ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Predicting lncRNA–Disease Associations Based on a Dual-Path Feature Extraction Network with Multiple Sources of Information Integration

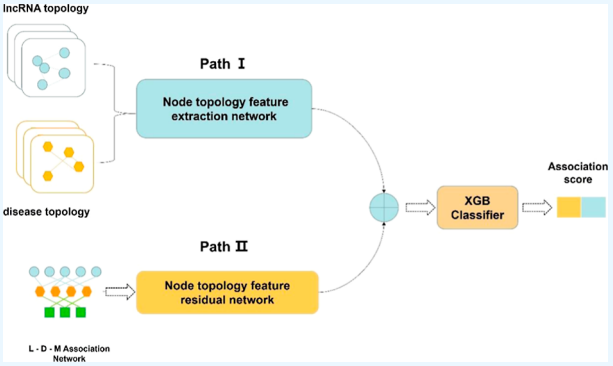
Dengju Yao,* Binbin Zhang, Xiaojuan Zhan, Bo Zhang, and Xiang Kui Li

Cite This: *ACS Omega* 2024, 9, 35100–35112

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identifying the associations between long noncoding RNAs (lncRNAs) and disease is critical for disease prevention, diagnosis and treatment. However, conducting wet experiments to discover these associations is time-consuming and costly. Therefore, computational modeling for predicting lncRNA-disease associations (LDAs) has become an important alternative. To enhance the accuracy of LDAs prediction and alleviate the issue of node feature oversmoothing when exploring the potential features of nodes using graph neural networks, we introduce DPFELDA, a dual-path feature extraction network that leverages the integration of information from multiple sources to predict LDA. Initially, we establish a dual-view structure of lncRNAs and disease and a heterogeneous network of lncRNA-disease-microRNA (miRNA) interactions. Subsequently, features are extracted using a dual-path feature extraction network. In particular, we employ a combination of a graph convolutional network, a convolutional block attention module, and a node aggregation layer to perform multilayer topology feature extraction for the dual-view structure of lncRNAs and diseases. Additionally, we utilize a Transformer model to construct the node topology feature residual network for obtaining node-specific features in heterogeneous networks. Finally, XGBoost is employed for LDA prediction. The experimental results demonstrate that DPFELDA outperforms the benchmark model on various benchmark data sets. In the course of model exploration, it becomes evident that DPFELDA successfully alleviates the issue of node feature oversmoothing induced by graph-based learning. Ablation experiments confirm the effectiveness of the innovative module, and a case study substantiates the accuracy of DPFELDA model in predicting novel LDAs for characteristic diseases.



1. INTRODUCTION

Long noncoding RNAs (lncRNAs) are a class of RNAs that are more than 200 nucleotides in length and that cannot encode proteins.[1] lncRNAs are usually expressed at lower levels than microRNAs. However, they play very important roles in life-regulating activities, including transcriptional regulation, translational regulation, RNA processing regulation and cell life cycle regulation.[2,3] Recently, it has been found that lncRNAs are closely related to disease pathogenesis,[4] so the detection of more disease-associated lncRNAs may be a good way to understand the occurrence, development, prevention, and treatment of diseases at the molecular level.

In previous studies, researchers used living organisms to validate disease associations with lncRNAs. This method is not only time-consuming but also difficult to use for determining the association pairs of lncRNAs with diseases. To increase the validation accuracy in wet experiments and to reduce the cost of conducting experiments, researchers have used the experimental data summarized by previous researchers combined with state-of-the-art computer models for the preferential selection of disease–lncRNA association pairs.[5−7]

Previous studies have employed the LDA matrix along with its computed similarity matrix to explore the features of individual nodes. LRLSLDA[8] used the lncRNA similarity matrices as the input features of the nodes and reconstructs the association matrices for prediction using the least-squares method with Laplacian regularization. MFLDA[9] explored the matrix three-factor decomposition technique to derive the disease, lncRNA and implicit feature matrices, and the reconstructed predictive association matrix can be derived by multiplying these matrices. SIMCLDA[10] used principal component analysis (PCA) to extract key features and a neural-induced complementation matrix to explore potential lncRNA–disease associations. Regarding IPCARF,[11] association prediction was performed with the Random Forest

Received: June 8, 2024
Revised: July 4, 2024
Accepted: July 22, 2024
Published: July 30, 2024

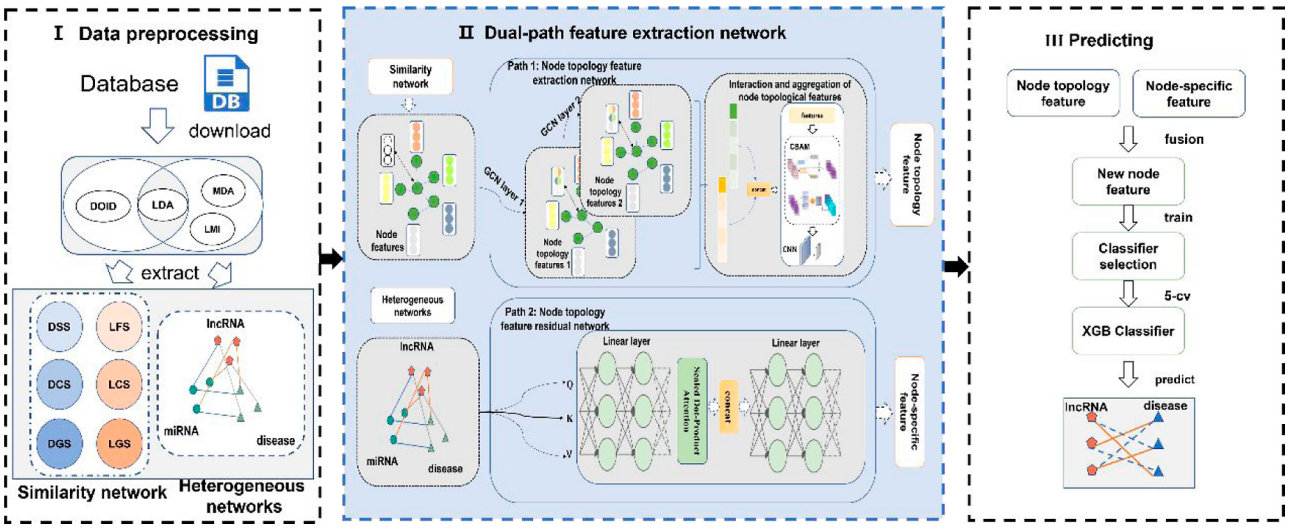


Figure 1. Flowchart of DPFELDA.

technique after feature extraction via PCA. GCHIRFLDA[12] is a method for extracting potential features using an autoencoder and combining it with a Random Forest classifier for prediction. SCCPMD[13] first performed similarity enhancement using logistic functions and then performs prediction of potential association pairs using the probabilistic matrix decomposition method with corrected similarity constraints. These traditional machine learning and matrix factorization methods have achieved acceptable performance, but they neglect the information regarding neighboring nodes that is available in both the LDA association network and the lncRNA−disease similarity network.

To fully leverage the information from neighboring nodes, LDA-LNSUBRW[14] used linear neighborhood similarity in the lncRNA network and disease network for unbalanced birandom walk and finally fused these components to produce the final predicted LDA. LRWRHLDA,[15] Laplace normalization of similarities and interactions between diseases, genes and noncoding RNAs, can be used to make a final prediction after several rounds of iterative training, and a similar method is MHRWR.[16] Birandom walks are also available in MSFUBRW,[17] where the interaction between lncRNAs and diseases was reconstructed using WKNKN[18] based on unbalanced birandom walks and used as a transfer matrix for the respective networks. Similar birandom walk methods are used in NCPBiRW[19] and lung cancer prediction.[20] Although existing network propagation random walk algorithms can utilize the neighbor information on nodes, the experimental process is cumbersome, and the resulting association scores are relatively low.

In recent years, deep learning models, due to their powerful representation learning ability, have greatly promoted the development of bioinformatics and are increasingly being used in various bioinformatics analysis tasks. For example, SpatialGlue[21] applied graph learning and attention mechanisms to the process of spatial transcriptomics data integration, unleashing the full potential of multimodal data. DrlGCL[22] used graph learning and contrastive learning methods to infer potential associations between drugs and diseases, not only improving the accuracy of association prediction, but also saving time and costs. PractiCPP[23] and IGT[24] both utilized Transformer technology to further understand drug delivery systems, which is beneficial for enhancing drug efficacy. This shows that the graph learning model and Transformer model in deep learning have good feature capturing ability. Graph convolutional neural networks (GCNs)[25] in deep learning excel at efficiently aggregating information from neighboring nodes, enriching the features of each node and enhancing the prediction accuracy of association pairs. MAGCNSE,[26] MMGCN,[27] and MAGCN[28] used multiview graph convolutional neural networks with feature-level attention mechanisms for feature extraction and ensemble learning classifiers for node classification. HGATLDA[29] and MCHNLDA[30] fully explored the topological features in the lncRNA and disease similarity network by using a graph attention network (GAT) and enhanced the learning of topological features by using techniques such as multihopping metapath information, a neural-induced complementation matrix and contrast learning. MGLDA,[31] NCPred,[32] and GTAN[33] used metapaths, convolutional neural networks and multiple attention mechanisms to extract features from multimodal lncRNA, disease and miRNA data and then classify the final association pairs using MLP classifiers. While these models can efficiently extract neighbor node features, GCN and GAT may suffer from a lack of node specificity, leading to information loss and hindering the accurate prediction of newly associated nodes.

