OXFORD

Briefings in Bioinformatics, 2022, 1–13

https://doi.org/10.1093/bib/bbac411

Problem solving protocol

ncRNAInter: a novel strategy based on graph neural network to discover interactions between lncRNA and miRNA

Hanyu Zhang, Yunxia Wang, Ziqi Pan, Xiuna Sun, Minjie Mou, Bing Zhang, Zhaorong Li, Honglin Li and Feng Zhu

Corresponding authors: Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. E-mail: zhufeng@zju.edu.cn; Honglin Li, Shanghai Key Laboratory of New Drug Design, East China University of Science and Technology, Shanghai 200237, China. E-mail: hlli@ecust.edu.cn

Abstract

In recent years, many studies have illustrated the significant role that non-coding RNA (ncRNA) plays in biological activities, in which lncRNA, miRNA and especially their interactions have been proved to affect many biological processes. Some *in silico* methods have been proposed and applied to identify novel lncRNA–miRNA interactions (LMIs), but there are still imperfections in their RNA representation and information extraction approaches, which imply there is still room for further improving their performances. Meanwhile, only a few of them are accessible at present, which limits their practical applications. The construction of a new tool for LMI prediction is thus imperative for the better understanding of their relevant biological mechanisms. This study proposed a novel method, ncRNAInter, for LMI prediction. A comprehensive strategy for RNA representation and an optimized deep learning algorithm of graph neural network were utilized in this study. ncRNAInter was robust and showed better performance of 26.7% higher Matthews correlation coefficient than existing reputable methods for human LMI prediction. In addition, ncRNAInter proved its universal applicability in dealing with LMIs from various species and successfully identified novel LMIs associated with various diseases, which further verified its effectiveness and usability. All source code and datasets are freely available at https://github.com/idrblab/ncRNAInter.

Keywords: RNA interaction, noncoding RNA, gene regulatory networks, machine learning, neural networks

Introduction

MicroRNAs (miRNAs) and long-noncoding RNAs (lncRNAs) are reported to play critical roles in diverse biological functions of a variety of organisms [1, 2], but the mechanisms underlying their molecular regulation remain elusive [3]. The miRNA regulates the expression of protein-coding genes by integrating into RNA-induced silencing complex and pairing with targeted messenger RNAs (mRNAs) [4, 5]. Within the sophisticated regulatory network of miRNA, lncRNA can function as a competing endogenous RNA (ceRNA) to bind with miRNA against mRNA [6], thereby interfering gene expression [7, 8]. Therefore, the lncRNA–miRNA interactions (LMIs) are found to be essential for the pathogenesis and drug resistance of various diseases [9–14], which indicates the great importance of the identification of novel LMIs [15, 16].

However, current experimental strategies such as RNA pull-down, Luciferase reporter assay, microarray and RT-PCR [17–20] applied for identifying individual LMI can only entail a slow-growing knowledge and present a limited overview of ceRNA network, which significantly hamper the advance of this research field [21]. Recently, the crosslinking immunoprecipitation sequencing (CLIP-seq) has been applied and greatly facilitates the studies on miRNA-related regulation [22], but this technique is still limited by its availability and high cost of spend [23, 24]. Therefore, it is urgently needed to have powerful computational tools to enable the high-throughput and effective discovery of new LMIs.

So far, many computational methods have been applied for miRNA-related and lncRNA-related research [25–28]. Based on these methods, some popular tools have been constructed and emerged for the discovery of novel LMIs [29, 30]. Some of them are based on traditional machine learning methods, such as GBCF [31]. The others are constructed by convolutional neural network (CNN) and recurrent neural network (RNN), such as LncMirNet [32], PmliPred [33], CIRNN [34] and preMLI [35]. The remaining are based on other deep learning (DL) methods, such

Hanyu Zhang is a PhD/MD candidate in the College of Pharmaceutical Sciences at Zhejiang University, China. He is interested in bioinformatics.
Yunxia Wang is a PhD/MD candidate in the College of Pharmaceutical Sciences at Zhejiang University, China. She is interested in bioinformatics.
Ziqi Pan is a PhD/MD candidate in the College of Pharmaceutical Sciences at Zhejiang University, China. He is interested in bioinformatics.
Xiuna Sun is a PhD/MD candidate in the College of Pharmaceutical Sciences at Zhejiang University, China. She is interested in bioinformatics.
Minjie Mou is a PhD/MD candidate in the College of Pharmaceutical Sciences at Zhejiang University, China. He is interested in bioinformatics.
Bing Zhang is a senior expert on big data/AI product and solution in Alibaba Cloud. His current focus includes Digital Health, Health Economics and Knowledge Graph.
Zhaorong Li is a senior expert on big data/AI product and solution in Alibaba Cloud. His current focus includes Digital Health, Health Economics and Knowledge Graph.
Honglin Li is a professor of Shanghai Key Laboratory of New Drug Design at East China University of Science and Technology, China. His research group (http://lilab-ecust.cn/) has been working in the fields of AI-based drug design.
Feng Zhu is a tenured professor in the College of Pharmaceutical Sciences at Zhejiang University, China. His research group (https://idrblab.org/) has been working in the fields of AI aided drug discovery.

Received: June 7, 2022. **Revised:** August 4, 2022. **Accepted:** August 23, 2022

as EPLMI [36], SLNPM [37], GCLMI [38], LNRLMI [39], LMNLMI [40], GEEL [41] and LMI-INGI [42]. However, all these methods are constructed using the RNA representation approach that merely based on sequence, expression profile or their similarity [40]. On the one hand, such approach cannot fully reflect the roles of physicochemical and structural properties in determining LMIs [43]. On the other hand, the similarity among molecules cannot completely represent their functional relevance since any slight variation in RNAs could lead to dramatical functional variation [44]. Moreover, the existing tools do not consider the problem of information leak [27] and do not utilize the power of DL to transmit information [42]. All these problems significantly limit the application of the existing tools, which ask for the construction of new tools with substantially improved RNA representation and extensively elevated model construction.

In this study, a new method named ncRNAInter was therefore constructed to enable the discovery of LMI based on DL strategy. This method was unique in (1) integrating a comprehensive strategy for RNA representation that considered not only the sequence but also the physicochemical and structural properties and (2) applying an advanced framework of graph neural network (GNN) that incorporated information leak avoidance and nonlinear updating for LMI prediction. To test the performance of this newly developed method, ncRNAInter was systematically evaluated from multiple angles and compared with state-of-the-art methods for LMI prediction, showing great superiority and strong credibility. Moreover, we appraised the potential applicability of ncRNAInter in various species and verified its adaptability to predict unknown LMIs associated with various diseases. Based on the analyses in this study, the ncRNAInter demonstrated a significantly enhanced performance in modern LMI research, meanwhile achieving extraordinary robustness and universal applicability, and could thus be considered as a good complement to other existing methods in the related research community. The ncRNAinter tool is open-source, which makes the results of this study fully reproducible.

Materials and methods

Benchmark datasets and data collection

Numerous methods for human LMI prediction, including LMI-INGI [42], GEEL [41], LncMirNet [32], EPLMI [36], etc., use experimentally validated RNA interactions from LncRNASNP2 database [45] to conduct their research. Likewise, in this study, we collected and utilized 18 595 experimentally validated LMIs from this widely used benchmark dataset to conduct our research, in which lncRNAs are annotated by ENSEMBL project [46]. Sequence information of miRNAs was obtained from miRbase-v22.1 [47], while sequence information of lncRNAs was obtained from GENCODE-v38 (the hg38/GRCh38 genome assembly version) [48, 49] through mapping the symbols of each individual transcript from LncRNASNP2 to these databases thoroughly. Furthermore, since the data from LncRNASNP2 are reported for individual transcripts, the interactions of similar splicing variants transcribed from the same gene could lead to redundancies, which might artificially inflate the performance of the method. To reduce this concern, redundant sequences (i.e. identity ≥ 0.9 [50]) were discarded using CD-HIT [51], which uses a short word filter to avoid unnecessary alignments and has been widely used as a clustering algorithm [52]. After sequence information matching and data screening, invalid data were removed and 13 800 practicable RNA pairs remained as positive dataset, including 266 miRNAs and 1499 lncRNAs. Random sampling on accessible RNAs against the positive pairs was applied to build a negative dataset of the same size as the positive set. The final balanced dataset with 27 600 LMIs was randomly split out 10% for testing using stratified sampling, and the remaining 24 840 pairs were used for training and validation. The statistics of the human LMIs data (benchmark 1) are illustrated in Supplementary Table 1, available online at http://bib.oxfordjournals.org/.

