite graph embedding is fundamental to predict the potential edge in the bipartite graph. Before bipartite embedding was proposed, many studies contributed to the work of homogeneous graph embedding (Cui et al., 2018; Cai et al., 2018). Although these methods work well, they are not suitable for embedding the construction of bipartite graphs.

To remedy the problem, increasing explorations on heterogeneous graphs have been proposed. The reconstruction-based method with graph convolutional matrix completion works pretty well on standard collaborative filtering benchmarks (Berg et al., 2017).

The structure of the bipartite graph network has been iterated many times. Metapath2vec applied scalable node representation in heterogeneous networks (Dong et al., 2017). BiNE proposed a random walk generator to generate representation vectors and also combined explicit relations and implicit relations (Gao et al., 2020). BiRank proposed a method to integrate bipartite graph structure and node representation (He et al., 2016). BiGI integrated the embedding of two node types into local-global representation, which also proposed the bipartite embedding applied to deep learning (Cao et al., 2021).

Gilmer captured node representations by using the features of neighboring nodes to train the neural network (Gilmer et al., 2017). DMGI (Park et al., 2020) utilized the infomax objective to heterogeneous graphs. It splits the heterogeneous graph into homogeneous graphs and applies the infomax objective to this task.

In this study, we integrate homogeneous networks and heterogeneous networks to construct mutual optimization model through bipartite graph embedding. The heterogeneous features combine the association information to obtain the bipartite graph features of each node. The representations are input into the homogeneous network established based on GraphSAGE and matched with interaction to update the bipartite features. Finally, the bipartite embedding is input into the logistic regression classifier to calculate the link categories. In summary, the main advantages of BiHo-GNN are as follows:

1. BiHo-GNN can capture the feature of the lncRNA-protein interactions and distinguish the disparate nodes, which can lower the negative effect of lncRNA-protein homogenization on link prediction.
2. BiHo-GNN combines the advantages of heterogeneous and homogeneous networks, which uses the heterogeneous network to generate bipartite graph features. The bipartite embedding is a feature composed of two types of node prototype representations.
3. Our homogeneous network based on GraphSAGE can iterate bipartite embedding from heterogeneous network to form a feedback process.

2 Materials and methods

2.1 Datasets

There are four datasets collected in this study. These datasets are NPInter2.0 (Yuan et al., 2014), NPInter3.0_H (Hao et al., 2016), NPInter3.0_M (Hao et al., 2016), RPI2241 (Muppirala et al., 2011).

NPInter 2.0 database includes 10,412 experimentally demonstrated functional lncRNA-protein interactions, containing 4,636 RNAs and 449 proteins, which were extracted from the UniProt database (UniProt Consortium, 2014) and the NONCODE database (Bu et al., 2012). NPInter 3.0 is an upgraded dataset of ncRNA-sequence interactions. We only use two pieces of data for processing. NPInter3.0_H is from the *homo sapiens* specie part, composed of 7,317 lncRNA-protein interaction pairs, 1874 RNAs, and 118 proteins. NPInter3.0_M is the musculus species subset of NPInter 3.0, involving 1847 experimentally verified lncRNA-protein

TABLE 1 Introduction of four datasets used in this paper.

Datasets	Species	Interactions	RNAs	Proteins
NPInter2.0	—	10,412	4,636	449
NPInter3.0_H	<i>Homo sapiens</i>	7,317	1874	118
NPInter3.0_M	<i>Mus musculus</i>	1847	1939	60
RPI2241	—	2,241	842	2,043

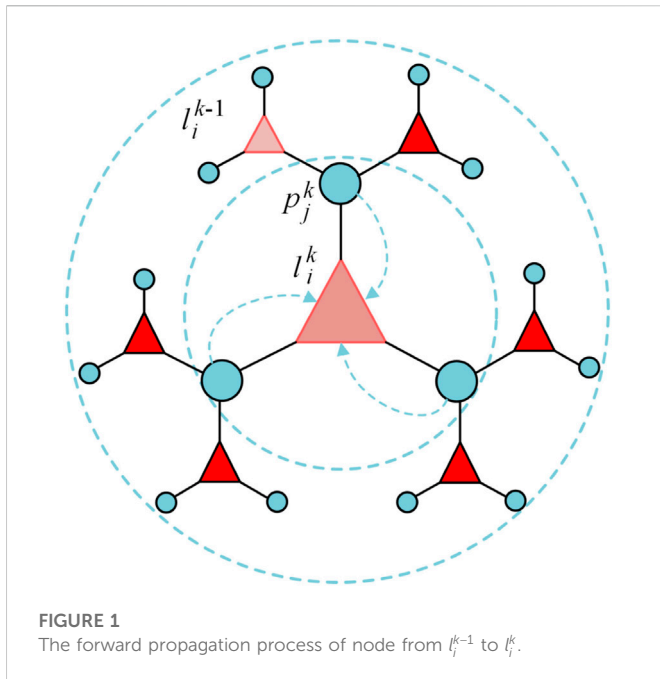


FIGURE 1
The forward propagation process of node from l_i^{k-1} to l_i^k .

interactions. These interactions contain 1939 RNAs and 60 proteins. RPI2241 and the above datasets differ in data acquisition. RPI2241 is acquired based on 3D atom coordinates and algorithm inference (Lewis et al., 2010), containing 2,241 interactions, 838 RNAs, and 2040 proteins. These four datasets have an exact number of each type of node. At the same time, the bipartite embedding model in data preprocessing may lead to filtering out a few low-frequency nodes. The specific number of each item is shown in Table 1. Since these four datasets only marked positive samples, we randomly selected negative samples with the same number of positive samples in the data sets that have not been verified to be associated.