In summary, first, we found that most of the current LDA association prediction models are using graph learning to explore the topological features of the nodes for diseases and lncRNAs in homogeneous or heterogeneous networks, then the obtained topological features of the disease and lncRNA are processed using feature fusion techniques in deep learning. However, they ignore the fact that a node's associative relationships with other types of nodes in a heterogeneous network may contain information specific to the node itself when constructing a topological network using interaction relationships such as disease-LncRNA and so on. Adding node-specific features on top of node topological features will effectively alleviate the problem of vanishing node specificity caused by graph learning. Therefore, we propose a new model (DPFELDA) which combines a node topology feature extraction network with a node topology feature residual network. The key contributions of DPFELDA include the following aspects:

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

First, we construct a dual-path node feature extraction network to extract the full node features of lncRNAs and diseases, and this network comprises a node topology feature extraction network and a node topology residual network. The former extracts comprehensive information from neighboring nodes with diverse topologies in the dual perspective of lncRNA and disease, while the latter learns distinctive node features from the overall information on lncRNA and disease.

Second, the node topology feature extraction network leverages the two-view structure of diseases and lncRNAs as input. It extracts neighborhood node information from lncRNAs and diseases using GCNs on various network structures. The enhancement and interaction of multilayered node topological information are facilitated through channel attention and spatial attention in the convolutional block attention module (CBAM) framework. Subsequently, a node aggregation layer (NAL) aggregates these topological features.

Finally, we design a node topology residual network within the dual-path feature extraction network to address the issue of vanishing node specificity. It employs the global attention mechanism of the Transformer to extract specific features for each node, facilitating enhanced information complementarity for the node topology feature extraction network.

2. MATERIALS AND METHODS

As shown in Figure 1, we developed a novel disease−lncRNA association prediction model, DPFELDA, comprising three phases: data preprocessing, dual-path feature extraction, and a prediction model. During the data preprocessing stage, the similarity matrix between lncRNAs and diseases was calculated based on the association matrix and the Disease-Ontology ID (DOID),[34] utilizing three association matrices [the LncRNA−Disease Association Matrix (LDM), MiRNA−Disease Association Matrix (MDM), and LncRNA−MiRNA Interaction Matrix (LMM)] to construct the disease-lncRNA−miRNA heterogeneous network. In the dual-path feature extraction stage, the network is bifurcated into a node topological feature extraction network and a node topological feature residual extraction network. The former operates with a dual-view structure employing lncRNA and disease similarity network as input, utilizing techniques such as CBAM and NAL for processing node topological features. Moreover, the node topology feature residual extraction network takes the mutual relationships in the lncRNA-disease-miRNA heterogeneous network as input. It utilizes the Transformer to extract specific features of nodes, which serve as compensatory features for the nodes. In the prediction model, the features extracted from both networks are combined and input into the final classifier XGBoost for LDA prediction.

2.1. Data Sets. We utilized four data sets to assess the model's performance. First, Data set 1, from MFLDA,[9] comprises 240 lncRNAs and 412 diseases involving 495 miRNAs. This data set encompasses 2697 lncRNA−disease associations, 13,562 miRNA−disease associations, and 1002 miRNA−lncRNA interactions. Data set 1 was used to investigate the feasibility and generalizability of the model.

Data set 2 from LDAformer.[35] It incorporated experimentally validated associations of lncRNAs with diseases sourced from the Lnc2Cancer v3.0[36] and LncRNADisease v2.0[37] databases. Additionally, for a more comprehensive analysis of the nodal features and specific features of lncRNAs and diseases, they utilized HMDD v3.2[38] to incorporate 8540 association relationships between 316 diseases and 295 miRNAs, leading to the identification of 2108 lncRNA−miRNA interactions from the starBase v2.0[39] database. Subsequently, Data set 2 was optimized based on the model parameters, and its performance was compared against that of other baseline models. Furthermore, we conducted model exploration using this data.

Data set 3 differs slightly from Data set 1. Several existing prediction models for LDA utilize this data set3 as a benchmark database. To reaffirm the superiority of DPFELDA, we performed additional model comparisons using this data set. Additional data set details are presented in Table 1.

Table 1. Information on the Dataset

	lncRNA	disease	miRNA	LDA	MDA	MLI
Data set1	240	412	495	2697	13 562	1002
Data set2	665	316	295	3833	8540	2108
Data set3	240	405	495	2687	13 359	1002
Data set4	636	204	262	4748	7696	2091

Additionally, we constructed the data set4 by updating the data sources from lncRNAdisease v2.0 to lncRNAdisease v3.0 and HMDD v3.2 to HMDD v4.0 based on the data set2. It is mainly used for model robustness studies.

In Table 1, LDA represents the association between lncRNAs and disease, MDA represents the relationship between miRNAs and disease, and MLI represents the interaction relationship between miRNAs and lncRNAs.

2.2. Methods. *2.2.1. Similarity Matrix Calculation. 2.2.1.1. Disease Semantic Similarity.* Disease−lncRNA association prediction is based on the hypothesis proposed by Chen and Yan[8] that similar diseases tend to be associated with functionally similar lncRNAs. Therefore, we obtained disease semantic information from Disease Ontology[34] and used the disease ontology to represent the parent−child relationships between diseases in the data structure of the directed acyclic graph. We calculated the semantic similarity matrix of diseases according to Wang's method[40] and the semantic contribution score of diseases to diseases according to the directed acyclic graph. Assuming that D is the ancestor node of the DAG and D' is the child node of D, the semantic contribution score S_{D1} of D_1 node in the DAG is

$$\begin{cases} S_{D_1}(D) = 1 & \text{if } D = D_1 \\ S_{D_1}(D) = \max\{0.5 \times S_{D_1}(D') | D' \int \text{children of } D_1\} & \text{if } D \neq D_1 \end{cases} \quad (1)$$

After the contribution scores is obtained, the semantic score S_{v1} is calculated for D_1

$$S_{V1}(D_1) = \sum_{D \in T(D_1)} S_{D_1}(D) \quad (2)$$

$T(D)$ represents the DAG topology of disease D_1, where D is a child node of D_1.

Finally, the semantic similarity of the two diseases is calculated with the following formula

$$\mathrm{DSS}(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (S_{d(i)}(t) + S_{d(j)}(t))}{S_v(d(i)) + S_v(d(j))} \quad (3)$$

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

Table 2. Performance of the Compared Methods on dataset1 and dataset2 by 5cv2[a]

metric	data set	HOPEXGB	SDLDA	GAEMCLDA	IPCARF	LDAformer	GCLMTP	NAGTLDA	DPFELDA
ACC	data set1	0.8988	0.9870	0.4897	0.9824	0.8794	0.7510	0.9431	**0.9956**
	data set2	0.8069	0.9852	0.4895	0.9857	0.9054	0.8403	0.9387	**0.9943**
precision	data set1	0.9531	0.8594	0.0101	0.8579	0.0427	0.0169	0.7086	**0.9796**
	data set2	0.8882	0.7970	0.0060	0.7825	0.0763	0.0197	0.6560	**0.9427**
recall	data set1	0.7987	0.3772	**0.9444**	0.4263	0.8863	0.7498	0.6422	0.8565
	data set2	0.6152	0.3776	**0.8699**	0.3210	0.7596	0.8618	0.6945	0.7349
MCC	data set1	0.8693	0.7314	0.4831	0.5975	0.1774	0.0866	0.6432	**0.9139**
	data set2	0.7354	0.5430	0.4931	0.4957	0.1516	0.1156	0.6289	**0.8297**
*F*1-score	data set1	0.8691	0.5242	0.0201	0.5691	0.0811	0.0329	0.6727	**0.9139**
	data set2	0.7269	0.5122	0.0118	0.4551	0.0661	0.0347	0.6738	**0.8259**
AUC	data set1	0.8988	0.9812	0.9324	0.9261	0.9561	0.8231	0.9381	**0.9967**
	data set2	0.8069	0.9560	0.8676	0.9132	0.9271	0.9175	0.9348	**0.9888**
AUPR	data set1	0.7667	0.7990	0.0372	0.6836	0.3568	0.0658	0.7629	**0.9671**
	data set2	0.5534	0.6040	0.0244	0.5762	0.2273	0.0996	0.7381	**0.9222**

[a]Bolding indicates that the evaluation indicator has reached its optimal value under the baseline model.

where t represents the union of the two diseases in terms of topological structure.

2.2.1.2. LncRNA Functional Similarity. We calculated the functional similarity matrix of lncRNAs based on the known LDA matrix and the computed semantic similarity matrix of diseases by Wang's method[5] with the following formula

$$\mathrm{LFS}(l_1, l_2) = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} \max_{1\le j\le n_2}(\mathrm{DSS}(d_{1i}, d_{2j})) + \sum_{j=1}^{n_2} \max_{1\le i\le n_1}(\mathrm{DSS}(d_{1i}, d_{2j}))\right] \tag{4}$$

For lncRNAs l_1 and l_2, if there are n_1 diseases associated with l_1 and n_2 diseases associated with l_2, DSS denotes the semantic similarity score between the two diseases, and d_{1i} and so on represent the disease types.