To better support our method and conduct more thorough analyses, two extra benchmark datasets of other species were prepared for additional training and verification. Plant LMI benchmark dataset (benchmark 2) was derived from the same dataset constructed by PmliPred [33], which includes 15 000 positive and negative LMIs from *Arabidopsis thaliana*, *Glycine max* and *Medicago truncatula*. Virus LMI benchmark dataset (benchmark 3) was collected from ViRBase v3.0 [53]. After similar process of sequence information matching and redundancy filtering, a total of 9652 interactions between virus miRNA and host (*Homo sapiens*) lncRNA were sorted as positive dataset. Then, an equal number of negative pairs were randomly sampled out, resulting in a final balanced dataset with 19 304 LMIs. The statistics of benchmark 2 and benchmark 3 are also illustrated in Supplementary Table 1, available online at http://bib.oxfordjournals.org/.

The comprehensive strategy for RNA representation

A comprehensive strategy for RNA representation is adopted in this study, differing from those existing sequence-based, expression-based or similarity-based methods. This strategy facilitates the deep learning model to distinguish specific characters of RNAs according to general rather than partial information, which limits their model's performance to a certain level [54]. This innovative strategy could avoid the negative influence of incomplete feature representation.

There have already been many researches encoding RNA sequence-intrinsic features to conduct related tasks such as RNA classification [55, 56], RNA coding potential prediction [52, 57, 58] and physiological function annotation [59]. Though physicochemical and secondary structural properties play important roles in RNA biological functions [28, 55], they were seldom considered in related research, especially LMI prediction. Therefore, in this study, RNA features including not only sequence-intrinsic but also physico-chemical and secondary structural properties are utilized for RNA feature representation [60, 61]. Specifically, a total of five codon-related features including Fickett score and stop codon-related properties; 31 ORF-related features including basic ORF properties, entropy density profile scores and Hexamer scores on ORF; 7 GC-related features including Guanine/Cytosine content properties; 126 transcript-related features including UTR-related properties, Basic transcript property, *K*-mer scores, CTD descriptors, entropy density profile scores and Hexamer scores on transcripts; 13 physicochemical property features including Pseudo-protein properties and EIIP spectrum scores; 9 Secondary structure features including multi-scale secondary scores and secondary structure descriptors are used in this study, as illustrated in Table 1. These diversified features are calculated on individual nucleotide sequences and then learned by DL model to conclude the patterns for LMI prediction. To the best of our knowledge, this is the most comprehensive representation strategy that has ever been used in the LMI prediction problem. All features applied in this study have been commonly accepted in influential studies on non-coding RNA-related researches such as CPPred [52], LncFinder [55], LncADeep [59], Seq-SymRF [62], PmliPEMG [63] and NCResNet [64]. The

Table 1. Comprehensive strategy of RNA intrinsic features applied in this study

Feature group (*type*)	Feature subgroup	Number of features
Codon-related (*sequence-intrinsic*)	Fickett score	1
	Stop codon-related properties	4
ORF-related (*sequence-intrinsic*)	Basic ORF properties	4
	EDP scores on ORF	20
	Hexamer scores on ORF	7
GC-related (*sequence-intrinsic*)	GC content properties	7
Transcript-related (*sequence-intrinsic*)	UTR-related properties	4
	Basic transcript property	1
	K-mer ($k = 3$)	64
	CTD descriptors	30
	EDP scores on transcript	20
	Hexamer scores on transcript	7
Physicochemical property (*physico-chemical*)	Pseudo-protein properties	5
	EIIP spectrum scores	8
Secondary structure (*secondary structural*)	Multi-scale secondary scores	6
	Secondary structure descriptors	3

detailed information and concrete calculation procedure of the features are illustrated in Supplementary Method 1, available online at http://bib.oxfordjournals.org/.

Novel architecture of the applied deep learning strategy

Nowadays, DL belonging to machine learning (ML) has become a popular technique for many complicated problems in the field of biochemistry and molecular biology [65–70]. Classical DL methods such as CNN have been used for Euclidean structure data and achieved great success [71–74]. However, most biological data such as protein–ligand interaction network are from non-Euclidean spaces in the practical applications [75], where classical ML and DL models might be unsuitable [76]. As one of the most emerging methods initially created to deal with non-Euclidean data [77, 78], GNN has made remarkable achievements in single-cell classification [79], RNA–protein interaction prediction [27], synthetic lethality prediction [80], polypharmacy side effect prediction [81], lncRNA target prioritizing [82] and so on.

Therefore, ncRNAInter optimizes a graph neural network framework of GraphSAGE [83] incorporating information leak avoidance and non-linear updating, in order to fully capture the RNA feature information and RNA interaction knowledge to predict LMIs. As illustrated in Figure 1, ncRNAInter mainly consists of four components:

RNA feature encoding and graph building

ncRNAInter encodes all RNAs into feature vectors in 191 dimensions based on the comprehensive strategy for RNA representation, as illustrated in Figure 1A. Based on benchmark 1, ncRNAInter builds a graph consisting of 1765 nodes (266 miRNAs and 1499 lncRNAs) and 24 840 edges (24 840 interactions). Node representations are set as corresponding RNA feature vectors. Edge weights are set as '1' when the interaction is positive but '0' when the interaction is negative or indeterminate.

Message transmission and non-linear node updating

This process mainly contains two modules, one is neighbor sampling and information aggregating and the other is non-linear node updating, as illustrated in Figure 1B.

When dealing with a specified edge ij, node i and node j will be activated to sample their source nodes as neighbors for message transmission. The linked neighbors $N(i)$ will multiply their respective node features $h_{N(i)}^{k-1}$ with the corresponding edge weights $w_{E(i)}$, and the calculation result will be considered as the neighbor messages $m_{N(i)}^{k-1}$ for node i (1).

$$m_{N(i)}^{k-1} = h_{N(i)}^{k-1} \times w_{E(i)} \tag{1}$$

Then, node i's feature h_i^{k-1} together with neighbor messages $m_{N(i)}^{k-1}$ will be put into an optional reduce function AGG( ), such as SUM (summing all messages), MEAN (taking the average of all messages), MAX (picking the max of all messages) and MIN (picking the min of all messages) (2).

$$a_i^k = \mathrm{AGG}\left(h_i^{k-1}, m_{N(i)}^{k-1}\right) \tag{2}$$

The aggregated messages a_i^k will further be processed by a learnable neural network for non-linear node updating. The output h_i^k will be regarded as the updated hidden feature of node i (3):

$$h_i^k = \mathrm{ReLU}\left(W^{k-1}a_i^k + b^{k-1}\right) \tag{3}$$

Edge embedding classifier

As illustrated in Figure 1B, by concatenating the updated hidden features of the two linked nodes, the embedding of the specified interaction (edge) f_{ij}^k could be defined and then put into a learnable fully connected neural network. Finally, the two-class classification is conducted by calculating the probability of different categories. Cross entropy loss is employed to measure the probability distribution of the model assessments (4).

$$H(p,q) = -\sum_{i=(\mathrm{pos,neg})} p(i)\log q(i) \tag{4}$$

Information leak avoidance

In the process of graph learning, the problem of information leak may arise in some cases, which could result in possibly overestimated performances. For example, the message transmitted from edge ji will carry knowledge that should not have been exposed to its reverse edge ij. In addition, validation data could