2.2 Background

Let $G = (L, P, E)$ be the bipartite graph, where L and P are the set of RNAs and proteins, with E is the edges between RNA and protein sets. It is obvious that RNA and protein nodes are heterogeneous. Representation vectors are instrumental in their forward propagation in graph neural networks. For this bipartite graph with vertex sets $L = \{l_1, \dots, l_r\}$, $P = \{p_1, \dots, p_s\}$, where r and s denote the number of lncRNAs and proteins. Edges $E \subseteq L \times P$, bi-adjacency matrix $A \in \{0,1\}^{r \times s}$, where $A_{i,j} = 1$ when the RNA node $l_i \in L$ and the protein node $p_j \in P$ interact, and $A_{i,j} = 0$ when no interaction occurs.

Bipartite graph embedding maps graph data into a feature matrix based on a sample Bi-GNN by mutual iteration between molecule nodes.

2.3 Bipartite graph embedding

The bipartite graph embedding, which was proposed by Cao (Cao et al., 2021). Inspired by this work, we design the feature of lncRNA and protein nodes. In this part, we use Bi-GNN as the bipartite graph encoder to generate the molecular node representations and take these representations into the following network frame. For clarification of the notations, we use l_i and p_j to stand for the representation of RNAs and proteins node, respectively.

The key problem in the prediction of lncRNA-protein interaction is the utilization of neighbor nodes attribution and utilize the homogeneity of two molecular nodes efficiently. Bipartite graph encoder can learn each node feature from two-hop neighbors interaction. Taking l_i^{k-1} for example as illustration shown in Figure 1, each node forward propagation in k -th layer aggregates the embedding of two-hop neighbor nodes. The propagation of l_i^k, p_j^k is represented l_i^{k-1} via a Bi-GNN encoder:

$$p_j^k = \text{ReLU}(\text{Linear}(\hat{A}l_i^{k-1}W^{k-1}; l_i \in L)) \quad (1)$$

Protein nodes p_j^k are represented by upper RNA nodes l_i^{k-1}, l_i^k can be obtained by the Bi-GNN encoder based on p_j^{k-1} .

$$l_i^k = \text{ReLU}(\text{Linear}(\hat{A}p_j^{k-1}W^{k-1}; p_j \in P)) \quad (2)$$

where $\hat{A} = D^{-1/2}(A + I_n)D^{1/2}$, A is the adjacency matrix of the bipartite graph, D is the diagonal degree matrix of $A + I_n$, and W denotes the weights of the GNN encoder. The final RNA node embedding l_n^k is the cascading matrix of l_n^k and l_n^{k-1} as follows:

$$l_n^k = \text{Linear}\left(\begin{bmatrix} l_n^k & l_n^{k-1} \end{bmatrix}\right) \quad (3)$$

2.4 Network structure

Depending on the generated bipartite node embedding, in this section, we proposed a homogeneous network to learn RNA and protein embedding, which can capture the homogenous properties of bipartite embedding.

In homogenous network, both RNA and protein share the same node type and node representation. We integrate two types of nodes embedding as the new input $H \in \mathbb{R}^{(r+s) \times d}$ of the homogenous network via a simple composition operation:

$$H = [L \ P]^T \quad (4)$$

For each node type, we construct a bipartite node feature matrix under the interaction order, and the new node attribution X can be decoded with H_l and H_p .

$$X = H_l \odot H_p \quad (5)$$

where H_l and H_p are the node feature order of RNA and protein.

In this paper, we applied GraphSAGE (Hamilton et al., 2017) to aggregate information from the neighborhood representation, which can extract the bipartite embedding for previously neglected nodes,