2.2.1.3. Gaussian Interaction Profile Kernel Similarity for lncRNAs and Diseases. Due to the sparsity of the LFS and DSS similarity matrices, it may be challenging to explore potential internal associations fully through disease semantic and lncRNA functional similarities. This inherent difficulty could introduce bias into the prediction results. To counteract this limitation, we incorporated the Gaussian interaction profile kernel similarity.[41] The formula is calculated as follows

$$\mathrm{RGS}(r_1, r_2) = \exp(-b_i \left\|\mathrm{IP}(r_i) - \mathrm{IP}(r_j)\right\|^2) \tag{5}$$

$$b_i = \frac{b_i'}{\left(\frac{1}{m}\sum_{i=1}^{m} \|\mathrm{IP}(r_i)\|\right)} \tag{6}$$

where RGS stands for the Gaussian kernel similarity of disease or lncRNA, r_1 and r_2 stand for the two lncRNAs or diseases, $\mathrm{IP}(r_i)$ is a binary vector representing the ith row in the LDA matrix, and b_i is a parameter in the Gaussian kernel computation. After previous experimental investigations and validations, we chose to set the parameter to 1. The definition of notations is detailed in Table 8.

2.2.1.4. Disease and lncRNA Cosine Similarity. To enhance the richness of node features, this study employs cosine similarity based on Euclidean distance to augment the diversity of bioinformatic networks. This approach is beneficial for the feature extraction process using graph convolutional neural networks, enabling the acquisition of more comprehensive node topology information. The calculation formula is as follows

Table 3. Further Comparisons on Dataset 3

method	average AUC	average AUPR
GTAN (2022)	0.983	0.454
NCPred (2023)	0.984	0.640
MGLDA (2023)	0.987	0.512
GAIRD (2023)	0.988	0.649
AGLDA (2024)	0.988	0.684
DPFELDA	0.997	0.968

$$\mathrm{LCS}(i, j) = \frac{A_{ld}(i, :) \times A_{ld}(j, :)}{\|A_{ld}(i, :)\|\|A_{ld}(j, :)\|} \tag{7}$$

$$\mathrm{DCS}(i, j) = \frac{A_{ld}.\ T(i, :) \times {A_{ld}}^T(j, :)}{\|{A_{ld}}^T(i, :)\|\|{A_{ld}}^T(j, :)\|} \tag{8}$$

$\|\cdot\|$represents the modulus length, A_{ld} $(i,:)$ represents the ith column of the association matrix, and T represents the transpose symbol.

2.2.2. Representation of Association Information between lncRNAs and Diseases in Heterogeneous Networks. In this method, there are l lncRNAs, d diseases, and m miRNAs, so the association matrix and interaction matrix can be defined as

$$\begin{cases} A_{ld} \in R^{l\times d} \\ A_{md} \in R^{m\times d} \\ I_{lm} \in R^{l\times m} \end{cases} \tag{9}$$

The heterogeneous association information on lncRNAs with diseases in heterogeneous networks can be expressed as follows

$$l_r = [A_{ld} I_{lm}]$$
$$l_r \in R^{l\times(d+m)} \tag{10}$$

$$d_r = [{A_{ld}}^T {I_{md}}^T]$$
$$d_r \in R^{l\times(l+m)} \tag{11}$$

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

Table 4. Performance of DPFELDA and Its Variants of 5cv1 in Dataset 2

single-path network	AUC (%)	AUPR (%)	dual-path network	AUC (%)	AUPR (%)
GCN	96.91	79.65	GCN + Transformer	97.36	84.09
GCN + CBAM	98.07	81.16	GCN + CBAM + Transformer	98.29	86.47
GCN + NAL	97.66	85.00	GCN + NAL + Transformer	97.85	86.85
GCN + CBAM + NAL	98.63	91.25	GCN + CBAM + NAL + Transformer	98.96	92.43

Table 5. Performance of DPFELDA with Different Views of 5cv1 in Dataset 2

metric	AUC (%)	AUPR (%)
DCS + LCS	97.77	84.70
DSS + LFS	96.78	81.53
DGS + LGS	97.19	84.27
(DSS + LFS) & (DGS + LGS)	97.48	85.08
(DSS + LFS) & (DCS + LCS)	97.64	84.65
(DGS + LGS) & (DCS + LCS)	98.63	91.43
(DGS + LGS) & (DCS + LCS) & (DSS + LFS)	98.96	92.43

Table 6. Hyperparameter Settings for DPFELDA

hyperparameter	setting
transformer self-attention heads	(2, 4, 6, **8**)
transformer encoder layers	(**1**, 2, 3, 4)
GCN layers	(1, 2, **3**, 4, 5)
node feature output dimensions	(32, 64, **128**, 256)

A_{ld} represents the LDM, A_{md} represents the MDM, I_{lm} is the LMM, l_r represents the multivariate data feature of lncRNAs, d_r represents the disease multivariate data feature, and R represents real numbers.

2.2.3. Dual-path Feature Extraction. Based on the association information between lncRNAs and diseases in the dual-view structure and the lncRNA-disease-miRNA heterogeneous network, we propose a dual-path feature extraction model which combines the node topology feature extraction network with the node topology feature residual network, as shown in Figure 2.

2.2.3.1. Node Topology Feature Extraction Network. **Graph Convolutional Neural Network**: We utilized the GCN[25] as a theoretical motivation for extracting the topological features of lncRNAs associated with diseases, so we considered using a multilayer GCN with hierarchical propagation rules to explore the topological features of the network of diseases associated with lncRNAs, and the formula for the GCN is shown below

$$f(H^{(l)}, A) = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)}) \quad (12)$$

$$\hat{A} = A + I \quad (13)$$

where A is the similarity matrix, and the sizes of the similarity matrices for lncRNAs and diseases are $\boldsymbol{l} \times \boldsymbol{l}$ and $\boldsymbol{d} \times \boldsymbol{d}$, respectively. $\hat{A}$ is the self-connection of A, $\hat{D}$ is the degree matrix of $\hat{A}$, $H^{(l)}$ is the node feature, $W^{(l)}$ is the linear transformation matrix and σ is the ReLU activation function. In this method, a three-layer GCN is chosen to explore the topological features present in the lncRNA−disease similarity network, where the input and output features of each node are 128-dimensional.

CBAM: Taking as input the intermediate features, H_l, $H_d \in R^{C\times H\times F}$, of each similar network in which the GCN has explored the given lncRNA and disease, CBAM[42] sequentially infers a 1D channel attention matrix, $M_C \in R^{C\times 1\times 1}$, and a 2D spatial attention matrix, $M_S \in R^{1\times H\times W}$.

The following formula is used for the channel attention module

Table 8. Table of Abbreviations and Symbols

lncRNA	long noncoding RNA
GCN	graph convolutional network
CBAM	convolutional block attention module
NAL	node aggregation layer
l	LncRNA number
d	disease number
m	MiRNA number
$\|\cdot\|$	Modulo operation
$\|$	concatenate
$\otimes$	dot product

Table 7. Top 30 Predicted lncRNAs Associated with Hepatocellular Cancer in dataset1

rank	lncRNA	evidence	PMID	rank	lncRNA	evidence	PMID
1	MEG3	LncRNADisease	21625215	16	MIR194-2HG	LncRNADisease	33116574
2	H19	LncRNADisease	21489289	17	HNF1A-AS1	LncRNADisease	27084450
3	GAS5	LncRNADisease	26163879	18	ZEB1-AS1	LncRNADisease	25025236
4	MALAT1	LncRNADisease	26614531	19	PVT1	LncRNADisease	27495068
5	CDKN2B-AS1	LncRNADisease	25966845	20	SNHG1	LncRNADisease	27133041
6	TUG1	LncRNADisease	27339553	21	AFAP1-AS1	lnc2Cancer	26803513
7	UCA1	LncRNADisease	26551349	22	DANCR	LncRNADisease	27919960
8	LINC00687	UNKNOWN		23	HOTTIP	LncRNADisease	24114970
9	PANDAR	lnc2Cancer	26054684	24	HULC	LncRNADisease	27285757
10	DBH-AS1	LncRNADisease	26393879	25	CCAT2	LncRNADisease	28280353
11	LINC00602	UNKNOWN		26	NEAT1	LncRNADisease	28526689
12	LINC00974	lnc2Cancer	25476897	27	MIR7−3HG	LncRNADisease	24296588
13	HCCAT5	Literature	23314567	28	XIST	LncRNADisease	27776968
14	CYTOR	LncRNADisease	26356260	29	SPRY4-IT1	LncRNADisease	27899259
15	HOTAIR	lnc2Cancer	32062551	30	TRERNA1	lnc2Cancer	31012192