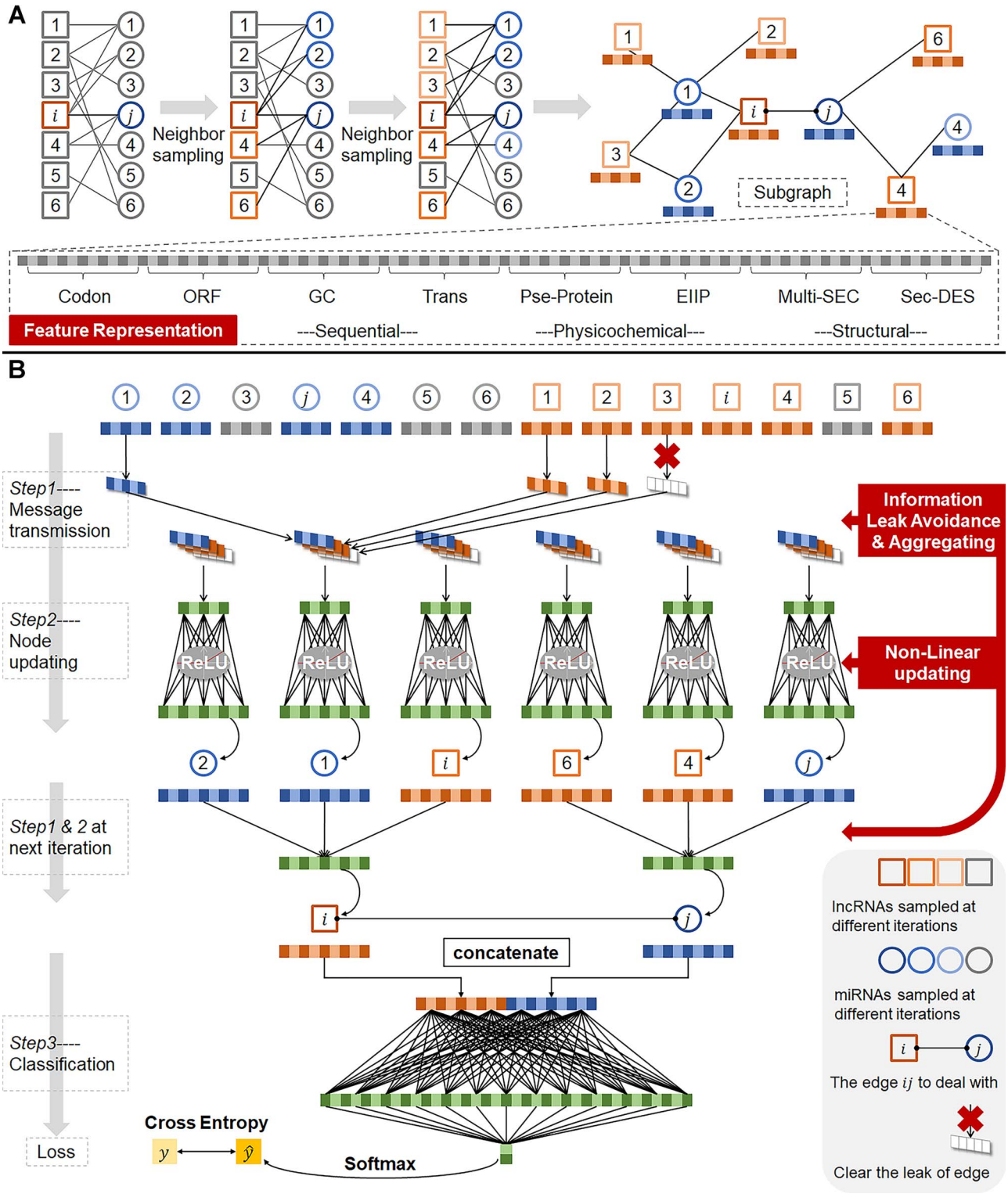

Figure 1. ncRNAInter workflow. (**A**) ncRNAInter first encodes all RNAs into feature vectors as node representations, based on RNA sequence-intrinsic, physico-chemical and secondary structural properties. When dealing with the interaction between RNA *i* and *j*, ncRNAInter conducts multiple iterations of neighbor sampling to construct a subgraph of edge *ij* for further processing. (**B**) All messages of neighbors are aggregated after sampling to be learned by a fully connected neural network. Updated node features will then be concatenated for edge classification. In order to avoid the risk of information leak, the reverse edge will be excluded and information in validation dataset will be cleared while training in each CV round.

be divulged through neighbor sampling in the training process of the cross-validation (CV) procedures [27]. In order to avoid the risk of information leak, when dealing with one edge, its reverse edge will be excluded. Additionally, information of edges belonging to validation dataset will be cleared in the process of aggregating and updating in each CV round. These measures ensure that no leaked information will be transmitted and exposed during the process of graph learning.

Parameter calibrations and model implementations

After multiple pre-trainings and adjustments, the number of the iterations of message transmission and node updating was set to '2' and the reduce function for information aggregation was set as MEAN. Softmax function was adopted as the activation function of the final layer for edge classifier [79]. Adaptive momentum estimation optimization (Adam) was adopted to optimize the GNN model [79]. Hyperparameters of learning rate and hidden dimension were set to 0.0005 and 256, as illustrated in Supplementary Figure 1, available online at http://bib.oxfordjournals.org/.

Experimental setup, evaluation criteria and system deployment

Stratified 5-fold CV was employed to evaluate the performance of ncRNAInter. In stratified 5-fold CV, the datasets are divided into five groups with the distribution of positive and negative samples as close as possible. In each CV round, one group is treated as validation set in rotation, while the other four are collected as training set.

The performance of ncRNAInter was evaluated according to the average values of accuracy (ACC), Matthews correlation coefficient (MCC), precision (PRE), recall (REC), specificity (SPC), F1 score (F1), area under the curve of receiver operating characteristic curve (AUROC) and precision-recall curve (AUPRC) [33] among 5-fold CVs. ncRNAInter is mainly implemented by pytorch 1.7.1 (https://pytorch.org/), scikit-learn 0.24.1 (https://scikit-learn.org/stable/) and dgl 0.6.1 (https://www.dgl.ai/). All scripts were written by Python 3.8.8. ncRNAInter was deployed on the computer with Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz, NVIDIA(R) Tesla(R) P100 16GB GPU and 263GB RAM on CentOS Linux release 7.9.2009 (Core).

Results and discussion

Evaluation of the predictive performance of ncRNAInter

ncRNAInter achieved satisfying performance with the average ACC of 0.9309, MCC of 0.8619, precision of 0.9342, recall of 0.9272, specificity of 0.9346, F1 score of 0.9307, AUROC of 0.9715 and the AUPRC of 0.9741 among 5-fold CVs on benchmark 1 for human LMI prediction, as the blue bar shown in Figure 2. We then used our testing dataset to evaluate our method, whose results achieved good robustness with the average ACC of 0.9251, MCC of 0.8504, precision of 0.9144, recall of 0.9380, specificity of 0.9122, F1 score of 0.9260, AUROC of 0.9726 and the AUPRC of 0.9748 among 5-fold CVs, as illustrated in Supplementary Table 2, available online at http://bib.oxfordjournals.org/. The standard deviations of all metrics among 5 folds turned out to be minimal, which further illustrated the stability of our method.

Moreover, since ncRNAInter built the original graph with all LMIs from benchmark 1 for training and validation, it is suspectable that validation datasets could be divulged in the training process. In this study, to avoid the risk of information leak and the possibly overestimated performances, the model will clear the information of LMIs from validation dataset in the process of aggregating and updating, as illustrated in Figure 1B. To further illustrate the influence of this measure, the training process of ncRNAInter with and without information leak avoidance was compared. As illustrated in Figure 3, the leaked information of validation data observably resulted in the inflated ACC values and a faster convergence speed while training, which amply demonstrates the necessity of information leak avoidance.

Graph neural network model contributes to LMI prediction

To probe the influence of the GNN framework applied in this study upon LMI prediction, we compared ncRNAInter with other classic machine learning methods including Support Vector Machine (SVM), Random Forest (RF) and CNN using the same strategy for RNA representation based on benchmark 1. The constructions of the classic machine learning methods used in this study are illustrated in Supplementary Method 2, available online at http://bib.oxfordjournals.org/. As shown in Figure 2, ncRNAInter made significant improvements among all metrics compared to these classic models.

Through such comparison, it is found that ncRNAInter using GNN model significantly improved the performance of LMIs prediction. The reasons may include but are not limited to the following analysis. First, the inputted RNA features were encoded from general perspectives, which lead the GNN model to learn from comprehensive RNA representations. This abundant information of RNA ensures that the GNN model will not miss any important factors for LMI and could fully extract RNA representations at high levels of abstraction through graph learning [84]. Second, the GNN model has an advantage in extracting the information of neighbor RNAs, allowing each RNA to use different parameters to weigh the information of its different neighbors. By extrapolating this strategy to more iterations between neighbor RNAs, the GNN model can learn edge- and neighbor-dependent weights to capture local detail [85], which further enriched the surrounding information of RNAs in the whole LMI network, and yet it was totally neglected by other machine learning methods. All these advantages of GNN contribute to the entire process of learning RNAs' intrinsic features and their surrounding information, which greatly helps our method to capture the vital information of LMIs and finally achieve the excellent performance in LMI prediction.

All in all, with both RNA intrinsic feature and RNA surrounding information captured in the learning process, the GNN model obviously contributes to LMI Prediction while using ncRNAInter.