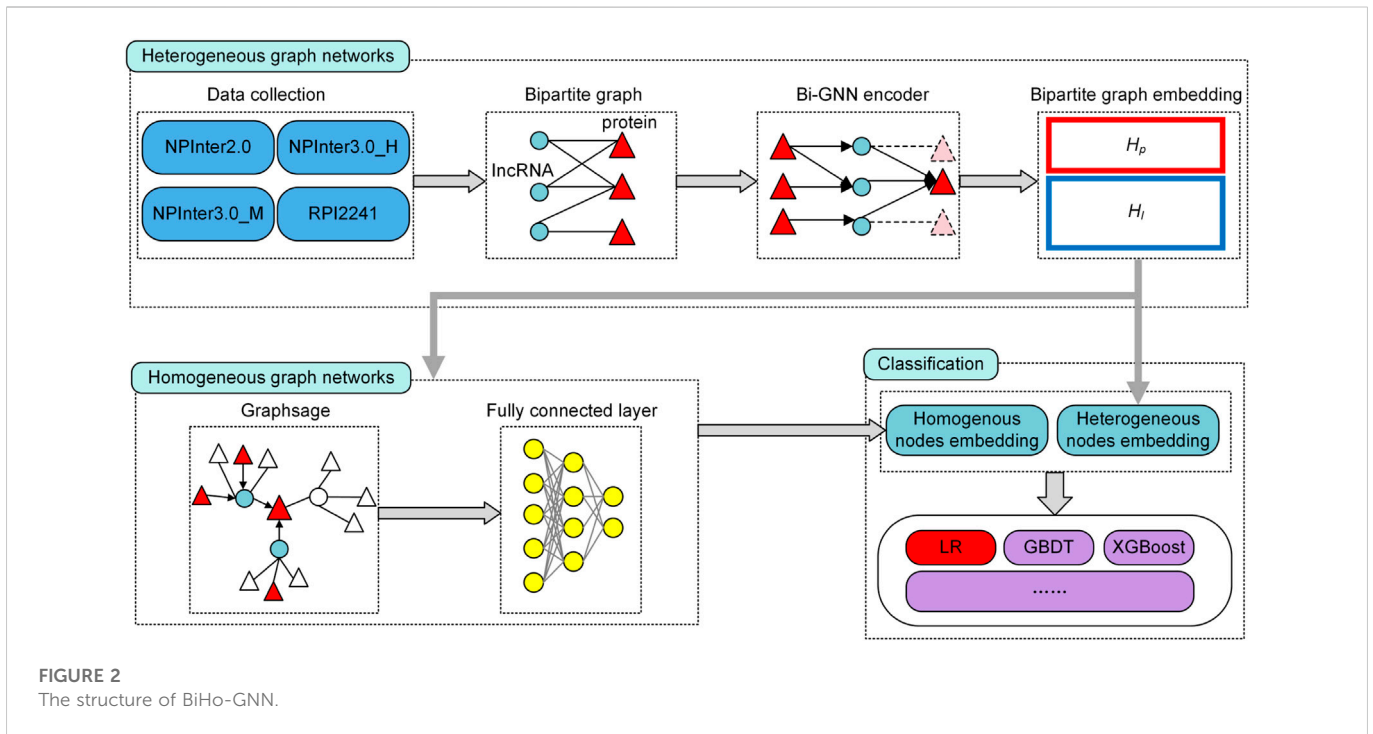


FIGURE 2 The structure of BiHo-GNN.

and we add three rectified linear layers as the final feature to the softmax function.

$$Z = X \cdot \Theta^{(0)} + \text{AGG}(X_q; x_q \in \mathcal{N}(x)) \cdot \Theta^{(1)} \quad (6)$$

where Θ denotes the weights matrix and $\mathcal{N}(x)$ denotes the q -hop neighbors of x .

2.5 Model training

From Eq. 4, we can obtain the lncRNA-protein prototype representation H , Bi-GNN encoder training loss function \mathcal{L}_b is defined as:

$$\mathcal{L}_b = -\frac{1}{|E| + |E'|} \sum_{i=1}^{|E|} \sum_{i=1}^{|E'|} (y_i [\log(\phi(h_i))] + y'_i [\log(1 - \phi(h'_i))]) \quad (7)$$

where y_i is the label of interaction and y'_i is the set of negative lncRNA-protein pairs. ϕ is a temporary classifier involved by a fully connected layer and a sigmoid function.

Through the optimization of node feature H in reconstruction loss \mathcal{L}_r , bipartite Graph embedding can be optimized in homogeneous networks.

$$\mathcal{L}_r = -\sum_{i=1}^{|E|} y_i \log x_i \quad (8)$$

The final loss function \mathcal{L} is composed of above two sections:

$$\mathcal{L} = \alpha \mathcal{L}_b + (1 - \alpha) \mathcal{L}_r \quad (9)$$

where α is a hyperparameter that balances tensor gradient descent between \mathcal{L}_b and \mathcal{L}_r . The procedure of BiHo-GNN is illustrated in Figure 2.

TABLE 2 Performance comparison of lncRNA-protein interaction prediction.

Method	AUC	AUPR	Recall	Precision	F1-score
BiHo-GNN	0.950	0.899	0.919	0.886	0.902
LncPNet	0.938	0.957	0.881	0.948	0.913
LPiGAC	0.936	0.822	0.669	0.832	0.742
LPISKF	0.909	0.685	0.623	0.643	0.633
RWR	0.826	0.581	0.566	0.535	0.550
LPBNI	0.852	0.624	0.634	0.533	0.579

Bold values are the best performance of each task.