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

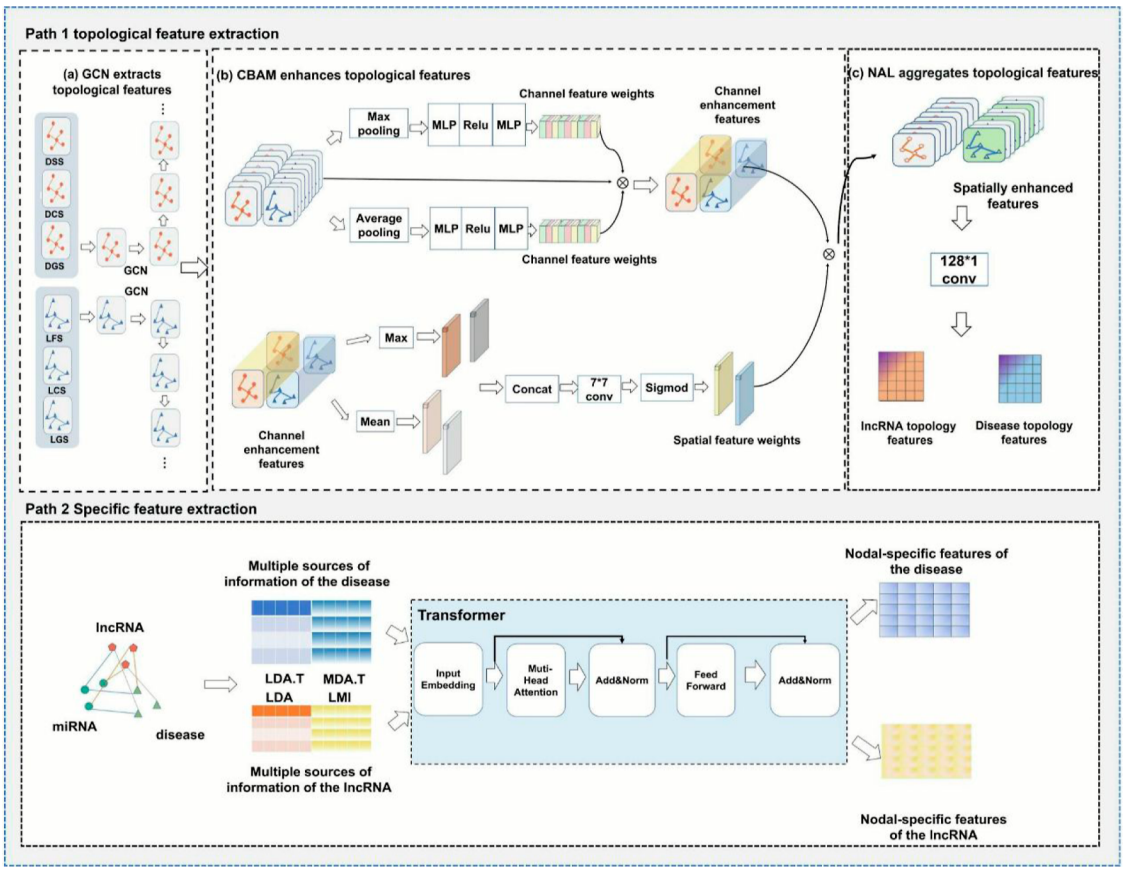


Figure 2. Presents the framework of the model, which consists of two parts. The first part is the node topology feature extraction network, which is responsible for extracting the node topology features. The node features from various convolutional layers of the GCN are enhanced using CBAM for channelwise and spatialwise feature enhancement. Then, the aggregated features are passed through the NAL for multichannel convolutional dimension reduction. The second part is the node topology feature residual network, which utilizes the heterogeneous association information between lncRNAs and diseases, along with the Transformer, to extract node-specific features. This network focuses on capturing the specific features of the nodes related to lncRNAs and diseases.

$$M_C(H) = \sigma_1(\mathrm{MLP}(\sigma_2(\mathrm{AvgPool}(H))) + \mathrm{MLP}(\sigma_2(\mathrm{maxPool}(H))))) \quad (14)$$

where H is the feature matrix of multiple channels, σ_1 is the ReLU activation function, σ_2 is the sigmoid activation function for the purpose of exploring more potential nonlinear features, MLP is the multilayer perceptron for extracting linear features between channels, AvgPool is the average pooling layer for obtaining the information on the global perceptron field of the channels, and MaxPool is used to extract the key features between channels. H is the intermediate features that have GCN explored and spliced together, F is the node feature dimension, H_l and H_d are the input features of lncRNA or disease for CBAM, the input feature of disease is $1 \times 9 \times \boldsymbol{d} \times 128$, the output is $1 \times 9 \times 1 \times 1$, the input feature of lncRNA is $1 \times 9 \times \boldsymbol{l} \times 128$, and the output is $1 \times 9 \times 1 \times 1$.

For the spatial attention module, the formula is shown below

$$M_s(H_1) = \sigma(f^{7\times7}[\mathrm{AvgPool}(H_1), \mathrm{maxPool}(H_1)]) \quad (15)$$

where $f^{7\times7}$ represents the convolution kernel of size 7×7, which is derived from the experimental validation of the model,[42] where H_1 is the input feature of the lncRNA or disease for CBAM, the input feature of the disease is $1 \times 9 \times \boldsymbol{d} \times 128$, the output is $1 \times 1 \times \boldsymbol{d} \times 128$, the input feature of the lncRNA is $1 \times 9 \times \boldsymbol{l} \times 128$, and the output is $1 \times 1 \times \boldsymbol{l} \times 128$.

Finally, we can obtain the feature matrix enhanced by CBAM features with the following formula

$$H_1 = M_c(H) \otimes H \quad (16)$$

$$H_2 = M_s(H_1) \otimes H_1 \quad (17)$$

where $\otimes$ represents the element-by-element multiplication, during which the attention values are broadcasted accordingly, and the channel attention is broadcasted along the channel dimension; otherwise, H_2 is the final feature-enhanced feature.

In addition, CBAM attention can be adaptively adjusted according to the characteristics of different views, thus improving the generalizability and adaptability of the model.

NAL: After the graph convolutional layer and convolutional attention, we obtain the feature matrix with node neighborhood topology information and interchannel interactions, which tend to introduce noise due to the high dimensionality of the obtained features. Therefore, in this method, a rectangular convolution kernel is chosen to aggregate only the individual node features, thus reducing the dimensionality of the data and maintaining the integrity and interpretability of the data to be explored. The formula is as follows

$$H_l = g(\mathrm{conv}, H_2^l) = \mathrm{conv}(h_2^{l1}, h_2^{l2}, \ldots, h_2^{l9}) \quad (18)$$

$$H_d = g(\mathrm{conv}, H_2^d) = \mathrm{conv}(h_2^{d1}, h_2^{d2}, \ldots, h_2^{d9}) \quad (19)$$

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

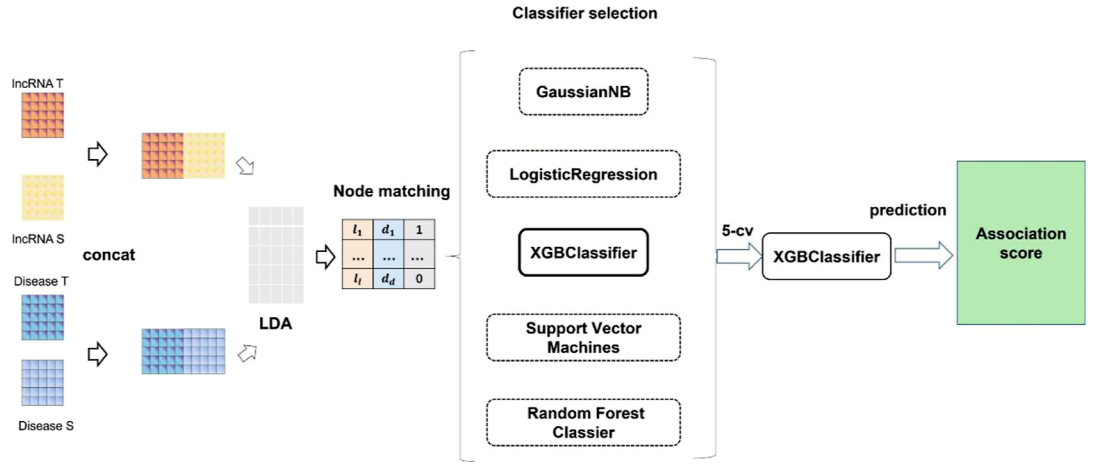


Figure 3. Feature splicing, node matching and final classification, where lncRNA T and disease T represent topology node features explored by the node topology feature extraction network, and lncRNA S and disease S represent node-specific features explored by the node topology feature residual network.

where H_l and H_d are the topological features of lncRNAs and diseases derived from the final training, the convolution kernel is 128 × 1 with a step size of 1, the number of input channels is 9, and the number of output channels is 128. h_2^{l1} represents the node features explored by different GCN layers and enhanced by CBAM for different lncRNA views.

