Molecular Therapy
Nucleic Acids
Original Article

Fusion of multi-source relationships and topology to infer lncRNA-protein interactions

Xinyu Zhang,[1] Mingzhe Liu,[1] Zhen Li,[2] Linlin Zhuo,[1] Xiangzheng Fu,[3] and Quan Zou[4]

[1]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China; [2]Institute of Computational Science and Technology, Guangzhou University, Guangzhou 510000, China; [3]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, China; [4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611730, China

Long non-coding RNAs (lncRNAs) are important factors involved in biological regulatory networks. Accurately predicting lncRNA-protein interactions (LPIs) is vital for clarifying lncRNA's functions and pathogenic mechanisms. Existing deep learning models have yet to yield satisfactory results in LPI prediction. Recently, graph autoencoders (GAEs) have seen rapid development, excelling in tasks like link prediction and node classification. We employed GAE technology for LPI prediction, devising the FMSRT-LPI model based on path masking and degree regression strategies and thereby achieving satisfactory outcomes. This represents the first known integration of path masking and degree regression strategies into the GAE framework for potential LPI inference. The effectiveness of our FMSRT-LPI model primarily relies on four key aspects. First, within the GAE framework, our model integrates multi-source relationships of lncRNAs and proteins with LPN's topological data. Second, the implemented masking strategy efficiently identifies LPN's key paths, reconstructs the network, and reduces the impact of redundant or incorrect data. Third, the integrated degree decoder balances degree and structural information, enhancing node representation. Fourth, the PolyLoss function we introduced is more appropriate for LPI prediction tasks. The results on multiple public datasets further demonstrate our model's potential in LPI prediction.

INTRODUCTION

Long non-coding RNA (lncRNA), defined as ncRNA molecules exceeding 200 nucleotides in length.[1] Numerous studies have established a close relationship between lncRNA and the development of certain human diseases.[2–4] Some lncRNAs are involved in biological processes like X chromosome silencing, chromatin modification, and nuclear transport, regulating gene expression across epigenetic, transcriptional, and post-transcriptional levels.[5] This influences disease-related proteins, contributing to the disease's onset and progression.[6] However, less than 1% of the identified genome comprises experimentally validated disease-related lncRNAs, whose biological functions warrant further exploration. Consequently, research on lncRNA's role in gene regulation, biological processes, and disease progression is gaining importance. lncRNA-protein interaction (LPI) prediction is a critical direction for lncRNA functional annotation and molecular structure analysis.[7] These approaches may unveil mysteries in epigenetics, address genetic queries in human biology, and aid in early disease detection.[8] Thus, developing an efficient and precise LPI prediction model is imperative.

Initially, LPI prediction primarily depended on biochemical experiments and analyses, necessitating costly equipment and substantial time. It is intuitive that lncRNAs with similar sequences might interact with identical proteins. This assumption led to sequence similarity methods using alignment algorithms like BLAST[9] or Smith-Waterman[10,11] to compute lncRNA or protein sequence similarity, thereby inferring potential LPIs from confirmed ones.[12] The gene co-expression approach leverages lncRNA co-expression data, analyzing lncRNA and protein co-expression patterns to infer potential LPIs. Common techniques involve correlation analysis,[13] cluster analysis,[14] and co-expression network construction.[15] Additionally, research indicates that the subcellular localization of lncRNA and proteins in cells can aid in LPI prediction.[16] Subcellular localization-based methods predict potential LPIs by analyzing the subcellular localization of lncRNA and proteins.[17] While these experimental and analytical methods can infer potential LPIs more accurately, they are constrained by time and financial resources. Consequently, this has fostered the application of computational methods in LPI prediction tasks.

Machine learning for association prediction initially emerged in recommendation systems. A prevalent method predicts potential user-item purchases (UTB) or rankings based on existing UTB and other relationships.[18] Ensemble learning models like CatBoost,[19] random forest,[20,21] and XGBoost,[22,23] which rely on feature splitting and multi-decision tree integration, predict user-interest items. The RWR model calculates item visits based on paths and recommends items to users based on visit

Received 18 December 2023; accepted 3 April 2024;
https://doi.org/10.1016/j.omtn.2024.102187.

Correspondence: Mingzhe Liu, School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China.
E-mail: liumz@wzut.edu.cn
Correspondence: Zhen Li, Institute of Computational Science and Technology, Guangzhou University, Guangzhou 510000, China.
E-mail: lizhen5000@gzhu.edu.cn
Correspondence: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611730, China.
E-mail: zouquan@nclab.net

similarity.[24] Likewise, these methods are applicable for potential LPI prediction. Liu et al. used matrix factorization and neighborhood information to predict potential LPIs.[25]

Additionally, network-based methods have been developed for potential LPI prediction. Zheng et al. created a protein similarity network from multi-source data like protein sequences, employing RF technology and known LPIs for potential LPI prediction (LPI-PPSN).[26] Shen et al. assessed lncRNA-protein similarity using methods like fast kernel learning (FKL), employing kernel ridge regression for potential LPI prediction (LPI-MFFKL).[27] Li et al. built a heterogeneous network from lncRNA similarity, protein similarity, and lncRNA-protein networks using RWR technology for potential LPI prediction (LPI-HN).[28] Ge et al. alternately updated lncRNA and protein node information in the LPI bipartite network, determining potential LPIs from interaction scores (LPI-BNI).[29] Xie et al. restructured the lncRNA and protein score network using LPN, lncRNA similarity, and protein similarity networks, predicting potential LPIs through second-order correlation (LPI-IBNRA).[30] Building on this, Zhou et al. employed similarity kernel fusion to integrate lncRNA and protein information for potential LPI prediction (LPI-SKF).[31] Shen et al. utilized Kronecker products to amalgamate multi-source data like lncRNA and protein similarity networks for semi-supervised potential LPI prediction (LPI-KTASLP).[32] The advent of machine learning technology has expedited LPI research, significantly reducing reliance on costly equipment. However, the performance of these models relies heavily on hand-designed features.