3 Results and discussions

3.1 Performance measures

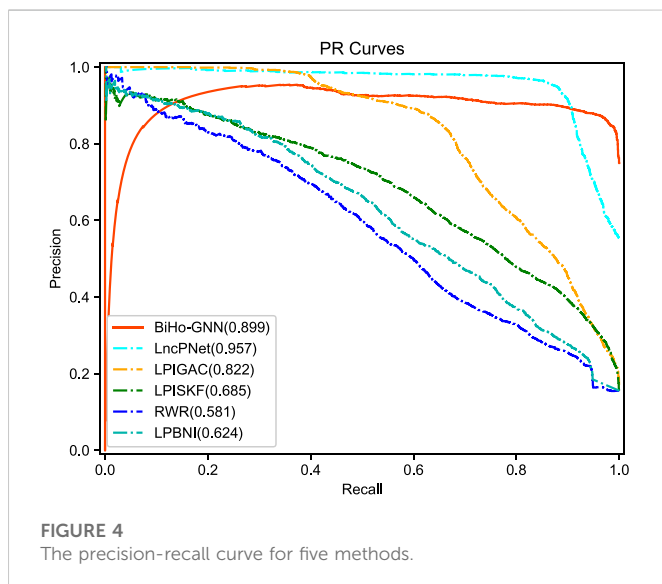
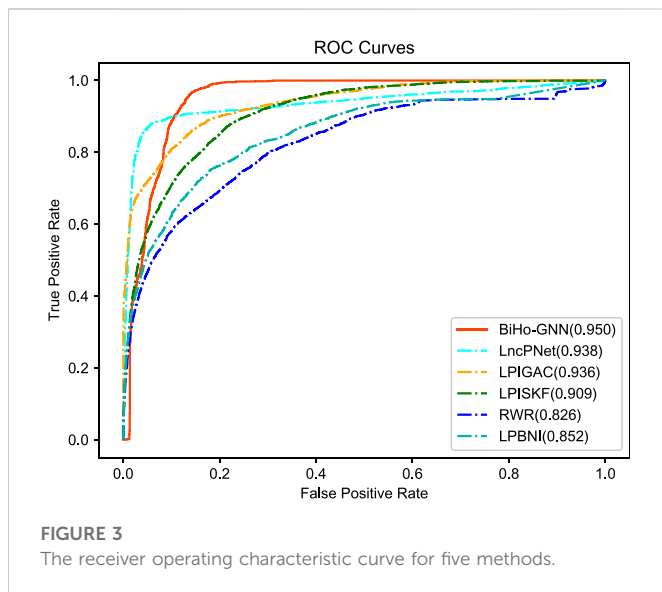
In this paper, we use recall, precision, F1-Score, AUC and AUPR to evaluate the performance of BiHo-GNN. Measurements are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives in the binary classification, respectively. Area under receiver operating characteristic curve (AUC) is utilized to measure the performance of the classifier with TP ratio and FP ratio. Area Under the Precision-Recall curve (AUPR) is used to evaluate the model with precision and recall.



3.2 Comparison with existing methods

We compare the proposed BiHo-GNN framework with five methods, including LPIGAC (Jin et al., 2021), LncPNet (Zhao et al., 2022), RWR (Random Walk with Restart) (Wiggins et al., 2016) and LPBNI (Ge et al., 2016) on NPInter2.0 (5:5). Table 2 demonstrates the performances of BiHo-GNN and the above methods.

From the table, BiHo-GNN achieves AUC of 95.0%, AUPR of 89.9%, Recall of 0.919%, Precision of 88.6%, and F1-score of 0.902. BiHo-GNN outperforms the other five methods under the same experimental conditions. In particular, the AUC and Recall values of BiHo-GNN are increased by 1.2% and 3.8% when compared with the highest evaluation indicators among the methods. ROC curves and PR curves for BiHo-GNN are illustrated on Figures 3, 4.

In Table 2, compared with methods, BiHo-GNN achieves robust performance.

3.3 Comparison with multiple classifiers

The classifier of the model plays a crucial role in the evaluation of the model. We conduct experiments on different classifiers in four data sets, we compare the SVM (Wang and Hu, 2005) classifier based on the linear kernel, the classic classifier XGBoost (Li et al., 2008), the gradient boosting decision tree (GBDT) (Ye et al., 2009), the random forest classifier based on bagging (Breiman, 2001), the k-nearest neighbor algorithm (KNN) (Sun and Huang, 2010) and the logistic regression (LR) method (Kleinbaum et al., 2002). Finally, we set LR as the final classifier. Table 3 indicates that the LR classifier outperforms the other six classifiers.

3.4 Performance analysis

Different from the general validation method of deep learning such as n-fold cross-validation according to the experimental parameters of the previous work (Gao et al., 2020), we split four lncRNA-protein interaction datasets into the 5:5 and 4:6, which denote the ratios of the training set and test set. This specific data set division rule limits the application of BiHo-GNN to datasets with a small amount of interactions. The performance of BiHo-GNN on four datasets is listed in Table 4.

We applied unique data partitioning methods that differed from traditional deep learning validation methods such as 10-fold cross-validation, because of the particularity of the heterogeneous graph, the training set and test set are not allowed to have a big difference in the amount of data, we divided each data set into 5:5 and 4:6. Algorithm convergence of the model under different data sets and partitions are shown in Figures 5, 6.

3.5 Parameter sensitivity

We randomly shuffled the dataset and validated BiHo-GNN on NPInter2.0 (5:5) to train BiHo-GNN for 50 epochs on each dataset. All training processes are run on Windows 11 operation system, a single NVIDIA GeForce RTX3060 GPU with 8 GB memory and Intel(R) Core(TM) i3-12100F CPU @ 3.30 GHz. The hyper-parameter in the model has an impact on the performance of the model. As shown in Table 5, the harmonic factor $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with step length 0.2 is selected for investigating the lncRNA-protein relationship. When $\alpha = 0.9$, our framework achieves the best performance on NPInter2.0 (5:5) and NPInter2.0 (4:6).

We implement BiHo-GNN with packages PyTorch 1.11.0 and PyTorch-geometric. Adam optimizer is adopted for gradient optimization. According to the scale of the dataset and the computation complexity of our framework, we set the feature dimension of the node to be 128. The learning rate is selected from $\{0.001, 0.005, 0.01, 0.05, 0.1\}$ to evaluate the model performance, and verification results show that the proper learning rate is 0.001. Moreover, the learning rate decay is the learning rate multiplied by a hyper-parameter. When the training loss rises, the learning rate

TABLE 3 Performance comparison of different classifiers.