2.2.3.2. Node Topology Feature Residual Network. In neighborhood feature aggregation, the effectiveness of the GCN network heavily relies on the degrees of both the central node and its neighboring nodes. Following the graph convolution formula 12, if nodes *A* and *B* are connected and the degree of node *B* is lower than that of node *A*, the features of node *A*'s aggregation can be diluted by those of node *B*. Consequently, node *A* may lose its unique characteristics, potentially leading to oversmoothing of node features. To address this issue, we employ the Transformer architecture, which utilizes the self-attention mechanism to learn global features while preserving the specificity of the node features.

In this method, we construct a node topology feature residual extraction network using the encoder layer of the Transformer[43] model. The encoder layer of the Transformer model comprises N identical stacked layers, with each layer consisting of two sublayers. The first sublayer employs a multihead self-attention mechanism, while the second sublayer is a feedforward neural network based on a fully connected layer. To harness the advantages of residual neural networks, we incorporate residual connections in each sublayer. Additionally, a layer of normalization is introduced. To ensure consistency in the residual connections, all sublayers in the model, including their embeddings, yield 128-dimensional outputs.

Regarding the Multi-Head Attention layer, in this method, the resulting compensated feature matrix is coded for dimensionality reduction using fully connected layers to explore potential linear features. The formula is as follows

$$\hat{b}_l = l_r W_{\ln c},\ W_{\ln c} \in R^{N_l \times N_f} \tag{20}$$

$$\hat{b}_d = d_r W_{\text{dis}},\ W_{\text{dis}} \in R^{N_d \times N_f} \tag{21}$$

where $\hat{b}_l$ and $\hat{b}_d$ are linear transformations of l_r and d_r. l_r and d_r are multivariate data features of lncRNA and disease, respectively. Matrices such as W are linearly transformed with respect to the features.

As the methodology for investigating compensatory features of lncRNAs and diseases remains consistent, the subsequent section will delineate the approach to exploring the node-specific attributes of an lncRNA node.

After obtaining the features after the completion of the transformation, the similarity between the nodes is determined, and the similarity is used to calculate the weight parameters of each feature with the following formula

$$s_{\text{in}} = \frac{\hat{b}_i \otimes \hat{b}_n}{\sqrt{d_k}} \tag{22}$$

where ⊗ represents the function matmul in Python. $\hat{b}_i$, $\hat{b}_n$ represent the node features of the *i*th and *n*th lncRNAs, and d_k is the scale factor, which is 128.

After determining the similarity between each node and each node, using the softmax function, the feature weights α are derived with the following formula

$$\alpha_{\text{in}} = \frac{\exp(s_{\text{in}})}{\sum_{m=1}^{l} \exp(s_{\text{im}})} \tag{23}$$

After calculating the weights, i.e., it is possible to obtain the node features containing new and richer information after the attention mechanism. The formula is as follows

$$\tilde{b}_l = \sum_{n=1}^{l} \alpha_{\text{in}} b_n + b_i \tag{24}$$

here, $\tilde{b}_i$ represents the node feature of the *i*th lncRNA derived through exploration, encompassing its unique node-specific details and integrating information from other nodes within the lncRNA perspective.

The idea of multihead attention is the same as the idea of self-attention; only the feature matrix is segmented. First, the self-attention mechanism is performed on each part of the segmentation, and then splicing is performed after the process of the attention mechanism. The feature splicing of lncRNAs is taken as an example, and the splicing formula is as follows

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

$$B_i = \Big\|_{h=1}^{H} \tilde{b}_h \tag{25}$$

‖ denotes splicing. B_i represents the features trained in the initial sublayer of the Transformer architecture, where H is designated as the eight-headed attention mechanism in this study.

In addition to the attention sublayer, each layer of the coding layer contains a fully connected feed-forward neural network that consists of two linear transformations with an activation function, as shown in the following equation

$$\mathrm{FFN}(B_{\mathrm{lnc}}) = \mathrm{ReLU}(B_{\mathrm{lnc}}W_1 + b_1)W_2 + b_2 \tag{26}$$

where W_1 and W_2 are the weight matrices of the fully connected layer and b_1 and b_2 are the biases.

2.2.3.3. Loss Function. To mitigate the reliance on known association nodes in the dual-path feature extraction network, we optimize the loss function by employing matrix multiplication. This approach reduces the two sets of lncRNAs acquired with the updated disease features into an association matrix. By utilizing the association matrix, we aim to obtain more precise potential features. The optimized formula for the loss function is presented below

$$\mathrm{LD}' = F_l \times F_d^T \tag{27}$$

$$\mathrm{loss} = \|\mathrm{LD}' - \mathrm{LD}\|_F^2 \tag{28}$$

2.2.4. Prediction Model. As shown in Figure 3, the prediction of lncRNA−disease association classification is accomplished based on the node features of lncRNAs and diseases obtained from the dual-path feature extraction network. These features are paired through association and labeled. We conducted comparisons with various classifiers, including the Naïve Bayes classifier,[44] logistic regression,[45] XGBoost,[46] support vector machine,[47] and random forest classifiers.[48] We ultimately selected the XGBoost classifier to carry out the classification task.

2.2.5. Experimental Environment. Our method was implemented using the torch_geometric package based on the PyTorch framework. The experiments were conducted on a Windows 10 operating system with an Intel(R) Core(TM) i7−8550U processor and 16 GB of RAM. The maximum number of epochs in our model was set to 500, and all the trainable parameters were optimized using the Adam optimizer with a learning rate of 0.001 and a weight decay rate of 0.05.

Since the data set is imbalanced, with a much larger number of negative samples than positive samples, it is crucial to evaluate the model's ability to retrieve true positive samples from the predicted positive samples. Here, positive samples are node pairs where 1 is located in the LDM, and negative samples are node pairs where 0 is located in the LDM. In our experiments, we evaluated the model using the following two approaches:

5-fold cross-validation (CV1): The positive and negative samples are divided into five equal parts, with one part randomly selected as the test set and the remaining four parts used as the training set. This evaluation method is mainly employed on Data set 2 in ablation experiments, parametric analysis, and other model building experiments.

5-fold cross-validation (CV2): Repeatedly perform 5-fold cross-validation experiments and select the group of experimental results with the smallest variance as the final result. This evaluation method is utilized for comparisons with baseline models, and model exploration in comparison with the substitution of the dual-path feature extraction network.

2.2.6. Evaluation Metrics. To better evaluate the model performance, based on the evaluation methods of existing studies,[35] we used the area under the receiver operating characteristic curve (AUC) and the area under the precision−recall curve (AUPR) as the comprehensive performance evaluation metrics of DPFELDA. Additionally, we used six other evaluation metrics, including accuracy, sensitivity, specificity, precision, *F*1-score, and Matthews correlation coefficient (MCC). These indicators were calculated as follows

$$\mathrm{accuracy} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{29}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{30}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{31}$$

$$F1 - \mathrm{score} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{32}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{33}$$

3. RESULTS

3.1. Performance Comparisons. To verify the validity of the model, we choose the state-of-the-art model in the literature that can work properly through debugging as the baseline model, and to ensure fairness, the model comparison experiments use the same data set with the same experimental environment. The baseline models include LDAformer,[35] which mainly utilizes multihop metapaths as original features and uses a Transformer for feature extraction. GAEMCLDA[49] uses matrix factorization and a graph autoencoder for association prediction of LDA. SDLDA,[50] exploring lncRNA-disease associations using matrix decomposition with fully connected layers. IPCARF[11] is an association prediction method based on a machine learning prediction of association methods using PCA and random forest for prediction. HOPEXGB,[51] exploring heterogeneous graphs using high-order proximity preserved embedding and using XGB to predict the association of lncRNAs with disease. GCLMTP[52] explores lncRNA and disease node features using graph Contrastive Learning and uses machine learning classifiers for association classification. NAGTLDA[53] utilized GCN to explore potential global and local features, and utilized Transformer's attention mechanism for feature fusion.

The baseline models in this study were constructed using traditional machine learning algorithms and deep learning algorithms. The parameters suggested in the literature were used to run each of these methods. Specifically, GAEMCLDA, LDAformer, NAGTLDA and GCLMTP incorporate the concept of learning graph structures. The evaluation metrics based on 5-CV2 averaging are shown Table 2. To establish the statistical distinction in the predictive performance of DPFELDA relative to the compared methods, we performed Wilcoxon test analyses on the AUC and AUPR evaluation

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

criteria of the baseline methods, as detailed in Supporting Information Table S4. The results revealed that all p-values were below 0.05, signifying a significant superiority of DPFELDA's performance over the other methods.

For two different data sets, Data set1 and Data set2, the same method shows certain differences between the databases. Considering that this may be due to the different ratio of positive to negative samples in the two data sets, with data set1 having a ratio of (1:37) and data set2 having a ratio of (1:54), it can be inferred that a sparse network structure is not conducive to association prediction by LDA. In the comparison of both data sets, DPFELDA demonstrated a significant superiority over the baseline model in both evaluations. In data set 1, DPFELDA outperforms the second-ranked SDLDA by 1.55% in AUC and 16.81% in AUPR. Regarding the evaluation metrics MCC and *F*1-score for the unbalanced data set, DPFELDA surpasses the second-ranked HOPEXGB by 4.46 and 4.48%, respectively. In data set 2, DPFELDA surpasses the second-ranked SDLDA by 3.28% in AUC and 31.82% in AUPR. Similarly, concerning the evaluation metrics MCC and *F*-score for the unbalanced data set, DPFELDA exceeds the second-ranked HOPEXGB by 9.43 and 9.9% respectively. The comparison of the baseline models demonstrates the superior performance of DPFELDA. In addition, the comparison with the dual data set in the baseline model highlights the improved suitability of DPFELDA for predicting correlations in unbalanced data sets compared to other LDA prediction models.