Deep learning technology has seen extensive use in LPI prediction research in recent years. Deep learning autonomously mines in-depth information, employing techniques like convolutional neural networks (CNNs), recurrent neural networks, autoencoders (AEs), LSTM, and others to extract lncRNA and protein sequence features and graph neural networks (GNNs) for structural or topological information. Wekesa et al. combined structural and sequence features, utilizing LSTM and attention mechanisms for potential LPI prediction (LPI-MALSTM).[33] Tian et al. employed AE technology to extract lncRNA and protein sequence features, using ensemble learning to infer potential LPIs (LPI-DF).[34] Zhang et al. utilized CNN technology for extracting lncRNA and protein sequence features, subsequently predicting potential LPIs (LPI-CNN).[35] Likewise, Zhou et al. used deep learning tools to extract features from lncRNA and protein sequences and applied the GBDT algorithm for potential LPI prediction (LPI-deepGBDT).[36] Li et al. integrated multi-source data using the FKL strategy and employed a GCN[37] to extract node representations (LPI-FKLGCN).[38] Jin et al. restructured the score matrix using a graph AE (GAE) and applied collaborative filtering to train the LPI predictor (LPI-GACF).[39] Shen et al. utilized a GCN and nuclear fusion technology to rebuild lncRNA and protein score matrices, integrating them to predict potential LPIs.[40]

Deep learning models efficiently and accurately predict potential LPIs, yet their performance is constrained by inherent limitations. Firstly, these models struggle with robust node representation extraction of lncRNA and proteins. Secondly, redundant or erroneous data in datasets often diminish model performance. We propose the FMSRT-LPI model, utilizing the GAE framework and a Bernoulli distribution-based path random masking strategy, effectively mitigating these issues and enabling accurate LPI prediction. Various experiments have corroborated the performance of the FMSRT-LPI model and the significance of its key modules. In summary, our contributions are as follows:

(1) We developed the FMSRT-LPI model for LPI prediction based on the GAE framework and a Bernoulli distribution-compliant path random masking strategy, yielding satisfactory results.
(2) Trained under the GAE framework, the FMSRT-LPI model assimilates multi-source relationships of lncRNA and protein nodes, along with topological information of LPN, learning robust node representations.
(3) We introduce a random path masking strategy aligned with the Bernoulli distribution, applied to the GAE framework, effectively minimizing redundant or erroneous information in LPNs and thus enhancing model performance.
(4) We introduced a degree decoder to improve node representation and assist in accurately reconstructing the LPN. Additionally, a new loss function, better adapted to the LPI prediction tasks, was introduced.

RESULTS

This section begins with an introduction to the relevant statistical information and evaluation metrics of the LPI datasets. The experimental results section presents a comparative analysis of the proposed model's performance against various benchmark models. Multiple parameter experiments were conducted to assess the impact of key modules and the sensitivity of model performance to various parameters. Furthermore, case studies were developed to explore several potentially valuable LPIs. In the experiments, default settings include a masking rate α of 0.4, a weight parameter β for both the edge and degree decoders of 0.006, and a GAE output node dimension of 128. Furthermore, all models partition the training and test sets at a 4:1 ratio. The study utilized a range of standard indicators as common classification tasks,[41–43] including AUC, AUPR, accuracy, precision, sensitivity, and F1-score.

Comparison with other models

Table 1 presents the performance metrics for various models applied to Zhang's dataset. Traditional machine learning algorithms, including CF and matrix factorization, exhibit suboptimal performance, potentially due to two factors. Firstly, these algorithms do not incorporate features like lncRNA and protein sequences, leading to inadequate informational input. Secondly, the limited number of known LPIs results in a high rate of false negatives and false positives. Ensemble learning approaches like XGBoost and RF, which utilize features of lncRNA and proteins, partially enhance prediction accuracy. In Zhang's dataset, the top five network-based methods demonstrate comparable efficacy to the ensemble learning approaches. The LPI-SKF[31] and LPI-MFFKL[27] models deploy kernel fusion strategies within multi-source networks, markedly enhancing performance.

Table 1. Results of multiple models on Zhang's dataset

Strategy	Method	AUC	AUPR	SEN	PRE	F1-score
Traditional	CF[17]	0.836	0.542	0.633	0.459	0.532
Ensemble learning	XGBoost[21]	0.8452	0.8213	0.6979	0.7126	0.7047
Ensemble learning	RF[19]	0.8531	0.835	0.7278	0.7236	0.7256
Ensemble learning	RWR[23]	0.826	0.581	0.566	0.535	0.55
Ensemble learning	LPI-HN[27]	0.838	0.548	0.648	0.494	0.56
Ensemble learning	LPI-BNI[28]	0.852	0.624	0.634	0.533	0.579
Network based	LPI-IBNRA[29]	0.866	0.684	0.599	0.653	0.624
Network based	LPI-KTASLP[31]	0.8686	0.6148	–	–	–
Network based	LPI-SKF[30]	0.909	0.685	0.623	0.643	0.633
Network based	LPI-MFFKL[26]	0.9063	0.695	–	–	–
Network based	LPI-MALSTM[32]	0.8517	0.8271	0.7166	0.7174	0.7165
Network based	LPI-DF[33]	0.8739	0.8562	0.7199	0.7285	0.7238
Deep learning	LPI-GACF[37]	0.936	0.822	0.669	0.832	0.742
Deep learning	LPI-FKLGCCN[36]	0.9502	0.5929	–	0.6041	0.5424
Deep learning	LPI-KCGCN[38]	0.9714	0.9216	0.6808	0.9858	0.8052
Deep learning	FMSRT-LPI	0.9896	0.9805	0.994	0.9706	0.9823

SEN, sensitivity; PRE, precision.

The deep learning models LPI-MALSTM[33] and LPI-DF[34] have attained performance on par with those of LPI-BNI,[29] LPI-IBNRA,[30] and LPI-KTASLP.[32] This can be ascribed to the deep learning models' capacity for automatic extraction of deeper features, rendering them equivalent to those utilizing multi-source network data. Additionally, the LPI-GACF,[39] LPI-FKLGCN,[38] and LPI-KCGCN[40] models integrate topological information, yielding relatively satisfactory outcomes. Building on this, the LPI-FKLGCN[38] and LPI-KCGCN[40] models incorporate kernel fusion technology to further augment their performance.

Table 2 displays all models' performance metrics on Zheng's dataset. Mirroring the earlier results, traditional machine learning algorithms exhibit the poorest performance. Multi-source network-based models perform better, and the incorporation of kernel fusion technology further enhances their performance. Deep learning models, which assimilate both topological and node-specific information, demonstrate the best performance.