Datasets	Classifier	AUC	AUPR	Recall	Precision	F1-score
NPIInter2.0 (5:5)	SVM	0.902	0.840	0.979	0.847	0.908
	XGBoost	0.640	0.637	0.285	0.980	0.442
	GBDT	0.556	0.553	0.117	0.957	0.209
	Random forest	0.937	0.906	0.946	0.930	0.938
	KNN	0.935	0.898	0.963	0.913	0.937
	LR	0.950	0.899	0.919	0.886	0.902
NPIInter2.0 (4:6)	SVM	0.879	0.814	0.958	0.827	0.887
	XGBoost	0.655	0.652	0.316	0.981	0.479
	GBDT	0.581	0.577	0.174	0.941	0.294
	Random forest	0.928	0.894	0.935	0.922	0.928
	KNN	0.931	0.862	0.955	0.911	0.932
	LR	0.944	0.893	0.920	0.901	0.899
NPIInter3.0_H (5:5)	SVM	0.816	0.763	0.792	0.831	0.811
	XGBoost	0.640	0.637	0.285	0.980	0.442
	GBDT	0.642	0.617	0.246	0.976	0.394
	Random forest	0.830	0.808	0.710	0.934	0.807
	KNN	0.803	0.772	0.698	0.889	0.782
	LR	0.913	0.909	0.587	0.911	0.714
NPIInter3.0_H (4:6)	SVM	0.829	0.807	0.709	0.933	0.805
	XGBoost	0.655	0.654	0.314	0.990	0.442
	GBDT	0.619	0.604	0.240	0.996	0.386
	Random forest	0.872	0.855	0.781	0.954	0.859
	KNN	0.855	0.825	0.793	0.916	0.843
	LR	0.923	0.935	0.682	0.927	0.786
NPIInter3.0_M (5:5)	SVM	0.825	0.741	0.997	0.741	0.850
	XGBoost	0.652	0.652	0.305	0.995	0.468
	GBDT	0.617	0.605	0.226	0.996	0.352
	Random forest	0.904	0.887	0.872	0.932	0.901
	KNN	0.915	0.876	0.924	0.907	0.915
	LR	0.866	0.791	0.945	0.747	0.835
NPIInter3.0_M (4:6)	SVM	0.733	0.653	0.976	0.656	0.785
	XGBoost	0.644	0.639	0.300	0.962	0.458
	GBDT	0.603	0.582	0.218	0.994	0.327
	Random forest	0.870	0.837	0.820	0.911	0.863
	KNN	0.883	0.835	0.895	0.875	0.884
	LR	0.803	0.679	0.741	0.709	0.725
RPI2241 (5:5)	SVM	0.630	0.576	0.908	0.584	0.711
	XGBoost	0.606	0.601	0.212	0.991	0.350
	GBDT	0.616	0.589	0.195	0.982	0.356

(Continued on following page)

TABLE 3 (Continued) Performance comparison of different classifiers.

Datasets	Classifier	AUC	AUPR	Recall	Precision	F1-score
	Random forest	0.937	0.893	0.985	0.898	0.940
	KNN	0.725	0.668	0.679	0.748	0.711
	LR	0.757	0.657	0.634	0.695	0.663
RPI2241 (4:6)	SVM	0.641	0.582	0.912	0.615	0.735
	XGBoost	0.625	0.613	0.241	0.971	0.416
	GBDT	0.626	0.612	0.253	0.964	0.375
	Random forest	0.914	0.905	0.981	0.892	0.958
	KNN	0.752	0.647	0.684	0.751	0.815
	LR	0.775	0.689	0.684	0.705	0.694

Bold values are the best performance of each task.

TABLE 4 Performance of BiHo-GNN on four different datasets.

Datasets	AUC	AUPR	Recall	Precision	F1-score
NPInter2.0 (5:5)	0.950	0.899	0.919	0.886	0.902
NPInter2.0 (4:6)	0.944	0.893	0.920	0.901	0.899
NPInter3.0_H (5:5)	0.913	0.909	0.587	0.911	0.714
NPInter3.0_H (4:6)	0.923	0.935	0.682	0.927	0.786
NPInter3.0_M (5:5)	0.866	0.791	0.945	0.747	0.835
NPInter3.0_M (4:6)	0.803	0.679	0.741	0.709	0.725
RPI2241 (5:5)	0.757	0.657	0.634	0.695	0.663
RPI2241 (4:6)	0.775	0.689	0.684	0.705	0.694

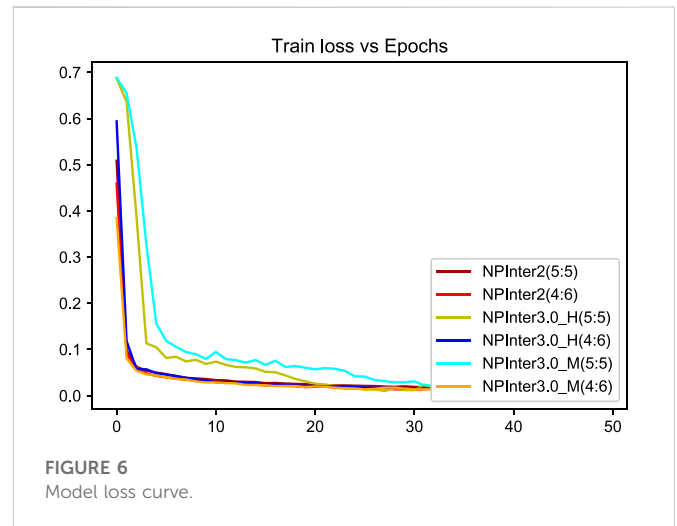


FIGURE 6
Model loss curve.