Data set 3 serves as a widely adopted benchmark data set, enabling the comparison of DPFELDA with a broader range of graph-structured learning LDA association prediction models. We also explore advanced methods, specifically GTAN,[33] NCPred,[32] MGLDA,[31] GAIRD,[54] and AGLDA,[55] which integrate additional deep learning techniques like graph learning and dual paths, showcasing strong performance. In Table 3, DPFELDA achieves the highest AUC, surpassing the second-best AGLDA by 0.8%. Additionally, DPFELDA demonstrates strong performance in AUPR, with most other prediction models having AUPR values between 0.4 and 0.6, while DPFELDA achieves a better 0.967, a notable 28.3% higher than the second-best AGLDA. These results highlight the superior predictive capabilities of DPFELDA.

3.2. Ablation Experiment. *3.2.1. Feature Extraction Network.* To validate the effectiveness of our innovative components, including the dual-path feature extraction network, CBAM, and NAL, we conducted ablation experiments in data set2. As shown in Table 4, the DPFELDA model, which combines the dual-path feature extraction network, CBAM, and NAL, achieved the best performance. Compared to the single-path feature extraction network, the dual-path network with residual connections for node topology features demonstrated an average improvement of 0.3725% in AUC and 3.195% in AUPR. Within the dual-path feature extraction network, removing the CBAM layer resulted in decreases of 1.11% in the AUC and 5.58% in the AUPR, while removing the NAL layer led to reductions of 0.67% in the AUC and 5.96% in the AUPR. Moreover, similar trends were observed in the single-path feature extraction network (after removing CBAM, the AUC and AUPR decreased by 1.03 and 6.25%, respectively; after removing NAL, the AUC and AUPR decreased by 0.56 and 10.97%, respectively).

3.2.2. Multiview Fusion. We explored the impact of similarity networks and their combinations on disease-lncRNA prediction in data set2 via a dual-pathway feature extraction network and the XGBoost classifier. As shown in Table 5, the effects of different similarity networks and their combinations on LDA prediction are not simply linear relationships. Compared with the DSS + LFS combination, the DCS + LCS combination achieved a greater AUC and AUPR (0.13 and 0.05%, respectively). Among the two-view combinations, DCS + LCS outperformed DSS + LFS, with higher AUC and AUPR values (0.99 and 3.17%, respectively), and DCS + LCS also outperformed DGS + LGS, with higher AUC and AUPR values (0.58 and 0.43%, respectively). Among the combinations of the two similarity networks, DGS + LGS and DCS + LCS achieved greater AUC and AUPR values than did DSS + LFS and DGS + LGS (1.15 and 6.35% and 0.99 and 6.78%, respectively). The combination of DGS + LGS, DCS + LCS and DSS + LFS yielded the best prediction performance.

3.3. Parameter Analysis. In this study, all the hyperparameters used in DPFELDA are referenced in the Materials and Methods Section. As shown in Table 6, we adjusted the important hyperparameters: the number of heads of the multihead attention and encoder layers in the Transformer and the number of graph convolution layers. In Supporting Information Figure S2, we chose 8 heads among (2, 4, 6, 8) for attention and 1 layer among (1, 2, 3, 4) for the encoder under the path of node topological features to extract the residual network. As shown in Supporting Information Figure S1, we selected 3 layers for graph convolution among (1, 2, 3, 4, 5) under the path of the node topological feature extraction network. In Supporting Information Figure S3, we selected two output dimensions for the topology extraction networks among (32, 64, 128, 256), and the final feature output dimensions for both networks were set to 128 under the dual-path feature extraction network.

3.4. Model Exploration. In addition, to validate the role of the dual path extraction network in DPFELDA, we explore the option of replacing the DPFELDA dual path feature extraction network with the graph attention (GAT) topological feature extraction network in data set2. The model details are shown in the Supporting Information Figure S5. Also, in order to validate the role of the topological feature residual network, we considered the option of using the GAT topological feature extraction network to replace the topological feature extraction network in the DPFELDA scheme. The model details are shown in the Supporting Information (Figure S6).

To validate whether the DPFELDA dual-path feature extraction network alleviates the issue of node-specific disappearance, we considered replacing the dual-path feature extraction network with the GAT topology feature extraction network, as shown in Supporting Information Table S3. The dual-path feature extraction network outperformed GAT with a 5.69% higher AUC and a 42.71% higher AUPR, and DPFELDA demonstrated better performance than GAT in other evaluation metrics. To verify the effectiveness of the topology residual network, we replaced DPFELDA topology feature extraction network with the GAT topology feature extraction network. The addition of the topology residual network on top of the GAT topology feature network resulted in a 3.21% increase in the AUC and a 29.51% increase in the AUPR. Other evaluation metrics also showed varying degrees of improvement, indicating the effectiveness of the topology residual feature extraction network. Inspired by the IPCRF model,[11] we considered four different dimensionality reduction techniques, as shown in Supporting Information Table S2:

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

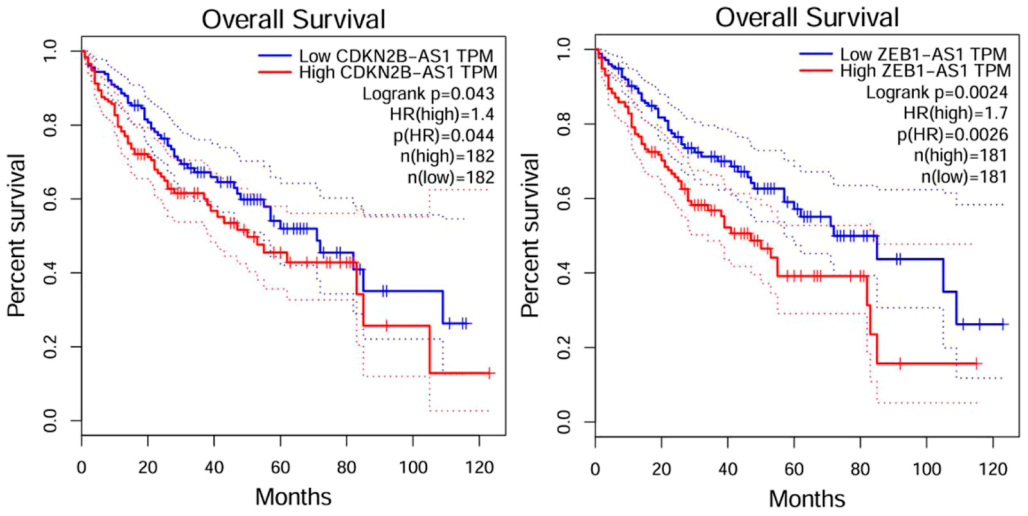


Figure 4. Survival analysis of hepatocellular cancer cells.

PCA[56] for principal component analysis, ICA[57] for independent component analysis, RP[58] for random projection, and NAL as a dimensionality reduction technique based on CNNs and fully connected layers. The results indicated that NAL was more helpful for node classification tasks.

3.5. Case Study. To investigate the model's feasibility, this paper conducts case study experiment. The experiment involves removing the feature columns related to a specific disease and excluding the disease from participating in the training of the classifier model during the training phase. Finally, the particular disease is used to make predictions with lncRNA associations to validate the accuracy of the association prediction results. Hepatocellular carcinoma, breast cancer, gastric cancer, and colorectal cancer were chosen for analysis and prediction in data set 1 and data set 2, respectively. For each disease within different data sets, the top 30 candidate lncRNAs were selected based on their prediction scores. To further validate the predictive ability of DPFELDA, we selected LncRNADisease 3.0,[59] lnc2Cancer v3.0, EVlncRNA s3,[60] and literature-based candidate association pairs for validation. Table 7 displays the candidate lncRNAs identified for hepatocellular carcinoma in data set 1, where 28 lncRNAs are associated with hepatocellular carcinoma. After validation, 26 candidate lncRNAs were confirmed to be associated with breast cancer (see Supporting Information Table S5), 28 with gastric cancer (see Supporting Information Table S6) and 28 with colorectal cancer (see Supporting Information Table S7). In the case study of data set2 with (Supporting Information Tables S8−S11), 30 candidate lncRNAs for breast, gastric, and hepatocellular carcinomas were experimentally validated, while 28 candidate lncRNAs for colorectal cancer were validated.

In data set1, in order to verify the accuracy of the experiment, we made use of TCGA data to do survival analysis of hepatocellular carcinoma with CDKN2B-AS1, ZEB1-AS1. As shown in Figure 4, the lower the expression of LncRNAs leading to hepatocellular carcinoma over time, the higher the survival probability of the patients. This also indicates that our prediction has a positive prognostic effect on the disease. In addition, as shown in Supporting Information Figure S9, we also did a survival analysis of breast cancer with STXBP5-AS1 in data set2. It was found that the lower the expression of LncRNAs leading to breast cancer over time, the higher the probability of patient survival.