The FMSRT-LPI model demonstrated superior performance on both Zhang's and Zheng's datasets. In Zhang's dataset, the FMSRT-LPI model outperformed the suboptimal LPI-KCGCN model by 1.72% in AUC, 5.89% in AUPR, 31.32% in recall, and 17.70% in F1-score. In Zheng's dataset, the FMSRT-LPI model surpassed the LPI-KCGCN model by 0.09% in AUC, 5.94% in AUPR, 24.22% in recall, and 13.18% in F1-score. The FMSRT-LPI model's precision index is marginally lower than that of the LPI-KCGCN model. The results confirm the FMSRT-LPI model's efficiency in predicting potential LPIs. The effectiveness of the FMSRT-LPI model may be due to its

Table 2. Results of multiple models on Zheng's dataset

Strategies	Methods	AUC	AUPR	SEN	PRE	F1-score
Traditional	CF[17]	0.8103	0.4267	–	0.3616	0.3222
Traditional	LPI-NRLMF[24]	0.8287	0.401	–	–	–
Traditional	LPI-PPSN[25]	0.9098	–	–	–	–
Traditional	LPI-KTASLP[31]	0.9152	0.7173	–	0.364	0.3488
Network based	RWR[23]	0.9282	0.2813	–	0.2864	0.3397
Network based	LPI-HN[27]	0.9315	0.2472	–	0.3913	0.3938
Network based	LPI-BNI[28]	0.9407	0.3336	–	–	–
Network based	LPI-MFFKL[26]	0.9669	0.7062	–	–	–
Network based	LPI-FKLGCCN[36]	0.9639	0.5212	0.1363	0.2362	–
Deep learning	LPI-KCGCN[38]	0.9907	0.9267	0.7377	0.9823	0.8426
Deep learning	FMSRT-LPI	0.9916	0.9861	0.9799	0.9609	0.9744

Table 3. Results of model using different GNN encoders (AUC %)

GNN encoders	Zhang's	Zheng's
SAGE	97.86 ± 0.34	97.45 ± 0.34
GAT	96.38 ± 0.39	97.16 ± 0.74
GCN	98.09 ± 0.41	97.69 ± 0.39

use of GAE technology, which integrates multi-source features of lncRNAs and proteins along with neighborhood topology. Employing a Bernoulli distribution to mask portions of the path can reduce the influence of superfluous or incorrect data. Furthermore, the model incorporates a degree decoder to aid the edge decoder in reconstructing the LPNs.

Impact of different GNN encoders

The FMSRT-LPI model allows for various GNN encoders, including SAGE, GAT, and GCN. To ascertain the influence of the GNN encoder on performance, experiments were conducted on Zhang's and Zheng's datasets, with results detailed in Table 3. It is evident that regardless of the chosen GNN encoder, the model yields satisfactory outcomes. However, the model exhibits marginally superior performance when employing GCN as the encoder. Thus, GCN may be selected as the default encoder when processing new datasets.

Impact of data fusion

We conducted a set of experiments to evaluate the model's performance with various data fusions. In Table 4, the FMSRT-LPI model fused sequence similarity, correlation, and expression features of lncRNA, along with sequence similarity, correlation, and Gene Ontology (GO) features of proteins, before splicing these features together. "Test1" indicates the model splices only the sequence correlation features of lncRNA and proteins. In this scenario, multi-source similarity features are not fused. "Test2" signifies the removal of sequence similarity features from the lncRNA and protein fusion features. "Test3" indicates the removal of sequence correlation features from the lncRNA and protein fusion features. "Test4" shows the removal of lncRNA expression and protein GO features from their respective fusion features.

Table 4 presents the model's performance with various data fusions. The model achieves optimal performance when fusing all lncRNA and protein features. Removing features from either source significantly reduces performance. Additionally, fusing any two features of lncRNA and protein results in a minimal performance gap. The model performs worse when only the correlation features of lncRNA and protein are inputted. This demonstrates that integrating multi-source data enhances the node representation of lncRNA and proteins, effectively improving model performance.

Ablation experiment

We conducted a series of experiments to investigate how node features and the topology of known LPNs influence model performance. In Table 5, "w/o GAE" denotes multi-source features fusing lncRNA and proteins, inputted into an MLP for training. "w/o fusion" signifies that the model eschews lncRNA and protein fusion features, instead randomly generating node embeddings. Additionally, we compare our findings with those from the LPI-KCGCN model.

Table 4. Impact of diverse data fusion on model results

Datasets	Data fusion	AUC	AUPR	SEN	PRE	F1-score
Zhang's	Test1	0.8964	0.9037	0.8634	0.8813	0.9084
Zhang's	Test2	0.9531	0.9569	0.9248	0.9406	0.9519
Zhang's	Test3	0.9594	0.9616	0.9316	0.9496	0.9635
Zhang's	Test4	0.9465	0.9461	0.9233	0.9431	0.9433
Zhang's	FMSRT-LPI	0.9896	0.9805	0.9940	0.9706	0.9823
Zheng's	Test1	0.9046	0.8976	0.8736	0.8837	0.8943
Zheng's	Test2	0.9641	0.9513	0.9458	0.9418	0.9586
Zheng's	Test3	0.9583	0.9586	0.9562	0.9454	0.9627
Zheng's	Test4	0.9479	0.9438	0.9468	0.9352	0.9487
Zheng's	FMSRT-LPI	0.9916	0.9861	0.9799	0.9609	0.9744

Table 5 displays the results of the ablation experiments. The lowest performance occurs when the model solely utilizes multi-source features of lncRNAs and proteins, disregarding known LPN topological information. Performance marginally declines when the model employs random node embeddings and leverages known LPNs for potential LPI prediction. This decline is more pronounced in Zheng's dataset. Based on the "Materials" section, Zhang's dataset has a sparsity of approximately 84.44%, and Zheng's dataset has a sparsity of approximately 94.93%. Hence, in datasets with high sparsity, topological information alone may be insufficient, but integrating lncRNA and protein node features can mitigate this issue. The LPI-KCGCN model achieves suboptimal performance, attributed to integrating multi-source features with known LPN topology. The FMSRT-LPI model outperforms others, likely due to its superior integration of multi-source features and known LPN topology within the GAE framework. Furthermore, the implemented path masking strategy mitigates the effect of noisy data, thereby enhancing model performance.