Comprehensive strategy for RNA representation contributes to LMI prediction

To probe the contribution of the comprehensive feature representation strategy for LMI prediction, all features used by ncRNAInter were analyzed based on importance scores computed from the permutation algorithm [86]. Through progressive scanning on features of different RNAs, feature importance scores were calculated based on the estimated error increase caused by permuting relevant feature values. The top 50 important features of lncRNAs as well as the top 50 important features of miRNAs are shown in Figure 4. In the case of lncRNA, ORF-related features, Secondary structure features and transcript-related features turned out to be the most important ones in LMI prediction, whereas physicochemical property features, GC-related features, codon-related features and transcript-related features ranked in the top in the case of miRNA. Different categories of features prove to play a part in different RNAs for LMI prediction, which indicates that neglecting any types of features may negatively impact the representation of certain types of RNAs and thereby the performance of LMI prediction. This further confirmed the indispensability of the comprehensive RNA representation strategy, which could help the GNN model to learn integral information of RNAs at the very beginning.

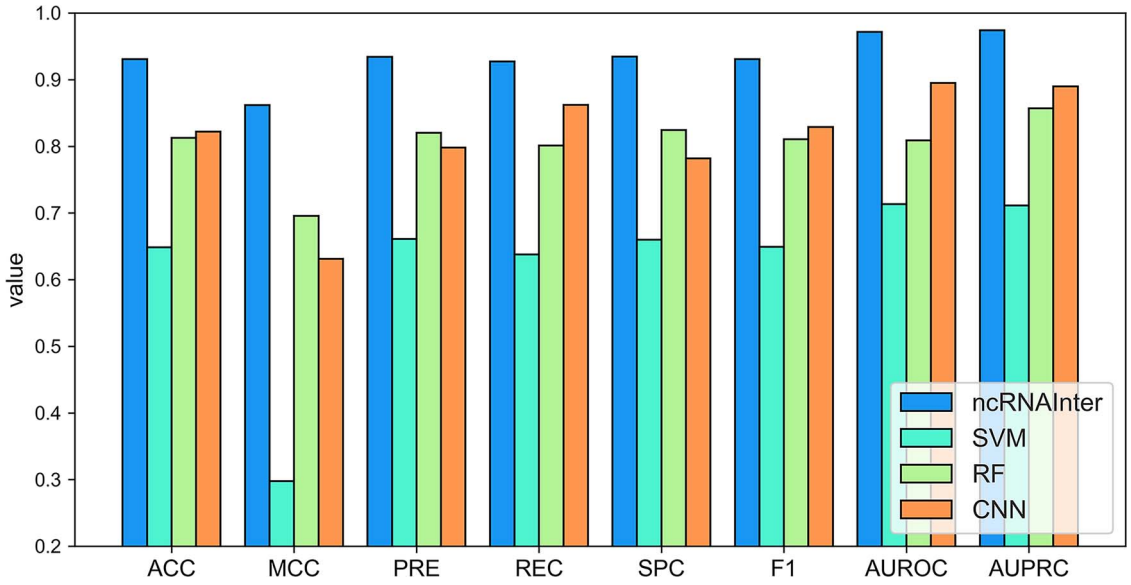

Figure 2. Model performance of ncRNAInter, SVM, RF and CNN using the same RNA representation strategy and dataset. ncRNAInter achieved the best performance among all metrics compared with the other classic machine learning methods.

All features applied in this study have been commonly accepted in influential studies on lncRNA-related research. For example, CPPred applied some types of features (including ORF related features, CTD related features) to identify coding potential in lncRNA and mRNA [52]. LncADeep applied some types of features (including UTR-related features, GC-related features, secondary structure-related features) to identify lncRNA and predict lncRNA–protein interactions [59]. LncFinder applied some types of features (including k-mer features, hexamer-related features, EIIP-related features) to discriminate lncRNA and mRNA [55]. NCResNet applied some types of features (including pseudo-protein features, Fickett score, codon-related features) to distinguish ncRNA and pcRNA [64]. However, some feature calculations might be deemed inappropriate for miRNA representation. For example, ORF-related features are commonly unable to be calculated on miRNA because of its short sequence (approximately 18–25 nucleotides in length) [87], and secondary structure features are more frequently used on the representation of pre-miRNA rather than mature miRNA in existing researches [62, 88, 89]. Therefore, two extra models were trained separately for appraising the influence of inappropriate feature calculations and comparing with the model using secondary structure features calculated on pre-miRNA. The first model excluded the inappropriate feature calculations of miRNA containing stop codon-related features, UTR-related features, pseudo-protein features, ORF-related features and secondary structure features; the second model used secondary structure features calculated on precursor sequences of miRNA. As a result, two extra models performed equally to the original model, as shown in Supplementary Figure 2, available online at http://bib.oxfordjournals.org/. The result of the first model indicates that since GNN can automatically sort out both contributory and inessential factors for LMI prediction, the ncRNAInter can restrict the influence of the features that have limited help for LMI prediction in the process of training, so that they won't interfere the LMI prediction; the result of the second model indicates that the choice of whether to use pre-miRNA or mature miRNA to calculate the secondary structure features does not have impact on LMI prediction. Therefore, the original calculation is applied, which can save the trouble of manual feature selection and pre-miRNA data preparation for the convenience of users, meanwhile maintaining the best predicting performance.

All in all, the integral RNA feature encoding is essential for LMI prediction and certain inappropriate feature calculations have little impact on model performance. Based on that, ncRNAInter eventually applies the comprehensive strategy for RNA representation, significantly improving the prediction of LMIs meanwhile ensuring the credibility and rationality of our method.

Comparison between ncRNAInter and existing methods

Here, ncRNAInter was tested to compete its performance with state-of-the-art methods, to assess the improvements of this new AI method made in predicting LMIs. The mainstream of most reputable methods for LMI prediction mainly included EPLMI [36], GBCF [31], SLNPM [37], GCLMI [38], LNRLMI [39], LMNLMI [40], GEEL [41], LncMirNet [32], LMI-INGI [42], PmliPred [33] and preMLI [35]. However, only four methods, LncMirNet [32], LMI-INGI [42], PmliPred [33] and preMLI [35], were accessible and repeatable for LMI prediction by far. In order to compare their performance with ours, the exact same human LMIs data (benchmark 1) were used to train LncMirNet, LMI-INGI, PmliPred (human) and preMLI (human). Their average ACC, MCC, precision, recall, specificity, F1 score, AUROC and AUPRC among 5-fold CVs were inspected with emphasis. As shown in Table 2a, ncRNAInter made significant improvements being 13.8% superior at ACC, 26.7% at MCC, 17.1% at precision, 4.9% at recall, 17.4% at specificity, 12.5% at F1 score, 8.2% at AUROC and 9.6% at AUPRC to the best one of four tools. All in all, ncRNAInter achieved remarkable superiority among all metrics compared to currently accessible methods.

Therefore, it can be concluded that ncRNAInter is a competent and rather competitive method in LMI prediction. In this study, not only the well-trained model is provided, but also our source code