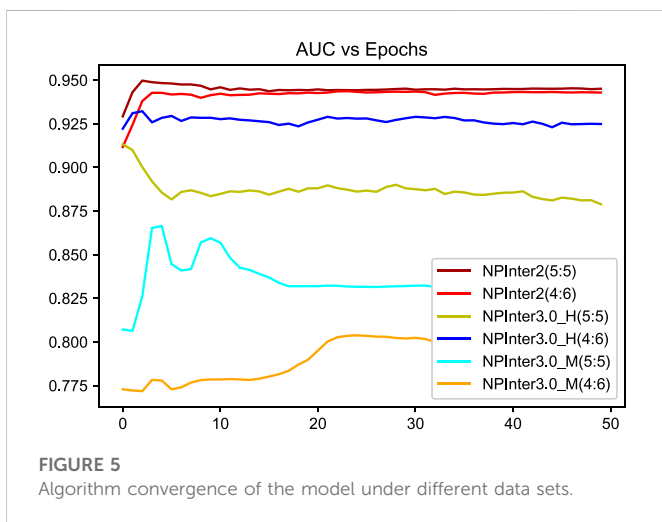


FIGURE 5
Algorithm convergence of the model under different data sets.

decay is set to 0.9, and the results with NPInter2.0 (5:5) and NPInter2.0 (4:6) are shown in Table 6.

Moreover, we evaluate different hidden dimensions of neural networks on BiHo-GNN. As shown in Table 7, when the hidden dimension at 128, BiHo-GNN achieves the best performance on

TABLE 5 Performance and standard deviation with five harmonic factor α .

Datasets	α	AUC	AUPR
NPInter2.0 (5:5)	0.1	0.9218 \pm 0.0037	0.8152 \pm 0.0041
	0.3	0.9347 \pm 0.0038	0.8451 \pm 0.0034
	0.5	0.9431 \pm 0.0062	0.8652 \pm 0.0045
	0.7	0.9429 \pm 0.0043	0.8632 \pm 0.0037
	0.9	0.9501 \pm 0.0046	0.8994 \pm 0.0025
NPInter2.0 (4:6)	0.1	0.9164 \pm 0.0055	0.8027 \pm 0.0043
	0.3	0.9244 \pm 0.0126	0.8371 \pm 0.0152
	0.5	0.9348 \pm 0.0036	0.8582 \pm 0.0046
	0.7	0.9326 \pm 0.0052	0.8569 \pm 0.0028
	0.9	0.9442 \pm 0.0042	0.8933 \pm 0.0058

Bold values are the best performance of each task.

NPInter2.0 (5:5) and NPInter2.0 (4:6), we set the hidden dimension to 128 to balance the cost of time and space of the model.

TABLE 6 Performance and standard deviation with four different learning rate.

Datasets	lr	AUC	AUPR
NPInter2.0 (5:5)	0.001	0.9357 ± 0.0039	0.8732 ± 0.0037
	0.01	0.9501 ± 0.0046	0.8994 ± 0.0025
	0.05	0.9285 ± 0.0024	0.8648 ± 0.0035
	0.1	0.9159 ± 0.0072	0.8541 ± 0.0083
NPInter2.0 (4:6)	0.001	0.9235 ± 0.0051	0.8528 ± 0.0024
	0.01	0.9442 ± 0.0042	0.8933 ± 0.0058
	0.05	0.9157 ± 0.0036	0.8426 ± 0.0024
	0.1	0.9014 ± 0.0039	0.8274 ± 0.0072

Bold values are the best performance of each task.

TABLE 7 Performance and standard deviation with seven different hidden dimensions.

Datasets	Hidden dimensions	AUC	AUPR
NPInter2.0 (5:5)	49	0.9325 ± 0.0057	0.8621 ± 0.0027
	64	0.9356 ± 0.0083	0.8694 ± 0.0138
	81	0.9423 ± 0.0034	0.8792 ± 0.0045
	100	0.9449 ± 0.0057	0.8649 ± 0.0084
	128	0.9501 ± 0.0046	0.8934 ± 0.0025
	144	0.9258 ± 0.0075	0.8493 ± 0.0064
	169	0.9136 ± 0.0063	0.8346 ± 0.0039
NPInter2.0 (4:6)	49	0.9285 ± 0.0047	0.8635 ± 0.0068
	64	0.9263 ± 0.0043	0.8644 ± 0.0118
	81	0.9385 ± 0.0085	0.8726 ± 0.0036
	100	0.9358 ± 0.0038	0.8685 ± 0.0079
	128	0.9442 ± 0.0042	0.8933 ± 0.0058
	144	0.9155 ± 0.0039	0.8495 ± 0.0075
	169	0.9025 ± 0.0085	0.8329 ± 0.0058

Bold values are the best performance of each task.

3.6 Case study

In this section, we mask lncRNA-protein interaction in database NPInter2.0 (Yuan et al., 2014) to infer possible potential associations and verify our result by publications. The predicted top 10 lncRNA-protein interaction is described in Table 8. From the table, we can observe EWSR1 is interacting with NONHSAG029787, NONHSAG008595, and NONHSAG008586, which is associated with non-small-cell Lung, lymphoma, and malignant glioma diseases (Paronetto et al., 2014). NONMMUG000162, NONHSAG055885, NONMMUG038556, and NONMMUG039105 can affect the transcription of the protein AGO4. Urinary bladder neoplasms, uterine cervical neoplasms, and thyroid cancer can be activated by the above interaction pairs (Yuan et al., 2014).