Table 5. Results of ablation experiment

Datasets	Models/metrics	AUC	AUPR	SEN	PRE	F1-score
Zhang's	w/o GAE	0.9432	0.8913	0.7578	0.9248	0.8706
Zhang's	w/o fusion	0.9814	0.9540	0.8864	0.9367	0.9465
Zhang's	LPI-KCGCN	0.9714	0.9216	0.6808	0.9858	0.8052
Zhang's	FMSRT-LPI	0.9896	0.9805	0.9940	0.9706	0.9823
Zheng's	w/o GAE	0.9540	0.8966	0.7638	0.9155	0.8619
Zheng's	w/o fusion	0.9658	0.9383	0.8934	0.9293	0.9334
Zheng's	LPI-KCGCN	0.9907	0.9267	0.7377	0.9823	0.8426
Zheng's	FMSRT-LPI	0.9916	0.9861	0.9799	0.9609	0.9744

Table 6. The lncRNAs predicted by the proposed model to be associated with proteins P53 and RABP, respectively

P53				RABP			
MIR205HG	confirmed	H19	confirmed	MALAT1	confirmed	NORAD	confirmed
NEAT1	confirmed	GAS5	confirmed	NEAT1	confirmed	H19	confirmed
P21	confirmed	MIR205HG	confirmed	HOTAIR	confirmed	BACE1-AS	confirmed
MEG3	confirmed	P53	confirmed	XIST	confirmed	HULC	confirmed
HOTAIR	confirmed	TP53TG1	confirmed	PANDAR	confirmed	UCA1	confirmed
MALAT1	confirmed	EPS	unconfirmed	ANRIL	confirmed	PCA3	unconfirmed
CCAT1	confirmed	ROR	confirmed	KCNQ1OT	confirmed	THOR	confirmed
XIST	confirmed	PINT	unconfirmed	TUG1	confirmed	FENDRR	confirmed
CCAT2	confirmed	ATB	confirmed	CCAT1	confirmed	ROR	confirmed
UCA1	confirmed	MALAT2	unconfirmed	MEG3	confirmed	MIR205HG	unconfirmed

Parameter experiment

The model's robustness was assessed by evaluating its parameter sensitivity on Zhang's and Zheng's datasets. The experiment focused on analyzing the effects of the masking ratio, decoder adjustment coefficient, and embedding size on model performance. Additionally, it offers guidance for parameter selection.

Effect of α

The masking ratio α dictates the extent to which the LPNs are masked, influencing the model's performance. Figure S1 illustrates the model's outcomes with various masking ratios α. The results indicate that the model's performance remains stable when α ranges between 0.1 and 0.5. Model performance begins to deteriorate when α exceeds 0.5. This suggests that suitable masking can mitigate the effects of redundant or incorrect data, enhancing model performance. Overmasking risks the loss of crucial information leading to a reduction in model performance. Therefore, a smaller masking ratio, such as the default 0.3, is preferable.

Effect of β

The adjustable weight parameter β specifies the degree decoder's role in reconstructing the LPNs. Figure S2 presents the model's results for various adjustment coefficients β. It is observed that the model's performance improves incrementally when β is within the range of [0, 0.006]. When β surpasses 0.006, there is a noticeable decline in model performance. This indicates that the degree decoder can effectively aid the edge decoder in reconstructing LPNs. However, an excessively high β will lead to an overemphasis on the degree decoder, increasing the proportion of redundant information. Consequently, selecting smaller adjustable parameters is advisable for efficient reconstruction of LPNs.

Effect of embedding size

Figure S3 depicts the model's performance across various embedding sizes. Evidentially, the embedding size is observed to influence the model's performance. The model performs consistently when the embedding size is set at or below 128. At embedding sizes greater than 128, the model's performance varies significantly, yet remains comparable across different datasets. Hence, a smaller embedding size, such as 128, is sufficient for the model to attain satisfactory results. This also contributes to reduced computational time and memory usage.

Case study

To assess the FMSRT-LPI model's efficacy in isolation scenarios, P53 protein[44] and poly(A)-binding protein (PABP)[45] were chosen for prediction and analysis. P53 protein serves as a tumor suppressor by inhibiting the unchecked proliferation of abnormal cells. Investigating P53 protein and its associated lncRNAs may reveal regulatory mechanisms beneficial to cancer treatment. PABP, a binding protein, typically attaches to poly(A) tails at both the 5′ and 3′ ends of RNA. PABP participates in the regulatory processes involving lncRNAs and other RNA types. Research on PABP and its related lncRNAs has uncovered potential regulatory mechanisms, which may aid in applications like drug repositioning.

Prior to model training, P53 protein, PABP, and their corresponding lncRNAs were excluded from the datasets. Subsequently, the FMSRT-LPI model was trained using the remaining data. For model prediction, P53 protein and PABP were reintroduced to the test set. The FMSRT-LPI model calculated and ranked the scores for all lncRNA-protein pairs with each of these two proteins. The top 20 lncRNAs were then selected and are documented in Table 6. The results confirmed that 17 lncRNAs, as predicted by the FMSRT-LPI model, interact with protein P53 and 18 with PABP, according to the real database.

Adopting a comparable approach, we assessed the model's performance on two lncRNAs: MALAT1[46] and XIST.[47] MALAT1, XIST, and their associated proteins were removed from the dataset prior to training the model. The remaining data were then utilized to train the FMSRT-LPI model. During prediction, MALAT1 and XIST were reintegrated into the test set. The trained FMSRT-LPI model was then used to calculate scores for lncRNA-protein pairs associated with these two lncRNAs. Subsequently, the top 20 proteins were identified and are listed in Table 7. The results revealed that, for MALAT1 and XIST, 18 and 17 of the predicted proteins,

Table 7. The proteins predicted by the proposed model to be associated with lncRNAs MALAT1 and XIST, respectively

MALAT1				XIST			
SRPK1	confirmed	CPSF1	confirmed	Lamin B1	confirmed	USP7	confirmed
SF2	confirmed	hnRNPA1	confirmed	YY1	confirmed	WTAP	confirmed
hnRNPC	confirmed	DNMT1	confirmed	CIZ1	confirmed	PCGF3	confirmed
FUS	confirmed	YB-1	confirmed	SAF-A	confirmed	PTBP1	confirmed
LMNB1	confirmed	SMAD2	confirmed	FUS	confirmed	HNRNPU	unconfirmed
EZH2	confirmed	PARP1	confirmed	hnRNP K	confirmed	CBX2	confirmed
HDAC9	confirmed	ZC3H7B	unconfirmed	SMC1A	confirmed	YB-1	unconfirmed
HuR	confirmed	ELAVL1	confirmed	RBFOX2	confirmed	EIF4A3	confirmed
P53	confirmed	HNRNPU	confirmed	SFPQ	confirmed	HDAC9	unconfirmed
ILF3	confirmed	PTBP1	unconfirmed	PSPC1	confirmed	CNOT1	confirmed

respectively, were verified in the database. This demonstrates the FMSRT-LPI model's effectiveness in identifying potential LPIs in isolation conditions.