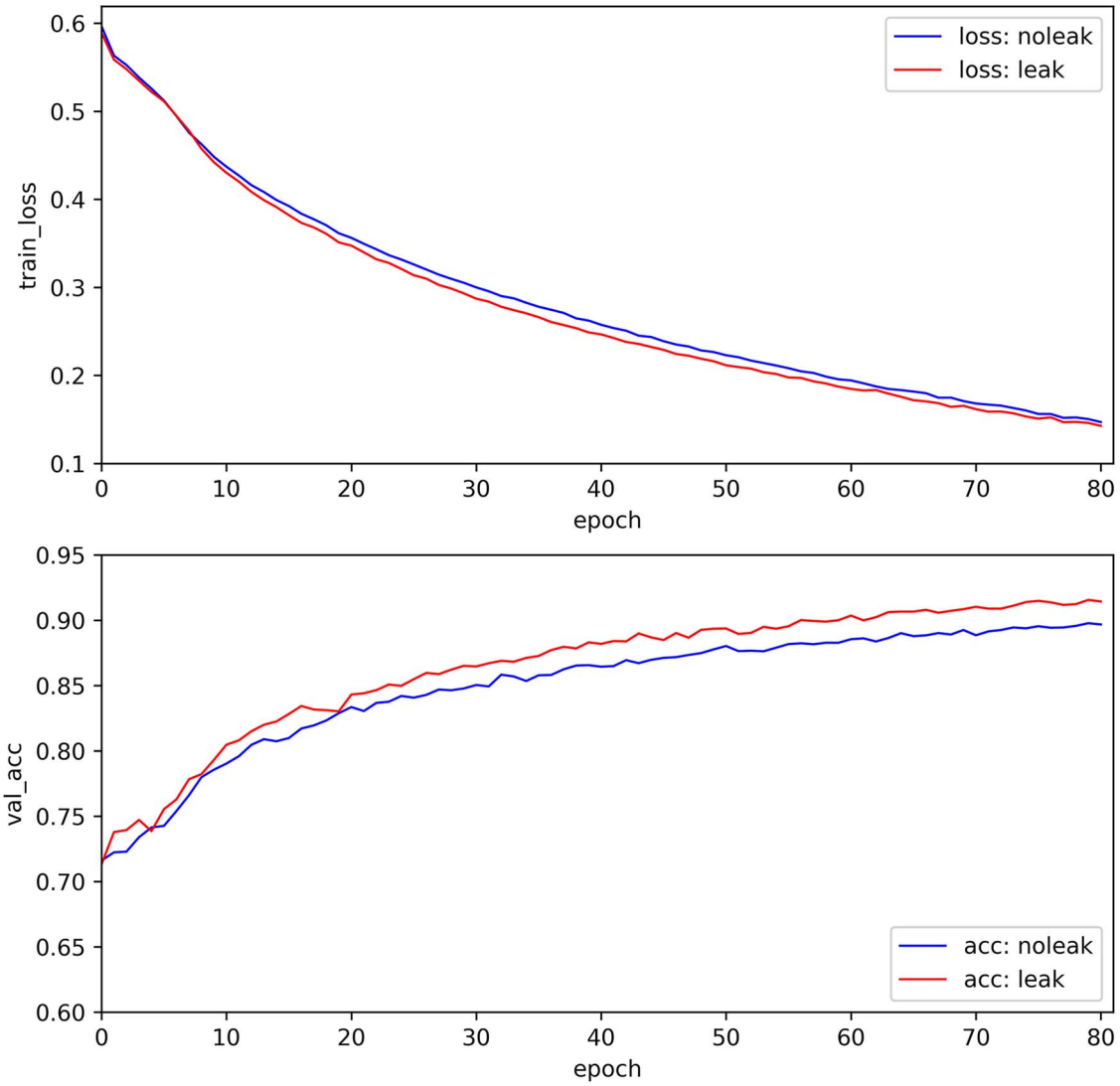

Figure 3. Comparison between the training process of ncRNAInter with and without information leak avoidance. In the process of model training, the model without information leak avoidance showed faster training loss reduction and validation ACC increase. Our strategy successfully precluded the model from this inflated consequence.

Table 2. The performance of ncRNAInter and other repeatable state-of-the-art methods using the benchmark datasets

	ACC	MCC	PRE	REC	SPC	F1	AUROC	AUPRC
(a)								
ncRNAInter	**0.9309**	**0.8619**	**0.9342**	**0.9272**	**0.9346**	**0.9307**	**0.9715**	**0.9741**
preMLI	0.8178	0.6394	0.7866	0.8724	0.7633	0.8273	0.8979	0.8888
PmliPred	0.8001	0.6801	0.7979	0.8040	0.7962	0.8008	0.8506	0.8515
LMI-INGI	0.6916	0.4427	0.6809	0.8842	0.5046	0.7422	0.8906	0.8729
LncMirNet	0.5479	0.0961	0.5000	0.5471	0.5489	0.5473	0.5693	0.5706
(b)								
ncRNAInter	**0.9429**	**0.8862**	**0.9345**	**0.9528**	**0.9331**	**0.9435**	**0.9863**	**0.9869**
PmliPred	0.9191	0.8511	0.9178	0.9228	0.9155	0.9193	0.9682	0.9607
(c)								
ncRNAInter	**0.9604**	**0.9211**	**0.9504**	**0.9716**	**0.9492**	**0.9608**	**0.9855**	**0.9809**

All methods were assessed based on 5-fold CV, and the performances reported were the average values among five CVs. The results of ncRNAInter were indicated in boldface. ACC, accuracy; MCC, Matthews correlation coefficient; PRE, precision; REC, recall; SPC, specificity; F1, F1 score; AUROC, area under the curve of receiver operating characteristic; AUPRC, area under the precision-recall curve. (a) Comparing ncRNAInter with LncMirNet, LMI-INGI, PmliPred and preMLI by repeating them based on the dataset of benchmark 1 for human LMI prediction. (b) Comparing ncRNAInter with PmliPred by repeating them based on the dataset of benchmark 2, which was originally reported in PmliPred for plant LMI prediction. (c) The performance of ncRNAInter trained on benchmark 3 for virus-related LMI prediction.

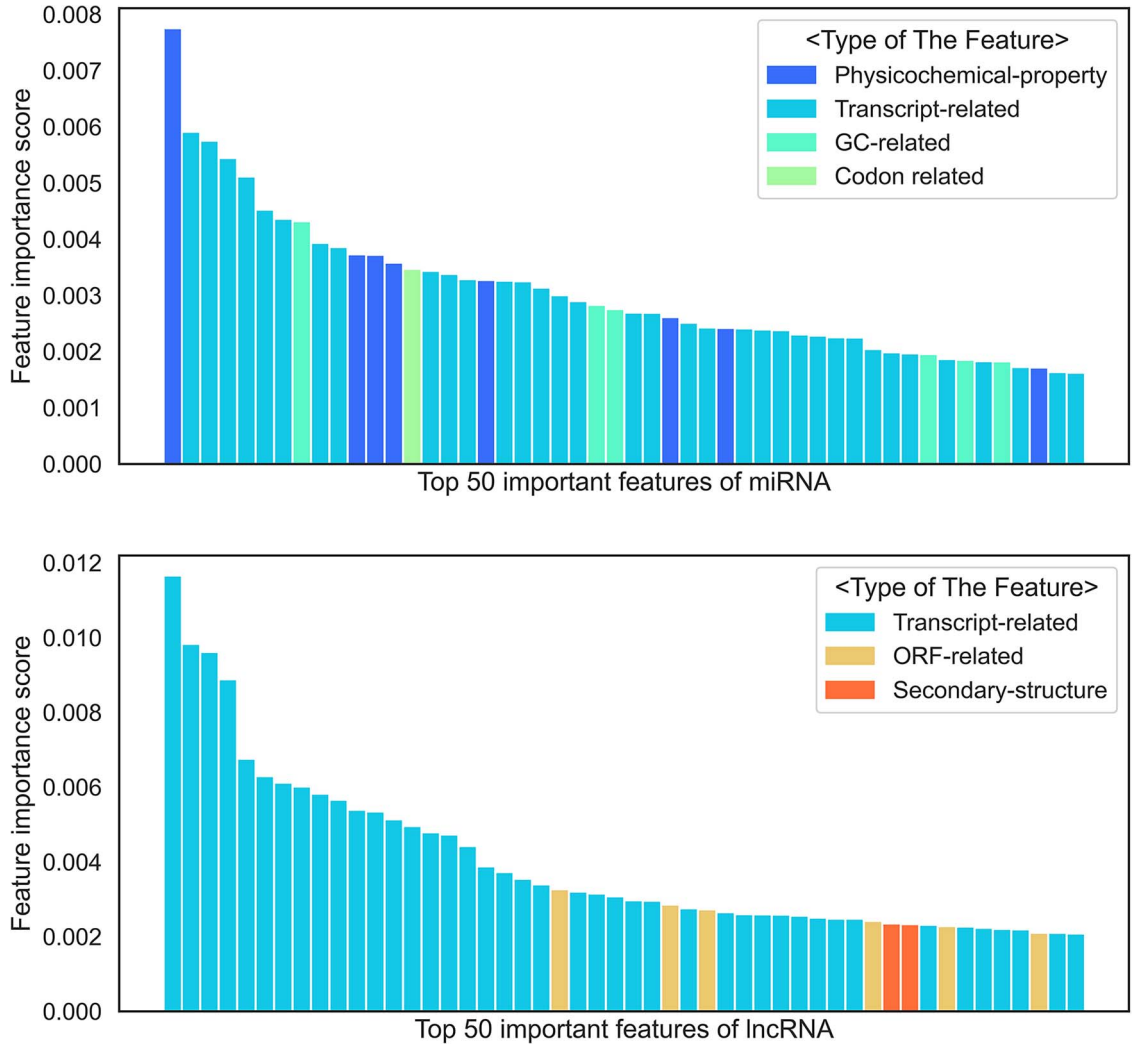

Figure 4. Rankings of feature importance of miRNA and lncRNA with top 50 important features presented. Different categories of features were found to play a part in different RNAs for LMI prediction, which indicated the indispensability of the comprehensive RNA representation strategy.

and datasets are completely released for users to fully repeat and utilize. It is believed that ncRNAInter has a very strong impetus for follow-up research.

Applying ncRNAInter to predicting LMIs from various species

All the previous discussions were focused on human LMIs. We are also intrigued to find out whether our method worked well on the LMI prediction for other species. To ulteriorly inspect these capabilities of ncRNAInter, the method was employed to predict LMIs associated with plants and viruses.