4. DISCUSSION

To alleviate the issues of node specificity loss in lncRNA disease association prediction, this study proposes an innovative DPFELDA prediction model. The model utilizes a dual-pathway model feature extraction network to enhance the representation of node features and combines it with an XGBoost classifier for disease association prediction. Through 5-fold cross-validation, the performance of the model was found to be superior to that of other existing LDA prediction models.

Several researchers have adopted the dual-pathway approach (MGLDA,[31] NCPred,[32] and GTAN[33]) for LDA prediction models. Similarly, our DPFELDA model also adopts the dual-pathway feature extraction approach. However, our model uses heterogeneous networks as input and enhances feature acquisition through a similar residual connection, improving the model's predictive performance.

To construct an accurate LDA prediction model, as shown in Supporting Information Table S1 and Figure S4, we compared several classifiers, including naive Bayes, random forest, support vector machines, logistic regression, and XGBoost. Among these classifiers, XGBoost exhibited the best performance, so we chose XGBoost as the classifier for our model.

The LDAformer model utilizes a heterogeneous graph of a similarity network and an association matrix to extract multihop metapath information for nodes via matrix multiplication. As shown in Table 2. The AUC of LDAformer in data set2 reaches 0.9271, but due to the influence of sample imbalance, the AUPR is only 0.2273. In DPFELDA, the node topology feature residual pathway also uses the association information on the heterogeneous graph as input. Under sample imbalance, the AUC and AUPR of the DPFELDA model constructed using the dual-pathway feature extraction approach are 0.0617 and 0.6949 greater than those of the LDAformer, respectively. This indicates that DPFELDA not only improves the predictive performance of LDA but also alleviates the low AUPR caused by sample imbalance.

GAEMCLDA is an LDA prediction model based on graph autoencoders (in data set2 AUC: 0.8676, AUPR: 0.0244). Consistent with the node topology feature extraction network of DPFELDA, both models use graph structure learning methods. As shown in Table 4, under the single pathway of the

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

node topology feature extraction network, utilizing a GCN model combined with various similarity networks to extract node topology features (in data set2 AUC: 96.91, AUPR: 79.65) performed better than the GAEMCLDA model. After adding the CBAM model, its performance improved. However, in the process of graph structure learning, the average feature smoothing of nodes may lead to a decrease in the model's predictive performance. As shown in Table 4, adding the node topology residual network in the dual-pathway network improves the performance of the model, indicating that the DPFELDA node topology residual network alleviates this problem to a certain extent and provides a new way to address the feature smoothing problem. It should be noted that GAEMCLDA compresses the disease feature matrix and the lncRNA feature matrix into a new LDA correlation matrix. It then evaluates the model using this new LDA matrix in comparison with the original LDA matrix. This behavior results in GAEMCLDA classifying a significant number of unknown samples as positive, consequently contributing to its high Recall value.

IPCARF utilizes the PCA technique for dimensionality reduction, which enhances the overall performance of the model. Inspired by the existing research on IPCARF, we NAL to the node topology feature extraction network. However, due to the limitations of the PCA technique, it cannot fully explore and aggregate the extracted node features. Therefore, we innovatively combined the CNN and fully connected layers for feature aggregation, as shown in Table 4 and Figure 2. Under the single-pathway network, the combination of GCN and NAL achieved an AUC and AUPR that were 7.32 and 33.63% greater than those of IPCARF (91.32 and 57.62%, respectively) in data set2. Furthermore, as shown in Table 4, under the dual-pathway network structure, the use of NAL or CBAM as innovative elements does not overshadow their role in improving model performance despite the increase in model complexity.

In order to verify the robustness of the model, we evaluated the generalization of the model using data set1, data set2, data set3, data set4, the results show that the model has a low dependency on the data and achieves a good performance on the four different data sets. See (Supporting Information Table S12).

However, the DPFELDA model still has some limitations. (1) To enhance the complexity of lncRNA-disease topology in constructing the lncRNA-disease view, we computed the similarity matrix between lncRNA and disease based on the existing association matrix. However, this approach results in the training features being reliant on the association matrix, potentially hindering the accurate prediction of new association pairs. (2) As the DPFELDA model is founded on a deep learning model, its learning process requires manual adjustments by individuals, and the hyperparameters within the model structure must also be fine-tuned for varying data volumes and features, thereby intensifying the training workload. In the future, we will continue to explore deep learning models suitable for generalization of lncRNA-disease association prediction and improve the adaptability of the models themselves to data input. Additionally, as bioinformatics progresses, we aim to incorporate multimodal biological data on disease and lncRNA as primary feature inputs, reducing reliance on association matrices. This approach will facilitate the exploration of novel lncRNA-disease associations and potentially aid wet experiments, further enhancing our understanding of the role of lncRNA in disease pathogenesis.

5. CONCLUSION

This paper presents a dual-pathway-based prediction model for lncRNA−disease associations, and through comparative experiments, it demonstrates the superiority of DPFELDA. We also conducted case studies and performed survival analysis experiments on the associations between lncRNAs and diseases, validating the accuracy of lncRNA−disease association prediction. These findings will facilitate the selection of diseases and lncRNAs for use in biological experiments. Further exploration of the role of lncRNAs in disease development will contribute to a better understanding of disease pathogenesis and provide improved treatment strategies.

■ ASSOCIATED CONTENT

Data Availability Statement

All code and data are available at https://github.com/ydkvictory/DPFELDA

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c05365.

Parametric analyses, classifier comparisons, model exploration, Wilcoxon test analysis, case studies with survivability analysis, Robustness experiments (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Dengju Yao − *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;* orcid.org/0000-0002-4974-3054; Email: ydkvictory@hrbust.edu.cn

Authors

Binbin Zhang − *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*

Xiaojuan Zhan − *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China; College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050, China*

Bo Zhang − *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*

Xiang Kui Li − *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;* orcid.org/0009-0009-7325-0258

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c05365

Author Contributions

D.Y. directed the research and revised the paper. B.Z. conceived and implemented the model, performed the experiments, and wrote the paper. X.Z. and B.Z. analyzed the experimental results and revised the paper. X.K.L. performed the experiments and revised the paper. All the authors have read and approved the final manuscript.

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

Funding

This work is supported by the National Natural Science Foundation of China (grant no. 62172128). The funding body did not play any role in the design of the study; the collection, analysis, or interpretation of the data; or the writing of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for suggestions that helped improve the paper substantially.

■ REFERENCES

(1) Kapranov, P.; Cheng, J.; Dike, S.; Nix, D. A.; Duttagupta, R.; Willingham, A. T.; Stadler, P. F.; Hertel, J.; Hackermüller, J.; Hofacker, I. L.; Bell, I.; Cheung, E.; Drenkow, J.; Dumais, E.; Patel, S.; Helt, G.; Ganesh, M.; Ghosh, S.; Piccolboni, A.; Sementchenko, V.; Tammana, H.; Gingeras, T. R. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* **2007**, *316* (5830), 1484−1488.

(2) Yao, R.-W.; Wang, Y.; Chen, L.-L. Cellular Functions of Long Noncoding RNAs. *Nat. Cell Biol.* **2019**, *21* (5), 542−551.

(3) Kim, T.-K.; Shiekhattar, R. Diverse Regulatory Interactions of Long Noncoding RNAs. *Curr. Opin. Genet. Dev.* **2016**, *36*, 73−82.

(4) Fang, Y.; Fullwood, M. J. Roles, Functions, and Mechanisms of Long Non-Coding RNAs in Cancer. *Genomics, Proteomics Bioinf.* **2016**, *14* (1), 42−54.

(5) Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the Human microRNA Functional Similarity and Functional Network Based on microRNA-Associated Diseases. *Bioinformatics* **2010**, *26* (13), 1644−1650.

(6) Liu, M.-X.; Chen, X.; Chen, G.; Cui, Q.-H.; Yan, G.-Y. A Computational Framework to Infer Human Disease-Associated Long Noncoding RNAs. *PLoS One* **2014**, *9* (1), No. e84408.

(7) Signal, B.; Gloss, B. S.; Dinger, M. E. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. *Trends Genet.* **2016**, *32* (10), 620−637.

(8) Chen, X.; Yan, G.-Y. Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* **2013**, *29* (20), 2617−2624.

(9) Fu, G.; Wang, J.; Domeniconi, C.; Yu, G. Matrix Factorization-Based Data Fusion for the Prediction of lncRNA-Disease Associations. *Bioinformatics* **2018**, *34* (9), 1529−1537.

(10) Lu, C.; Yang, M.; Luo, F.; Wu, F.-X.; Li, M.; Pan, Y.; Li, Y.; Wang, J. Prediction of lncRNA-Disease Associations Based on Inductive Matrix Completion. *Bioinformatics* **2018**, *34* (19), 3357−3364.

(11) Zhu, R.; Wang, Y.; Liu, J.-X.; Dai, L.-Y. IPCARF: Improving lncRNA-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and a Random Forest Classifier. *BMC Bioinf.* **2021**, *22* (1), 175.