TABLE 8 The predicted top 10 potential lncRNA-protein interaction pairs with BiHo-GNN.

lncRNA	Protein	Confirmed	PubMed
NONHSAG029787	EWSR1	Yes	24813895
NONHSAG008595	EWSR1	Yes	22955616
NONHSAG008586	EWSR1	Yes	22955616
NONMMUG000162	AGO4	Yes	29167373
NONHSAG055885	AGO4	No	—
NONMMUG038556	AGO4	Yes	29167373
NONMMUG039105	AGO4	Yes	29167373
NONHSAG008584	EWSR1	Yes	22955616
NONHSAG008517	EWSR1	Yes	22955616
NONHSAG008516	EWSR1	Yes	22955616

4 Conclusion

lncRNAs are responsible for the regulation of many critical biological processes, such as protein transcription. These two molecular interaction information are closely related to multiple human diseases. It is a significant work to predict potential lncRNA-protein interaction and to study heterogeneous network learning.

In this paper, we propose the novel framework BiHo-GNN for predicting lncRNA-protein interaction. BiHo-GNN utilized bipartite embedding generated by Bi-GNN Encoder. Our work first integrates bipartite graph neural networks and homogeneous graph networks, which strongly verifies the feasibility of heterogeneous graph networks in predicting lncRNA-protein interaction and similar link prediction problems.

Model performance comparison and case study show that BiHo-GNN outperforms state-of-the-art methods on all selected datasets in this paper. Compared with other models using bipartite graph features, BiHo-GNN can well integrate the features of homogeneous networks and heterogeneous networks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Publicly available datasets were analyzed in this study. NPInter2.0 database can be found https://github.com/zhanglabNKU/BiHo-GNN/tree/main/BiHo/dataset_preprocessing/dataset, NPInter3.0 database can be found <http://bigdata.ibp.ac.cn/npinter4/download/>, RPI2241 database can be found https://github.com/zhanglabNKU/BiHo-GNN/tree/main/BiHo/dataset_preprocessing/dataset. Full codes of the BiHo-GNN project are available at our GitHub repository <https://github.com/zhanglabNKU/BiHo-GNN>.

Author contributions

HZ and YM: substantial contributions to the conception and design of the work; YM, HZ, CJ, and CK: drafting the work; the acquisition, analysis or interpretation of data for the work; HZ: revising the work critically for important intellectual content; providing approval for publication of the

content. All the authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This work was supported by key project of the Natural Science Foundation of Tianjin City, China (21JCZDJC00140) and National Natural Science Foundation of China (61973174).

Acknowledgments

We sincerely appreciate our laboratory fellow Mengyao He for his inspiration of this work.