Application of ncRNAInter in plant LMIs

In order to evaluate the application possibility of ncRNAInter in plant LMI prediction, plant LMI data (benchmark 2) derived from the same dataset constructed by PmliPred [33] were used to train ncRNAInter (plant). As reported, PmliPred is a reputable tool originally for predicting plant LMIs. It has strong capability of predicting LMIs for plants and it is necessary to completely repeat PmliPred for comparison. As shown in Table 2b, ncRNAInter (plant) achieved better performance than PmliPred (plant) did with ACC of 0.9429, MCC of 0.8862 and AUROC of 0.9863 while the latter achieved ACC of 0.9191, MCC of 0.8511 and AUROC of 0.9682. Moreover, the results of the PmliPred (plant) were rather close to its reported results in original publication with deviations of ACC being approximately 1%, which was basically the same and further illustrated the reproducibility and reliability of this method. Such repeatable tools are demanded in the academic field because their ability to facilitate subsequent reconstruction and utilization for researchers. However, in this case, ncRNAInter still achieved modest improvement of about 1–5%, which convincingly proved its potential to play a role in the LMI prediction for plants.

Application of ncRNAInter in virus-related LMIs

In the pursue of evaluating the application possibility of ncRNAInter in broader fields, another attempt of applying the method to predict the interactions between virus miRNA and host (*H. sapiens*) lncRNA was carried out. In this study, virus LMI data (benchmark 3) were used to train ncRNAInter (virus). As shown in Table 2c, ncRNAInter (virus) achieved ACC of 0.9604, MCC of 0.9211 and AUROC of 0.9855. Although there were no methods originally designed for virus-related LMI prediction available so far to statistically compare with, the satisfying results of ncRNAInter (virus) sufficed to show the competency of our method to predict virus-related LMIs.

Table 3. Examples of new LMIs identified in this study associated with various diseases which were experimentally validated in publications

Disease	lncRNA (gene name)	miRNA	Reference
Alzheimer's disease	ENST00000501122.2 (NEAT1)	miR-16	[90]
	ENST00000501122.2 (NEAT1)	miR-195	
	ENST00000501122.2 (NEAT1)	miR-15a	
	ENST00000425595.5 (HOTAIR)	miR-107	
	ENST00000453875.5 (HOTAIR)		
	ENST00000554988.1 (Rpph1)	miR-122	
Parkinson's disease	ENST00000537068.5 (SNHG1)	miR-7	
	ENST00000537925.5 (SNHG1)		
	ENST00000545688.5 (SNHG1)		
	ENST00000540865.5 (SNHG1)		
Multiple sclerosis	ENST00000425595.5 (HOTAIR)	miR-136-5p	
	ENST00000453875.5 (HOTAIR)		
Colorectal cancer	ENST00000602587.5 (XIST)	miR-133a-3p	[95]
	ENST00000417942.5 (XIST)		
	ENST00000437681.1 (SNHG3)	miR-539	[96]
Breast cancer	ENST00000449469.5 (SNHG17)	miR-124-3p	[97]
	ENST00000424235.1 (SNHG17)		
	ENST00000414142.5 (SNHG17)		
Gallbladder cancer	ENST00000411861.5 (H19)	miR-342-3p	[98]
	ENST00000412788.5 (H19)		
	ENST00000436715.5 (H19)		
Prostate cancer	ENST00000453875.5 (HOTAIR)	miR-193a	[99]
	ENST00000425595.5 (HOTAIR)		
	ENST00000424518.5 (HOTAIR)		
Ovarian cancer	ENST00000456876.1 (HOXD-AS1)	miR-186-5p	[100]
	ENST00000436126.5 (HOXD-AS1)		
	ENST00000425005.5 (HOXD-AS1)		
Endometrial cancer	ENST00000501122.2 (NEAT1)	miR-361	[101]
Cholangiocarcinoma	ENST00000424518.5 (HOTAIR)	miR-204-5p	[102]
	ENST00000425595.5 (HOTAIR)		
	ENST00000453875.5 (HOTAIR)		
	ENST00000427868.6 (LINC00665)	miR-424-5p	[103]
	ENST00000590622.5 (LINC00665)		

All in all, our method proved its superiority and universality to conduct LMI prediction for various species, greatly broadening its application scenarios for LMI prediction.

Discovery of new LMIs from various disease benchmarks

The primary goal of our method is to discover new LMIs that were previously unknown or mistakenly categorized and then guide further downstream studies. Thus, we carried out two analyses based on benchmark 1 to predict unknown interactions involved in neurodegenerative diseases (NDDs) and cancers using well-trained ncRNAInter.

Moreno-García *et al.* [90] had reviewed the NDD-associated miRNA–ceRNAs networks experimentally validated to date and reported 74 related lncRNA/miRNA axes, where we screened out 13 axes whose information was excluded in our original training and validation datasets, as illustrated in Supplementary Table 3, available online at http://bib.oxfordjournals.org/. These axes were appraised by the well-trained ncRNAInter. As a result, the model identified 12 new LMIs, including NEAT1/miR-16, NEAT1/miR-195, NEAT1/miR-15a, HOTAIR/miR-107, Rpph1/miR-122 in Alzheimer's disease (AD) [91, 92], SNHG1/miR-7 in Parkinson's disease (PD) [93] and HOTAIR/miR-136-5p in multiple sclerosis (MS) [94], as illustrated in Table 3. In addition, as illustrated in Supplementary Table 3, available online at http://bib.oxfordjournals.org/, 22 cancer-associated lncRNA/miRNA axes whose information was excluded in our original training and validation datasets were obtained through a preliminary literature survey. These experimentally validated axes were appraised by the well-trained ncRNAInter. As a result, the model identified 21 new LMIs, including XIST/miR-133a-3p and SNHG3/miR-539 in colorectal cancer [95, 96], SNHG17/miR-124-3p in breast cancer [97], H19/miR-342-3p in gallbladder cancer [98], HOTAIR/miR-193a in prostate cancer [99], HOXD-AS1/miR-186-5p in ovarian cancer [100], NEAT1/miR361 in endometrial cancer [101], HOTAIR/miR-204-5p and LINC00665/miR-424-5p in cholangiocarcinoma [102, 103], as illustrated in Table 3.

Through these analyses, ncRNAInter has proved its effectiveness to identify LMIs associated with various major diseases. Moreover, we took a further step to evaluate the ability of ncRNAInter to predict unexplored LMIs associated with cancer by conducting another practical application. We organized the data of the 20 different miRNAs included in the 22 cancer-related lncRNA/miRNA axes mentioned above to predict all potential LMIs of the 20 miRNAs. As a result, excluding 1950 LMIs originally existing in the datasets, 2422 novel interactions were predicted as positive interactions out of a total of 29 980 potential LMI pairs, which implies there are a great amount of unknown LMIs to be experimentally validated.

All in all, by predicting the potential LMIs, ncRNAInter is able to identify the most possible interactions, helping researchers sort out candidates to perform further experimental validations, which substantially reduces their original workload. In this case, our method will provide insights and draw directions to discover novel LMIs.

Conclusion

In this study, we proposed a novel method, ncRNAInter, which was applied to identify new LMIs. It proposed an innovative comprehensive strategy for RNA representation and utilized graph neural network algorithm to conduct feature propagation and aggregation, contributing to LMI prediction. ncRNAInter achieved superior performance and robustness compared with existing methods, meanwhile ensuring the credibility and rationality of its algorithm. Moreover, ncRNAInter has universal applicability in various species, which is considered to have the potential to play a role in numerous application scenarios of LMI prediction. In addition, a certain number of novel LMIs associated with various major diseases were successfully identified, which verified its effectiveness and usability. Nevertheless, limited by the lack of experimentally validated data, future works of the experimental identification on unknown LMIs need to be put on schedule. In all, ncRNAInter, with its outstanding performance and broad applicability, has the potential to predict possible LMIs, giving insights into future research.

Data availability

The implemented code and experimental dataset are available online at https://github.com/idrblab/ncRNAInter.

Author Contributions

F.Z. and H.L. conceived the idea and supervised the work. H.Z. performed the research and wrote the scripts. H.Z., Y.W., Z.P., X.S., M.M., B.Z. and Z.L. prepared and analyzed the data. F.Z., H. L and H.Z. wrote the manuscript. All authors reviewed and approved the manuscript.