DISCUSSION

Accurate LPI prediction aids in understanding lncRNA regulatory mechanisms, discovering novel biomarkers, and drug repositioning. While current network-based and deep learning methods more accurately identify potential LPIs, they encounter two primary challenges. First, the topological information and the information of the lncRNA and the protein itself are not well absorbed at the same time, precluding the extraction of robust node representations. Second, these methods struggle with managing redundant or erroneous data.

Consequently, our research introduces the GAE framework and a probabilistic distribution-based path masking strategy culminating in the FMSRT-LPI model designed for precise identification of potential LPIs. Operating within the GAE framework, the FMSRT-LPI model concurrently processes multi-source lncRNA and protein information and LPI network topology, extracting robust node representations. Additionally, we incorporate a Bernoulli distribution-based path masking strategy and random walk technique to obscure redundant paths in the LPI network, diminishing the impact of superfluous or incorrect data. A novel loss function is also introduced to optimize the model for LPI classification tasks. Ultimately, we conducted comparative, ablation, parameter studies, and case analyses to validate the model's performance and the significance of its key components.

Overall, the FMSRT-LPI model accurately identifies potential LPIs, enhancing understanding of lncRNA functions and disease mechanisms. Future work will focus on exploring key lncRNAs and their targets for significant diseases affecting human health, aiming to inform treatment strategies.

MATERIALS AND METHODS

Our proposed LPI prediction model, FMSRT-LPI, utilizes the GAE framework and a Bernoulli distribution-based random path masking strategy to address issues of inadequate node feature extraction and data redundancy. Operating within the GAE framework, the FMSRT-LPI model encompasses initial feature extraction, LPN construction, and random path masking, as well as GNN encoder and decoder modules. Subsequently, the model and its related modules are detailed.

Materials

Our study evaluates the performance of the proposed model alongside other comparative models using two publicly available datasets. The first dataset, compiled by Zhang et al., encompasses 4,158 LPIs sourced from the NPInter v.2.0 database.[48] This dataset includes interactions involving 27 proteins and 990 lncRNAs.[49] The second dataset, assembled by Zheng et al., comprises 4,470 LPIs featuring 84 proteins and 1,050 lncRNAs. Detailed sequence information and additional data for these interactions can be retrieved from the NONCODE[50] and SUPERFAMILY[51] databases or accessed from previously published studies (https://github.com/6gbluewind/LPI-KCGCN).

Model architecture

Figure 1 depicts the FMSRT-LPI model architecture, comprising four primary steps. Step A involves fusing the sequence, similarity, and expression information of lncRNA with the sequence, similarity, and GO information of protein, integrating existing LPIs to construct an initial LPN. In step B, a portion of the path in the LPN is masked following the Bernoulli distribution. Step C selects a GNN encoder to extract features from the masked LPN. Step D utilizes a degree decoder and edge decoder to collaboratively reconstruct the LPN, obtaining node representations within the graph. Subsequently, the inner product of the lncRNA-protein pair is computed to determine the existence of an LPI.

Initial node representations

We gather three-dimensional data encompassing sequence similarity, nucleotide sequences, and lncRNA expression alongside similar data for protein sequence similarity, nucleotide sequences, and GO. The average fusion strategy[52] is applied to amalgamate these multi-source datasets, thereby constructing the initial features of lncRNA and proteins.