References

- Berg, R. v. d., Kipf, T. N., and Welling, M. (2017). Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., et al. (2012). Noncode v3. 0: Integrative annotation of long noncoding rnas. *Nucleic acids Res.* 40, D210–D215. doi:10.1093/nar/gkr1175
- Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* 30, 1616–1637. doi:10.1109/tkde.2018.2807452
- Cao, J., Lin, X., Guo, S., Liu, L., Liu, T., and Wang, B. (2021). “Bipartite graph embedding via mutual information maximization,” in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 635–643.
- Chen, X., and Yan, G.-Y. (2013). Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* 29, 2617–2624. doi:10.1093/bioinformatics/btt426
- Chen, X., Zhou, Z., and Zhao, Y. (2018). Ellpmda: ensemble learning and link prediction for mirna–disease association prediction. *RNA Biol.* 15, 807–818. doi:10.1080/15476286.2018.1460016
- Cui, P., Wang, X., Pei, J., and Zhu, W. (2018). A survey on network embedding. *IEEE Trans. Knowl. Data Eng.* 31, 833–852. doi:10.1109/tkde.2018.2849727
- Dong, Y., Chawla, N. V., and Swami, A. (2017). “metapath2vec: Scalable representation learning for heterogeneous networks,” in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 135–144.
- Gao, M., He, X., Chen, L., Liu, T., Zhang, J., and Zhou, A. (2020). Learning vertex representations for bipartite networks. *IEEE Trans. Knowl. Data Eng.* 34, 379–393. doi:10.1109/tkde.2020.2979980
- 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, 62–71. doi:10.1016/j.gpb.2016.01.004
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). “Neural message passing for quantum chemistry,” in International conference on machine learning (PMLR), 1263–1272.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of rna-binding protein and microRNA target sites by par-clip. *Cell* 141, 129–141. doi:10.1016/j.cell.2010.03.009
- Hamilton, W., Ying, Z., and Leskovec, J. (2017). “Inductive representation learning on large graphs,” in *Advances in neural information processing systems*, 30.
- Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., et al. (2016). Npinter v3. 0: an upgraded database of noncoding rna-associated interactions. *Database* 2016, baw057. doi:10.1093/database/baw057
- He, X., Gao, M., Kan, M.-Y., and Wang, D. (2016). Birank: Towards ranking on bipartite graphs. *IEEE Trans. Knowl. Data Eng.* 29, 57–71. doi:10.1109/tkde.2016.2611584
- Huang, Z.-A., Huang, Y.-A., You, Z.-H., Zhu, Z., and Sun, Y. (2018). Novel link prediction for large-scale mirna–lncrna interaction network in a bipartite graph. *BMC Med. Genomics* 11, 113–127. doi:10.1186/s12920-018-0429-8
- Jin, C., Shi, Z., Zhang, H., and Yin, Y. (2021). “Predicting lncrna–protein interactions based on graph autoencoders and collaborative training,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), 38–43.
- Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
- Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*. New York, NY: Springer.
- Lewis, B. A., Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., et al. (2010). Pridb: a protein–rna interface database. *Nucleic acids Res.* 39, D277–D282. doi:10.1093/nar/gkq1108
- Li, X., Wang, L., and Sung, E. (2008). Adaboost with svm-based component classifiers. *Eng. Appl. Artif. Intell.* 21, 785–795. doi:10.1016/j.engappai.2007.07.001
- 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
- Muppurala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting rna–protein interactions using only sequence information. *BMC Bioinforma.* 12, 489. doi:10.1186/1471-2105-12-489
- Park, C., Kim, D., Han, J., and Yu, H. (2020). Unsupervised attributed multiplex network embedding. *Proc. AAAI Conf. Artif. Intell.* 34, 5371–5378. doi:10.1609/aaai.v34i04.5985
- Paronetto, M. P., Bernardis, I., Volpe, E., Bechara, E., Sebestyén, E., Eyra, E., et al. (2014). Regulation of fas exon definition and apoptosis by the ewing sarcoma protein. *Cell Rep.* 7, 1211–1226. doi:10.1016/j.celrep.2014.03.077
- Pibouin, L., Villaudy, J., Ferbus, D., Muleris, M., Prospéri, M.-T., Remvikos, Y., et al. (2002). Cloning of the mrna of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. *Cancer Genet. Cytogenet.* 133, 55–60. doi:10.1016/s0165-4608(01)00634-3
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. neural Netw.* 20, 61–80. doi:10.1109/TNN.2008.2005605
- Schaukowitz, K., and Kim, T.-K. (2014). Emerging epigenetic mechanisms of long non-coding rnas. *Neuroscience* 264, 25–38. doi:10.1016/j.neuroscience.2013.12.009
- Shen, Z.-A., Luo, T., Zhou, Y.-K., Yu, H., and Du, P.-F. (2021). Npi-gnn: predicting ncRNA–protein interactions with deep graph neural networks. *Briefings Bioinforma.* 22, bbab051. doi:10.1093/bib/bbab051
- Sun, S., and Huang, R. (2010). “An adaptive k-nearest neighbor algorithm,” in 2010 seventh international conference on fuzzy systems and knowledge discovery (IEEE), 91–94.
- UniProt Consortium (2014). Activities at the universal protein resource (uniprot). *Nucleic acids Res.* 42, D191–D198. doi:10.1093/nar/gkt1140
- Wang, H., and Hu, D. (2005). “Comparison of svm and ls-svm for regression,” in 2005 International conference on neural networks and brain (IEEE), 279–283.
- Wang, Q., Hu, X., Du, M., Lu, Z., Yan, K., Zhao, D., et al. (2020). Znr1-as1 promotes nasopharyngeal carcinoma cell invasion and metastasis by regulating the mir-335–rock1 axis. *Oncotargets Ther.* 13, 4779–4790. doi:10.2147/OTT.S250028
- Wang, Z., Li, T.-E., Chen, M., Pan, J.-J., and Shen, K.-W. (2020). mir-106b-5p contributes to the lung metastasis of breast cancer via targeting cnn1 and regulating rho/rock1 pathway. *Aging (Albany NY)* 12, 1867–1887. doi:10.18632/aging.102719
- Wiggins, B. S., Saseen, J. J., Page, R. L., Reed, B. N., Sneed, K., Kostis, J. B., et al. (2016). Recommendations for management of clinically significant drug–drug interactions with statins and select agents used in patients with cardiovascular disease: a scientific statement from the American heart association. *Circulation* 134, e468–e495. doi:10.1161/CIR.0000000000000456

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.

- Xu, X., Nie, J., Lu, L., Du, C., Meng, F., and Song, D. (2021). Linc00337 promotes tumor angiogenesis in colorectal cancer by recruiting dnmt1, which suppresses the expression of cnn1. *Cancer Gene Ther.* 28, 1285–1297. doi:10.1038/s41417-020-00277-2
- Yang, G., Lu, X., and Yuan, L. (2014). Lncrna: a link between rna and cancer. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.* 1839, 1097–1109. doi:10.1016/j.bbagr.2014.08.012
- Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009). “Stochastic gradient boosted distributed decision trees,” in Proceedings of the 18th ACM conference on Information and knowledge management, 2061–2064.
- Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2014). Npinter v2. 0: an updated database of ncRNA interactions. *Nucleic acids Res.* 42, D104–D108. doi:10.1093/nar/gkt1057
- Zhang, M., and Chen, Y. (2018). “Link prediction based on graph neural networks,” in *Advances in neural information processing systems* 31.
- 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
- Zhao, G., Li, P., Qiao, X., Han, X., and Liu, Z.-P. (2022). Predicting lncrna–protein interactions by heterogeneous network embedding. *Front. Genet.* 12, 814073. doi:10.3389/fgene.2021.814073
- Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). Lpi-deepgbd: a multiple-layer deep framework based on gradient boosting decision trees for lncrna–protein interaction identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8