Key Points

- This study constructed a new method named ncRNAInter to discover novel LMIs based on DL algorithm of GNN, which incorporated information leak avoidance and non-linear updating.
- This study integrated, for the first time, a comprehensive strategy for RNA representation to predict LMIs, which considered the sequence, physicochemical and structural properties.
- The ncRNAInter is robust and credible, which demonstrates a significantly enhanced performance in LMI prediction compared with other state-of-the-art methods.
- The ncRNAInter proves to have universal applicability in various species and various diseases, which greatly extends its application scenarios and gives insights into future research.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Natural Science Foundation of China (81872798 and U1909208); Natural Science Foundation of Zhejiang Province (LR21H300001); Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); 'Double Top-Class' University Project (181201*194232101); Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University.

References

1. Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov* 2017;**16**:203–22.
2. Cheng L, Wang P, Tian R, *et al.* LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res* 2019;**47**:D140–4.
3. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 2014;**157**:77–94.
4. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011;**12**:861–74.
5. Tian L, Wang S-L. Exploring miRNA sponge networks of breast cancer by combining miRNA-disease-lncRNA and miRNA-target networks. *Curr Bioinform* 2021;**16**:385–94.
6. Salmena L, Poliseno L, Tay Y, *et al.* A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? *Cell* 2011;**146**:353–8.
7. Zhang Y, Tao Y, Liao Q. Long noncoding RNA: a crosslink in biological regulatory network. *Brief Bioinform* 2018;**19**:930–45.
8. Wang N, Li Y, Liu S, *et al.* Bioinformatics analysis and validation of differentially expressed MicroRNAs with their target genes involved in GLP-1RA facilitated osteogenesis. *Curr Bioinform* 2021;**16**:928–42.
9. Liu H, Deng H, Zhao Y, *et al.* LncRNA XIST/miR-34a axis modulates the cell proliferation and tumor growth of thyroid cancer through MET-PI3K-AKT signaling. *J Exp Clin Cancer Res* 2018;**37**:279.
10. Wang Y, Yang L, Chen T, *et al.* A novel lncRNA MCM3AP-AS1 promotes the growth of hepatocellular carcinoma by targeting miR-194-5p/FOXA1 axis. *Mol Cancer* 2019;**18**:28.
11. Chen W, Li Q, Zhang G, *et al.* LncRNA HOXA-AS3 promotes the malignancy of glioblastoma through regulating miR-455-5p/USP3 axis. *J Cell Mol Med* 2020;**24**:11755–67.
12. He Y, Jiang X, Duan L, *et al.* LncRNA PKMYT1AR promotes cancer stem cell maintenance in non-small cell lung cancer via activating Wnt signaling pathway. *Mol Cancer* 2021;**20**:156.
13. Tang W, Wan S, Yang Z, *et al.* Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
14. Yang H, Qi C, Li B, *et al.* Non-coding RNAs as novel biomarkers in cancer drug resistance. *Curr Med Chem* 2022;**29**:837–48.
15. Matsui M, Corey DR. Non-coding RNAs as drug targets. *Nat Rev Drug Discov* 2017;**16**:167–79.
16. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA

using biological interaction networks. *Brief Bioinform* 2016;**17**:193–203.
17. Fan M, Li X, Jiang W, *et al.* A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. *Exp Ther Med* 2013;**5**:1143–6.
18. Ni W, Zhang Y, Zhan Z, *et al.* A novel lncRNA uc.134 represses hepatocellular carcinoma progression by inhibiting CUL4A-mediated ubiquitination of LATS1. *J Hematol Oncol* 2017;**10**:91.
19. Zhang X, Wang S, Wang H, *et al.* Circular RNA circNRIP1 acts as a microRNA-149-5p sponge to promote gastric cancer progression via the AKT1/mTOR pathway. *Mol Cancer* 2019;**18**:20.
20. Zhang M, Weng W, Zhang Q, *et al.* The lncRNA NEAT1 activates Wnt/β-catenin signaling and promotes colorectal cancer progression via interacting with DDX5. *J Hematol Oncol* 2018;**11**:113.
21. Veneziano D, Marceca GP, Di Bella S, *et al.* Investigating miRNA-lncRNA interactions: computational tools and resources. *Methods Mol Biol* 2019;**1970**:251–77.
22. Li JH, Liu S, Zhou H, *et al.* starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;**42**:D92–7.
23. Pan Z, Zhou S, Zou H, *et al.* MCNN: multiple convolutional neural networks for RNA-protein binding sites prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**PP**:1. https://doi.org/10.1109/TCBB.2022.3170367.
24. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.
25. Chen L, Heikkinen L, Wang C, *et al.* Trends in the development of miRNA bioinformatics tools. *Brief Bioinform* 2019;**20**:1836–52.
26. Cui F, Zhou M, Zou Q. Computational biology and chemistry special section editorial: computational analyses for miRNA. *Comput Biol Chem* 2021;**91**:107448.
27. Shen ZA, Luo T, Zhou YK, *et al.* NPI-GNN: predicting ncRNA-protein interactions with deep graph neural networks. *Brief Bioinform* 2021;**22**:bbab051.
28. Hu L, Xu Z, Hu B, *et al.* COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res* 2017;**45**:e2.
29. Rincón-Riveros A, Morales D, Rodríguez JA, *et al.* Bioinformatic tools for the analysis and prediction of ncRNA interactions. *Int J Mol Sci* 2021;**22**:11397.
30. Chen X, Yan CC, Zhang X, *et al.* Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017;**18**:558–76.
31. Huang ZA, Huang YA, You ZH, *et al.* Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph. *BMC Med Genomics* 2018;**11**:113.
32. Yang S, Wang Y, Lin Y, *et al.* LncMirNet: predicting LncRNA-miRNA interaction based on deep learning of ribonucleic acid sequences. *Molecules* 2020;**25**:4372.
33. Kang Q, Meng J, Cui J, *et al.* PmliPred: a method based on hybrid model and fuzzy decision for plant miRNA-lncRNA interaction prediction. *Bioinformatics* 2020;**36**:2986–92.
34. Zhang P, Meng J, Luan Y, *et al.* Plant miRNA-lncRNA interaction prediction with the ensemble of CNN and IndRNN. *Interdiscip Sci* 2020;**12**:82–9.
35. Yu X, Jiang L, Jin S, *et al.* preMLI: a pre-trained method to uncover microRNA-lncRNA potential interactions. *Brief Bioinform* 2022;**23**:bbab470.
36. Huang YA, Chan KCC, You ZH. Constructing prediction models from expression profiles for large scale lncRNA-miRNA interaction profiling. *Bioinformatics* 2018;**34**:812–9.
37. Zhang W, Tang G, Zhou S, *et al.* LncRNA-miRNA interaction prediction through sequence-derived linear neighborhood propagation method with information combination. *BMC Genomics* 2019;**20**:946.
38. Huang YA, Huang ZA, You ZH, *et al.* Predicting lncRNA-miRNA interaction via graph convolution auto-encoder. *Front Genet* 2019;**10**:758.
39. Wong L, Huang YA, You ZH, *et al.* LNRLMI: linear neighbour representation for predicting lncRNA-miRNA interactions. *J Cell Mol Med* 2020;**24**:79–87.
40. Hu P, Huang YA, Chan KCC, *et al.* Learning multimodal networks from heterogeneous data for prediction of lncRNA-miRNA interactions. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:1516–24.
41. Zhao C, Qiu Y, Zhou S, *et al.* Graph embedding ensemble methods based on the heterogeneous network for lncRNA-miRNA interaction prediction. *BMC Genomics* 2020;**21**:867.
42. Zhang L, Liu T, Chen H, *et al.* Predicting lncRNA-miRNA interactions based on interactome network and graphlet interaction. *Genomics* 2021;**113**:874–80.
43. Wang XW, Liu CX, Chen LL, *et al.* RNA structure probing uncovers RNA structure-dependent biological functions. *Nat Chem Biol* 2021;**17**:755–66.
44. Fu T, Li F, Zhang Y, *et al.* VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res* 2022;**50**:D1417–31.
45. Miao YR, Liu W, Zhang Q, *et al.* lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res* 2018;**46**:D276–80.
46. Howe KL, Achuthan P, Allen J, *et al.* Ensembl 2021. *Nucleic Acids Res* 2021;**49**:D884–91.
47. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62.
48. Frankish A, Diekhans M, Jungreis I, *et al.* GENCODE 2021. *Nucleic Acids Res* 2021;**49**:D916–23.
49. Lee BT, Barber GP, Benet-Pagès A, *et al.* The UCSC genome browser database: 2022 update. *Nucleic Acids Res* 2022;**50**:D1115–22.
50. Kang YJ, Yang DC, Kong L, *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* 2017;**45**:W12–6.
51. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
52. Tong X, Liu S. CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res* 2019;**47**:e43.
53. Cheng J, Lin Y, Xu L, *et al.* ViRBase v3.0: a virus and host ncRNA-associated interaction repository with increased coverage and annotation. *Nucleic Acids Res* 2022;**50**:D928–33.
54. Bonidia RP, Sampaio LDH, Domingues DS, *et al.* Feature extraction approaches for biological sequences: a comparative study of mathematical features. *Brief Bioinform* 2021;**22**:bbab011.
55. Han S, Liang Y, Ma Q, *et al.* LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief Bioinform* 2019;**20**:2009–27.
56. Chantsalnyam T, Siraj A, Tayara H, *et al.* ncRDense: a novel computational approach for classification of non-coding RNA family by deep learning. *Genomics* 2021;**113**:3030–8.
57. Kong L, Zhang Y, Ye ZQ, *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007;**35**:W345–9.