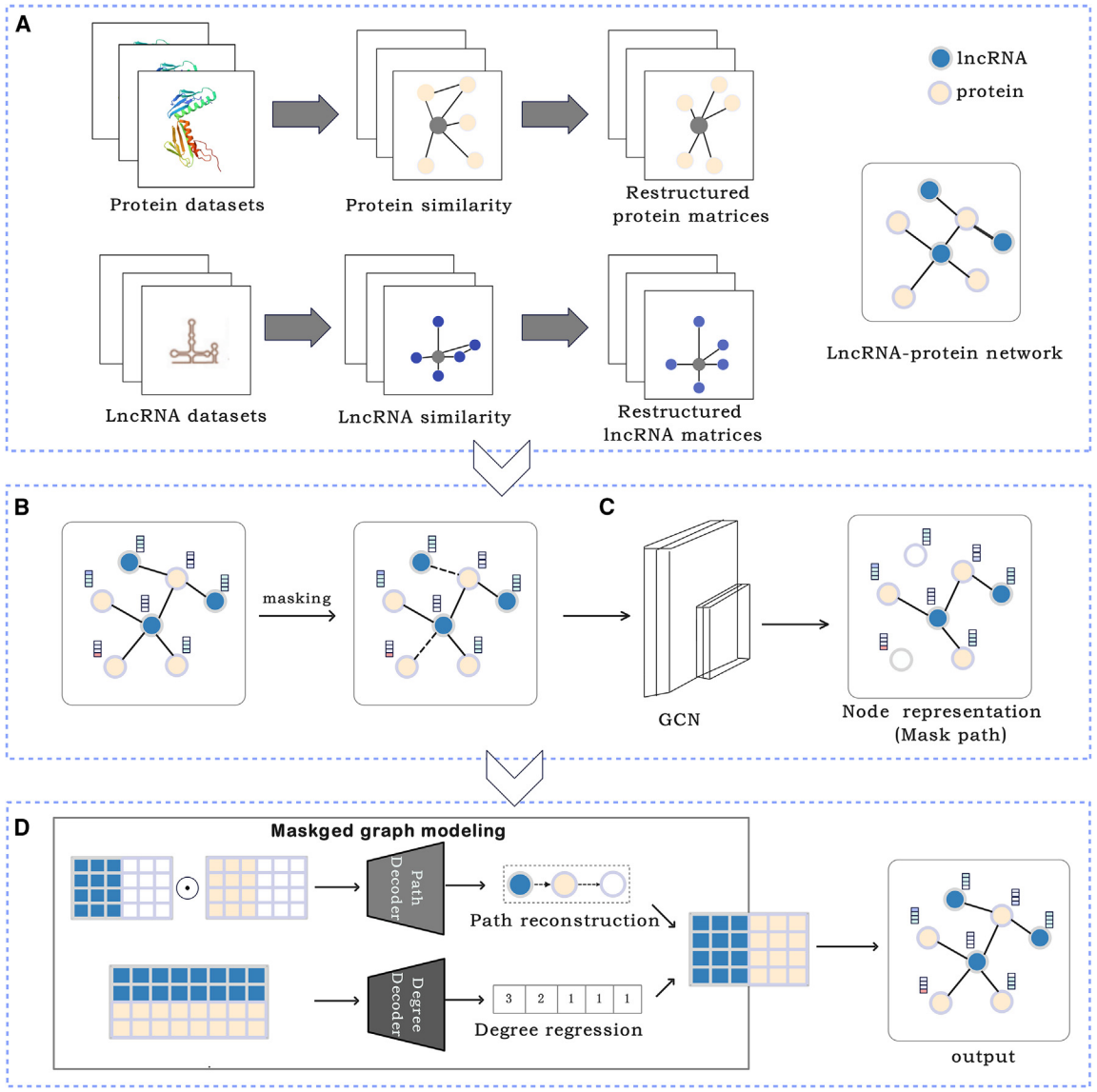

Figure 1. The FMSRT-LPI model's architecture comprises four primary modules

(A) LPN construction, (B) LPN partial masking, (C) node embedding extraction on LPN, and (D) LPN reconstruction.

Sequence similarity matrix

Assuming *sp* represents protein sequence and *sl* represents lncRNA sequence in the dataset, the sequence similarity between lncRNAs and proteins is calculated as follows, respectively:

$$S_L(l_i, l_m) = \frac{SW(sl_i, sl_m)}{\sqrt{SW(sl_i, sl_i)}\sqrt{SW(sl_m, sl_m)}}, \text{and} \quad \text{(Equation 1)}$$

$$Sp\left(p_t, p_j\right) = \frac{SW\left(sp_t, sp_j\right)}{\sqrt{SW\left(sp_t, sp_j\right)}\sqrt{SW\left(sp_t, sp_j\right)}}, \quad \text{(Equation 2)}$$

where $SW(\cdot, \cdot)$ refers to the Smith-Waterman algorithm.[10]

Sequence correlation matrix

Pse-PSSM[53] and CT[54] technologies are employed to represent protein and lncRNA sequences, respectively. The sequences of lncRNA and protein are denoted as *cl* and *cp*, respectively. Radial basis function technology is utilized to compute the relationship matrix between proteins and lncRNAs:

$$C_L(l_i, l_m) = \exp\left(-\sigma\|cl_i - cl_m\|^2\right), \text{and} \quad \text{(Equation 3)}$$

$$C_P\left(p_t, p_j\right) = \exp\left(-\sigma\|cp_t - cp_j\|^2\right). \quad \text{(Equation 4)}$$

lncRNA expression correlation matrix

lncRNA expression profile data are sourced from the NONCODE database,[50] and radial basis function technology is then applied to compute the lncRNA relationship matrix.

Protein GO correlation matrix

Protein-related GO data, gathered from the GOA database,[55] are utilized in the calculations to enhance the initial protein features.[26] Assuming that the GO data for a protein are denoted as gp, the relationship matrix of the protein is calculated using the Jaccard coefficient:

$$GC_P\left(p_t, p_j\right) = \frac{\left|gp_t \cap gp_j\right|}{\left|gp_t \cup gp_j\right|}, \quad \text{(Equation 5)}$$

where $gp_t \cap gp_j$ denotes the intersection of terms for proteins p_t and p_j and $gp_t \cup gp_j$ represents the union of terms for proteins p_t and p_j.

Fusion of multi-source relationships

Generally speaking, similarity data from a single source often cannot fully characterize lncRNA or proteins. Integrating similarity data from multiple sources provides complementary information, thereby enhancing lncRNA or protein representation. Common data fusion methods encompass summation, inner product, and Hadamard product. Additionally, multi-source data for both lncRNAs and proteins originate from similarity networks and exhibit a certain degree of homology. Therefore, our research employs simple summation averaging to fuse the multi-source similarity data of lncRNAs and proteins.

Masking strategy

Appropriate masking within a GAE can mitigate the effects of redundant or erroneous information on model performance. The common strategy masks certain LPIs based on a probability distribution, followed by reconstructing the LPN. This study introduces a masking strategy that randomly masks parts of the path in accordance with a probability distribution. Compared to the standard masking strategy, this approach of masking portions of the path effectively truncates the graph's higher-order structure, compelling the model to capture the graph's critical path. Let $G = (V_L, V_P, E)$ denote the initial LPN. $G_{existed} = (V_L, V_P, E_{existed})$ represents the remaining graph, with E_{masked} being the set of all edges on the masked path, and $V = V_L \cup V_P$, $E_{existed} = E - E_{masked}$. And we adopt the Bernoulli distribution and a random walk strategy to complete the E_{masked} collection:

$$E_{masked} = Randomly\left(S, l_{step}\right), \text{and} \quad \text{(Equation 6)}$$

$$S \sim Bernoulli(\alpha), \quad \text{(Equation 7)}$$

where S denotes the set of root nodes for the random walk, l_{step} signifies the path length, and α, ranging from [0,1], represents the masking rate. Initially, nodes are sampled on the LPN based on the Bernoulli distribution and the masking rate α. Subsequently, the masked path is determined via a random walk, and edges along this path are masked in the original LPN.

Indeed, masking edges connecting nodes of lower degrees can lead to information loss. However, the masking strategy is applied independently in each training round, with certain paths being re-masked each time. Consequently, sparse nodes experience minimal impact throughout the entire training process. As a result, the model's performance remains unaffected.