58. Wang L, Park HJ, Dasari S, *et al.* CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013;**41**:e74.
59. Yang C, Yang L, Zhou M, *et al.* LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* 2018;**34**:3825–34.
60. Chen Z, Zhao P, Li C, *et al.* iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res* 2021;**49**:e60.
61. Chen Z, Zhao P, Li F, *et al.* iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**:1047–57.
62. Li J, Chen X, Huang Q, *et al.* Seq-SymRF: a random forest model predicts potential miRNA-disease associations based on information of sequences and clinical symptoms. *Sci Rep* 2020;**10**:17901.
63. Kang Q, Meng J, Shi W, *et al.* Ensemble deep learning based on multi-level information enhancement and greedy fuzzy decision for plant miRNA-lncRNA interaction prediction. *Interdiscip Sci* 2021;**13**:603–14.
64. Yang S, Wang Y, Zhang S, *et al.* NCResNet: noncoding ribonucleic acid prediction based on a deep resident network of ribonucleic acid sequences. *Front Genet* 2020;**11**:90.
65. Zou J, Huss M, Abid A, *et al.* A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12–8.
66. Greener JG, Kandathil SM, Moffat L, *et al.* A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 2022;**23**:40–55.
67. Wang X, Li F, Xu J, *et al.* ASPIRER: a new computational approach for identifying non-classical secreted proteins based on deep learning. *Brief Bioinform* 2022;**23**:bbac031.
68. Zhu Y, Li F, Xiang D, *et al.* Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. *Brief Bioinform* 2021;**22**:bbaa299.
69. Cheng L, Hu Y, Sun J, *et al.* DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 2018;**34**:1953–6.
70. Lv Z, Wang P, Zou Q, *et al.* Identification of sub-Golgi protein localization by use of deep representation learning features. *Bioinformatics* 2020;**36**:5600–9.
71. Hong J, Luo Y, Mou M, *et al.* Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
72. Hong J, Luo Y, Zhang Y, *et al.* Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.
73. Li F, Chen J, Leier A, *et al.* DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* 2020;**36**:1057–65.
74. Liu Q, Chen J, Wang Y, *et al.* DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief Bioinform* 2021;**22**:bbaa124.
75. Shen H, Zhang Y, Zheng C, *et al.* A Cascade graph convolutional network for predicting protein-ligand binding affinity. *Int J Mol Sci* 2021;**22**:4023.
76. Chang J, Wang L, Meng G, *et al.* Local-aggregation graph networks. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:2874–86.
77. Li R, Yuan X, Radfar M, *et al.* Graph signal processing, graph neural network and graph learning on biological data: a systematic review. *IEEE Rev Biomed Eng* 2021;**PP**:1.
78. Wu Z, Pan S, Chen F, *et al.* A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst* 2021;**32**:4–24.
79. Shao X, Yang H, Zhuang X, *et al.* scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res* 2021;**49**:e122.
80. Wang S, Xu F, Li Y, *et al.* KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics* 2021;**37**:i418–25.
81. Bang S, Ho Jhee J, Shin H. Polypharmacy side effect prediction with enhanced interpretability based on graph feature attention network. *Bioinformatics* 2021;**37**:2955–62.
82. Zhao T, Hu Y, Peng J, *et al.* DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 2020;**36**:4466–72.
83. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst* 2017;**30**:1025–35.
84. Wen B, Zeng WF, Liao Y, *et al.* Deep learning in proteomics. *Proteomics* 2020;**20**:e1900335.
85. Isufi E, Gama F, Ribeiro A. EdgeNets:edge varying graph neural networks. *IEEE Trans Pattern Anal Mach Intell* 2021;**PP**:1. https://doi.org/10.1109/TPAMI.2021.3111054.
86. Shen WX, Zeng X, Zhu F, *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat Mach Intell* 2021;**3**:334–43.
87. Yao Q, Chen Y, Zhou X. The roles of microRNAs in epigenetic regulation. *Curr Opin Chem Biol* 2019;**51**:11–7.
88. Chen Q, Lan W, Wang J. Mining featured patterns of MiRNA interaction based on sequence and structure similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**:415–22.
89. Liao Y, Chen KH, Dong XM, *et al.* A role of pre-mir-10a coding region variant in host susceptibility to coxsackie virus-induced myocarditis. *Eur Rev Med Pharmacol Sci* 2015;**19**:3500–7.
90. Moreno-García L, López-Royo T, Calvo AC, *et al.* Competing endogenous RNA networks as biomarkers in neurodegenerative diseases. *Int J Mol Sci* 2020;**21**:9582.
91. Spreafico M, Grillo B, Rusconi F, *et al.* Multiple layers of CDK5R1 regulation in Alzheimer's disease implicate long non-coding RNAs. *Int J Mol Sci* 2018;**19**:2022.
92. Gu R, Wang L, Tang M, *et al.* LncRNA Rpph1 protects amyloid-β induced neuronal injury in SK-N-SH cells via miR-122/Wnt1 axis. *Int J Neurosci* 2020;**130**:443–53.
93. Cao B, Wang T, Qu Q, *et al.* Long noncoding RNA SNHG1 promotes neuroinflammation in Parkinson's disease via regulating miR-7/NLRP3 pathway. *Neuroscience* 2018;**388**:118–27.
94. Duan C, Liu Y, Li Y, *et al.* Sulfasalazine alters microglia phenotype by competing endogenous RNA effect of miR-136-5p and long non-coding RNA HOTAIR in cuprizone-induced demyelination. *Biochem Pharmacol* 2018;**155**:110–23.
95. Yu X, Wang D, Wang X, *et al.* CXCL12/CXCR4 promotes inflammation-driven colorectal cancer progression through activation of RhoA signaling by sponging miR-133a-3p. *J Exp Clin Cancer Res* 2019;**38**:32.
96. Dacheng W, Songhe L, Weidong J, *et al.* LncRNA SNHG3 promotes the growth and metastasis of colorectal cancer by regulating miR-539/RUNX2 axis. *Biomed Pharmacother* 2020;**125**:110039.
97. Du Y, Wei N, Hong J, *et al.* Long non-coding RNASNHG17 promotes the progression of breast cancer by sponging miR-124-3p. *Cancer Cell Int* 2020;**20**:40.
98. Wang SH, Ma F, Tang ZH, *et al.* Long non-coding RNA H19 regulates FOXM1 expression by competitively binding endoge-

nous miR-342-3p in gallbladder cancer. *J Exp Clin Cancer Res* 2016;**35**:160.

99. Ling Z, Wang X, Tao T, *et al.* Involvement of aberrantly activated HOTAIR/EZH2/miR-193a feedback loop in progression of prostate cancer. *J Exp Clin Cancer Res* 2017;**36**:159.
100. Dong S, Wang R, Wang H, *et al.* HOXD-AS1 promotes the epithelial to mesenchymal transition of ovarian cancer cells by regulating miR-186-5p and PIK3R3. *J Exp Clin Cancer Res* 2019;**38**: 1–13.
101. Dong P, Xiong Y, Yue J, *et al.* Long noncoding RNA NEAT1 drives aggressive endometrial cancer progression via miR-361-regulated networks involving STAT3 and tumor microenvironment-related genes. *J Exp Clin Cancer Res* 2019;**38**:295.
102. Lu M, Qin X, Zhou Y, *et al.* LncRNA HOTAIR suppresses cell apoptosis, autophagy and induces cell proliferation in cholangiocarcinoma by modulating the miR-204-5p/HMGB1 axis. *Biomed Pharmacother* 2020;**130**:110566.
103. Lu M, Qin X, Zhou Y, *et al.* Long non-coding RNA LINC00665 promotes gemcitabine resistance of cholangiocarcinoma cells via regulating EMT and stemness properties through miR-424-5p/BCL9L axis. *Cell Death Dis* 2021;**12**:72.