GNN encoder

GNNs are capable of mining the structural information of LPNs and extracting robust node representations. In this research, GCN,[37] GraphSAGE,[56] GIN,[57] and GAT[58] are selected as encoders to delve into the structural information of LPN. Using GCN as an example, message propagation is conducted on the LPN, aggregating and updating node representations of lncRNA and protein:

$$Z^l = \sigma\left(\widehat{A}Z^{l-1}W^l\right), \quad \text{(Equation 8)}$$

where Z^l represents the node representation matrix on the l-th GCN layer and W^l represents the corresponding trainable matrix. $\sigma(\cdot)$ represents the sigmoid activation function. $\widehat{A}$ represents the normalized lncRNA-protein adjacency matrix, which can be calculated as

$$\widehat{A} = Diag(A+E)^{-\frac{1}{2}} \cdot A \cdot Diag(A+E)^{-\frac{1}{2}}, \quad \text{(Equation 9)}$$

where A represents the original lncRNA-protein adjacency matrix. $(A + E)$ is defined as the adjacency matrix with self-loops, and $Diag(A + E)$ is defined as the degree matrix of $(A + E)$.

Dual decoders

We introduce a decoder architecture where edge decoders and degree decoders work together. The edge decoder is the most important part, and its purpose is to reconstruct the missing LPIs in LPN by calculating the scores of lncRNA-protein pairs. The edge decoder can be defined as

$$h\left(z_l, z_p\right) = \sigma\left(MLP\left(z_l \circ z_p\right)\right), \quad \text{(Equation 10)}$$

where z_l and z_p symbolize the node representations of lncRNA and protein as extracted by the GNN encoder, respectively. The above equation outlines that, initially, the Hadamard product of the lncRNA-protein pair is computed, followed by its input into the MLP, and ultimately, the sigmoid function calculates the pair's score. The inference process requires only the calculation of lncRNA-protein pair prediction scores using the trained model to classify these pairs.

Additionally, the degree decoder compares node information pre- and post-masking, compelling the training model to align closely with the pre-masking topology. The degree decoder is defined as follows:

$$g_{\varphi}(d_i) = MLP(d_i), \quad \text{(Equation 11)}$$

where d_i denotes the degree of the node i.

Optimization

The FMSRT-LPI model primarily reconstructs the LRN via collaborative efforts of the edge and degree decoders. Consequently, the model's optimization objective comprises two components: (1) the loss incurred by the edge decoder and (2) the regression loss of the degree decoder. To tailor the model for LPI prediction tasks, the PolyLoss function[59] is employed to calculate the loss for the edge decoder:

$$L_{edge} = -\log(P_{label}) + \eta(1 - P_{label}), \quad \text{(Equation 12)}$$

where P_{label} represents the predicted probability of the true label of lnRNA-protein pairs.

For the degree decoder, we utilize mean-squared error to compute the loss between a node's degree in the masked graph G_{masked} and its predicted degree:

$$L_{\deg\ ree} = \frac{1}{|V|}\sum_{i \in v}\|g_{\varphi}(d_i) - \deg ree_{mask}(v)\|_F^2, \quad \text{(Equation 13)}$$

where $degree_{mask}(v)$ denotes the degree of node v in the masked graph G_{masked}.

Consequently, the optimization objective is defined as

$$L = L_{edge} + \beta L_{\deg\ ree}, \quad \text{(Equation 14)}$$

where β symbolizes an adjustable weight.

DATA AND CODE AVAILABILITY

Our code and data are publicly available in the GitHub repository: https://github.com/2014402680/FMSRT-LPI.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2024.102187.

ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (nos. 62002111 and 62372158).

AUTHOR CONTRIBUTIONS

X.Z. was responsible for conducting experiments and writing the first draft, M.L. was responsible for experimental guidance and manuscript revision, X.F. and L.Z. were responsible for conducting ablation experiments and checking the grammar, and Z.L. and Q.Z. were responsible for manuscript revision.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Schaukowitch, K., and Kim, T.K. (2014). Emerging epigenetic mechanisms of long non-coding RNAs. Neuroscience *264*, 25–38.
2. Ma, Y. (2022). DeepMNE: deep multi-network embedding for lncRNA-disease association prediction. IEEE J. Biomed. Health Inform. *26*, 3539–3549.
3. Li, C.H., and Chen, Y. (2013). Targeting long non-coding RNAs in cancers: progress and prospects. Int. J. Biochem. Cell Biol. *45*, 1895–1910.
4. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118.
5. Dhanasekaran, K., Kumari, S., and Kanduri, C. (2013). Noncoding RNAs in chromatin organization and transcription regulation: an epigenetic view. Subcell. Biochem. *61*, 343–372.
6. Ferrè, F., Colantoni, A., and Helmer-Citterich, M. (2016). Revealing protein–lncRNA interaction. Brief. Bioinform. *17*, 106–116.
7. Smith, M.A., and Mattick, J.S. (2017). Structural and functional annotation of long noncoding RNAs. Bioinformatics *1526*, 65–85.
8. Li, X., Wu, Z., Fu, X., and Han, W. (2014). lncRNAs: insights into their function and mechanics in underlying disorders. Mutat. Res. Rev. Mutat. Res. *762*, 1–21.
9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403–410.
10. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.
11. Reichert, T.A., Cohen, D.N., and Wong, A.K. (1973). An application of information theory to genetic mutations and the matching of polypeptide sequences. J. Theor. Biol. *42*, 245–261.
12. Zhao, J., Sun, J., Shuai, S.C., Zhao, Q., and Shuai, J. (2023). Predicting potential interactions between lncRNAs and proteins via combined graph auto-encoder methods. Brief. Bioinform. *24*, bbac527.
13. Guo, X., Chang, Q., Pei, H., Sun, X., Qian, X., Tian, C., and Lin, H. (2017). Long noncoding RNA–mRNA correlation analysis reveals the potential role of HOTAIR in pathogenesis of sporadic thoracic aortic aneurysm. Eur. J. Vasc. Endovasc. Surg. *54*, 303–314.
14. Yuan, Q., Guo, X., Ren, Y., Wen, X., and Gao, L. (2020). Cluster correlation based method for lncRNA-disease association prediction. BMC Bioinf. *21*, 180–214.
15. Wang, P., Fu, H., Cui, J., and Chen, X. (2016). Differential lncRNA-mRNA co-expression network analysis revealing the potential regulatory roles of lncRNAs in myocardial infarction. Mol. Med. Rep. *13*, 1195–1203.
16. Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. PLoS Comput. Biol. *14*, e1006616.
17. Yang, R., Gao, S., Fu, Y., and Zhang, L. lncSLP: An Ensemble Method with Multi-Source Sequence Descriptors to Predict lncRNA Subcellular Localizations from Imbalanced Data. Available at SSRN 4515036.
18. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In 10th International Conference on World Wide Web.
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. *31*.
20. Pal, M. (2005). Random forest classifier for remote sensing classification. Int. J. Rem. Sens. *26*, 217–222.
21. Hasan, M.M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020). HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics *36*, 3350–3356.
22. Chen, T., and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In 22nd ACM International Conference on Knowledge Discovery and Data Mining.

23. Hasan, M.M., Basith, S., Khatun, M.S., Lee, G., Manavalan, B., and Kurata, H. (2021). Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief. Bioinform. *22*, bbaa202.

24. Gan, M. (2014). Walking on a user similarity network towards personalized recommendations. PLoS One *9*, e114662.

25. Yan, C., Wang, J., Ni, P., Lan, W., Wu, F.-X., and Pan, Y. (2019). DNRLMF-MDA: predicting microRNA-disease associations based on similarities of microRNAs and diseases. IEEE/ACM Trans. Comput. Biol. Bioinform. *16*, 233–243.

26. Zheng, X., Wang, Y., Tian, K., Zhou, J., Guan, J., Luo, L., and Zhou, S. (2017). Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. BMC Bioinf. *18*, 420–518.

27. Shen, C., Ding, Y., Tang, J., and Guo, F. (2018). Multivariate information fusion with fast kernel learning to kernel ridge regression in predicting LncRNA-protein interactions. Front. Genet. *9*, 716.

28. 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.

29. Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding RNA–protein interactions. Dev. Reprod. Biol. *14*, 62–71.

30. Xie, G., Wu, C., Sun, Y., Fan, Z., and Liu, J. (2019). Lpi-ibnra: Long non-coding rna-protein interaction prediction based on improved bipartite network recommender algorithm. Front. Genet. *10*, 343.

31. Zhou, Y.-K., Hu, J., Shen, Z.-A., Zhang, W.-Y., and Du, P.-F. (2020). LPI-SKF: predicting lncRNA-protein interactions using similarity kernel fusions. Front. Genet. *11*, 615144.

32. Shen, C., Ding, Y., Tang, J., Jiang, L., and Guo, F. (2019). LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. IEEE Access *7*, 13486–13496.

33. Wekesa, J.S., Meng, J., and Luan, Y. (2020). Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction. Genomics *112*, 2928–2936.

34. Tian, X., Shen, L., Wang, Z., Zhou, L., and Peng, L. (2021). A novel lncRNA–protein interaction prediction method based on deep forest with cascade forest structure. Sci. Rep. *11*, 18881.

35. Zhang, S.-W., Zhang, X.-X., Fan, X.-N., and Li, W.-N. (2020). LPI-CNNCP: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. Anal. Biochem. *601*, 113767.

36. Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. BMC Bioinf. *22*, 479–524.

37. Kipf, T.N., and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.

38. Li, W., Wang, S., and Guo, H. (2021). LPI-FKLGCN: Predicting LncRNA-Protein Interactions Through Fast Kernel Learning and Graph Convolutional Network. In International Symposium on Bioinformatics Research and Applications.

39. Jin, C., Shi, Z., Zhang, H., and Yin, Y. (2021). Predicting lncRNA-protein interactions based on graph autoencoders and collaborative training. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

40. Shen, C., Mao, D., Tang, J., Liao, Z., and Chen, S. (2024). Prediction of LncRNA-Protein Interactions Based on Kernel Combinations and Graph Convolutional Networks. IEEE J. Biomed. Health Inform. *28*, 1937–1948.

41. Wang, R., Zhou, Z., Wu, X., Jiang, X., Zhuo, L., Liu, M., Li, H., Fu, X., and Yao, X. (2023). An effective plant small secretory peptide recognition model based on feature correction strategy. J. Chem. Inf. Model. *64*, 2798–2806.

42. Zhou, Z., Zhuo, L., Fu, X., Lv, J., Zou, Q., and Qi, R. (2024). Joint masking and self-supervised strategies for inferring small molecule-miRNA associations. Mol. Ther. Nucleic Acids *35*, 102103.

43. Zhuo, L., Wang, R., Fu, X., and Yao, X. (2023). StableDNAm: towards a stable and efficient model for predicting DNA methylation based on adaptive feature correction learning. BMC Genom. *24*, 742.

44. Hassin, O., and Oren, M. (2023). Drugging p53 in cancer: one protein, many targets. Nat. Rev. Drug Discov. *22*, 127–144.

45. Schäfer, I.B., Yamashita, M., Schuller, J.M., Schüssler, S., Reichelt, P., Strauss, M., and Conti, E. (2019). Molecular basis for poly (A) RNP architecture and recognition by the Pan2-Pan3 deadenylase. Cell *177*, 1619–1631.e21.

46. Kim, J., Piao, H.-L., Kim, B.-J., Yao, F., Han, Z., Wang, Y., Xiao, Z., Siverly, A.N., Lawhon, S.E., Ton, B.N., et al. (2018). Long noncoding RNA MALAT1 suppresses breast cancer metastasis. Nat. Genet. *50*, 1705–1715.

47. Chen, C.-K., Blanco, M., Jackson, C., Aznauryan, E., Ollikainen, N., Surka, C., Chow, A., Cerase, A., McDonel, P., and Guttman, M. (2016). Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. Science *354*, 468–472.

48. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. *313*, 903–919.

49. 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.

50. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, D98–D103.

51. 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.

52. He, J., Chang, S.-F., and Xie, L. (2008). Fast kernel learning for spatial pyramid matching. In 2008 IEEE Conference on Computer Vision and Pattern Recognition.

53. Chou, K.-C., and Shen, H.-B. (2007). MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Commun. *360*, 339–345.

54. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA *104*, 4337–4341.

55. Wan, S., Mak, M.-W., and Kung, S.-Y. (2013). GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. J. Theor. Biol. *323*, 40–48.

56. Hamilton, W.L., Ying, Z., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA.

57. Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How Powerful are Graph Neural Networks? In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.

58. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

59. Leng, Z., Tan, M., Liu, C., Cubuk, E.D., Shi, X., Cheng, S., and Anguelov, D.P. (2022). A Polynomial Expansion Perspective of Classification Loss Functions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.12511.