(12) Yao, D.; Zhang, T.; Zhan, X.; Zhang, S.; Zhan, X.; Zhang, C. Geometric Complement Heterogeneous Information and Random Forest for Predicting lncRNA-Disease Associations. *Front. Genet.* **2022**, *13*, 995532.

(13) Lin, L.; Chen, R.; Zhu, Y.; Xie, W.; Jing, H.; Chen, L.; Zou, M. SCCPMD: Probability Matrix Decomposition Method Subject to Corrected Similarity Constraints for Inferring Long Non-Coding RNA-Disease Associations. *Front. Microbiol.* **2023**, *13*, 1093615.

(14) Xie, G.; Jiang, J.; Sun, Y. LDA-LNSUBRW: lncRNA-Disease Association Prediction Based on Linear Neighborhood Similarity and Unbalanced Bi-Random Walk. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2020**, *19*, 1−997.

(15) Wang, L.; Shang, M.; Dai, Q.; He, P. Prediction of lncRNA-Disease Association Based on a Laplace Normalized Random Walk with Restart Algorithm on Heterogeneous Networks. *BMC Bioinf.* **2022**, *23* (1), 5.

(16) Zhao, X.; Yang, Y.; Yin, M. MHRWR: Prediction of lncRNA-Disease Associations Based on Multiple Heterogeneous Networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2021**, *18* (6), 2577−2585.

(17) Dai, L.; Zhu, R.; Liu, J.; Li, F.; Wang, J.; Shang, J. MSF-UBRW: An Improved Unbalanced Bi-Random Walk Method to Infer Human lncRNA-Disease Associations. *Genes* **2022**, *13* (11), 2032.

(18) Toprak, A.; Eryilmaz, E. Prediction of miRNA-Disease Associations Based on Weighted [Formula: See Text]-Nearest Known Neighbors and Network Consistency Projection. *J. Bioinf. Comput. Biol.* **2021**, *19* (01), 2050041.

(19) Liu, Y.; Yang, H.; Zheng, C.; Wang, K.; Yan, J.; Cao, H.; Zhang, Y. NCP-BiRW: A Hybrid Approach for Predicting Long Noncoding RNA-Disease Associations by Network Consistency Projection and Bi-Random Walk. *Front. Genet.* **2022**, *13*, 862272.

(20) Guo, Z.; Hui, Y.; Kong, F.; Lin, X. Finding Lung-Cancer-Related lncRNAs Based on Laplacian Regularized Least Squares With Unbalanced Bi-Random Walk. *Front. Genet.* **2022**, *13*, 933009.

(21) Long, Y.; Ang, K. S.; Sethi, R.; Liao, S.; Heng, Y.; Van Olst, L.; Ye, S.; Zhong, C.; Xu, H.; Zhang, D.; Kwok, I.; Husna, N.; Jian, M.; Ng, L. G.; Chen, A.; Gascoigne, N. R. J.; Gate, D.; Fan, R.; Xu, X.; Chen, J. Deciphering Spatial Domains from Spatial Multi-Omics with SpatialGlue. *Nat. Methods* **2024**.

(22) Luo, Y.; Shan, W.; Peng, L.; Luo, L.; Ding, P.; Liang, W. A Computational Framework for Predicting Novel Drug Indications Using Graph Convolutional Network with Contrastive Learning. *IEEE J. Biomed. Health Inf.* **2024**, 1−10.

(23) Shi, K.; Xiong, Y.; Wang, Y.; Deng, Y.; Wang, W.; Jing, B.; Gao, X. PractiCPP: A Deep Learning Approach Tailored for Extremely Imbalanced Datasets in Cell-Penetrating Peptide Prediction. *Bioinformatics* **2024**, *40* (2), btae058.

(24) Liu, S.; Wang, Y.; Deng, Y.; He, L.; Shao, B.; Yin, J.; Zheng, N.; Liu, T.-Y.; Wang, T. Improved Drug-Target Interaction Prediction with Intermolecular Graph Transformer. *Briefings Bioinf.* **2022**, *23* (5), bbac162.

(25) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2017**, arXiv:1609.02907v4.

(26) Liang, Y.; Zhang, Z.-Q.; Liu, N.-N.; Wu, Y.-N.; Gu, C.-L.; Wang, Y.-L. MAGCNSE: Predicting lncRNA-Disease Associations Using Multi-View Attention Graph Convolutional Network and Stacking Ensemble Model. *BMC Bioinf.* **2022**, *23* (1), 189.

(27) Tang, X.; Luo, J.; Shen, C.; Lai, Z. Multi-View Multichannel Attention Graph Convolutional Network for miRNA-Disease Association Prediction. *Briefings Bioinf.* **2021**, *22* (6), bbab174.

(28) Wang, W.; Chen, H. Predicting miRNA-Disease Associations Based on lncRNA-miRNA Interactions and Graph Convolution Networks. *Briefings Bioinf.* **2023**, *24* (1), bbac495.

(29) Zhao, X.; Zhao, X.; Yin, M. Heterogeneous Graph Attention Network Based on Meta-Paths for lncRNA-Disease Association Prediction. *Briefings Bioinf.* **2022**, *23* (1), bbab407.

(30) Zhao, X.; Wu, J.; Zhao, X.; Yin, M. Multi-View Contrastive Heterogeneous Graph Attention Network for lncRNA-Disease Association Prediction. *Briefings Bioinf.* **2023**, *24* (1), bbac548.

(31) Xuan, P.; Zhao, Y.; Cui, H.; Zhan, L.; Jin, Q.; Zhang, T.; Nakaguchi, T. Semantic Meta-Path Enhanced Global and Local Topology Learning for lncRNA-Disease Association Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2023**, *20* (2), 1480−1491.

(32) Xuan, P.; Bai, H.; Cui, H.; Zhang, X.; Nakaguchi, T.; Zhang, T. Specific Topology and Topological Connection Sensitivity Enhanced Graph Learning for lncRNA-Disease Association Prediction. *Comput. Biol. Med.* **2023**, *164*, 107265.

(33) Xuan, P.; Zhan, L.; Cui, H.; Zhang, T.; Nakaguchi, T.; Zhang, W. Graph Triple-Attention Network for Disease-Related LncRNA Prediction. *IEEE J. Biomed. Health Inf.* **2022**, *26* (6), 2839−2849.

(34) Schriml, L. M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; Bisordi, K.; Campion, N.; Hyman, B.; Kurland, D.; Oates, C. P.; Kibbey, S.; Sreekumar, P.; Le, C.; Giglio, M.; Greene, C. Human Disease

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112

Ontology 2018 Update: Classification, Content and Workflow Expansion. *Nucleic Acids Res.* **2019**, *47* (D1), D955−D962.
(35) Zhou, Y.; Wang, X.; Yao, L.; Zhu, M. LDAformer: Predicting lncRNA-Disease Associations Based on Topological Feature Extraction and Transformer Encoder. *Briefings Bioinf.* **2022**, *23* (6), bbac370.
(36) Gao, Y.; Shang, S.; Guo, S.; Li, X.; Zhou, H.; Liu, H.; Sun, Y.; Wang, J.; Wang, P.; Zhi, H.; Li, X.; Ning, S.; Zhang, Y. Lnc2Cancer 3.0: An Updated Resource for Experimentally Supported lncRNA/circRNA Cancer Associations and Web Tools Based on RNA-Seq and scRNA-Seq Data. *Nucleic Acids Res.* **2021**, *49* (D1), D1251−D1258.
(37) Bao, Z.; Yang, Z.; Huang, Z.; Zhou, Y.; Cui, Q.; Dong, D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* **2019**, *47*, D1034−D1037.
(38) Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3.0: A Database for Experimentally Supported Human microRNA-Disease Associations. *Nucleic Acids Res.* **2019**, *47* (D1), D1013−D1017.
(39) Li, J.-H.; Liu, S.; Zhou, H.; Qu, L.-H.; Yang, J.-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acid Res.* **2014**, *42*, D92−D97.
(40) Wang, J. Z.; Du, Z.; Payattakool, R.; Yu, P. S.; Chen, C.-F. A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics* **2007**, *23* (10), 1274−1281.
(41) van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E. Gaussian Interaction Profile Kernels for Predicting Drug-Target Interaction. *Bioinformatics* **2011**, *27* (21), 3036−3043.
(42) Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I. S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arxiv:1807.06521v2.
(43) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; NeurIPS Proceedings, 2017.
(44) Rennie, J. D. M.; Shih, L.; Teevan, J.; Karger, D. R. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning*; ICML, 2003.
(45) Cramer, J. S. The Origins of Logistic Regression. In *Tinbergen Inst. Discuss. Pap.*; SSRN, 2002.
(46) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD, 2016, pp 785−794.
(47) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273−297.
(48) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(49) Wu, X.; Lan, W.; Chen, Q.; Dong, Y.; Liu, J.; Peng, W. Inferring LncRNA-Disease Associations Based on Graph Autoencoder Matrix Completion. *Comput. Biol. Chem.* **2020**, *87*, 107282.
(50) Zeng, M.; Lu, C.; Zhang, F.; Li, Y.; Wu, F.-X.; Li, Y.; Li, M. SDLDA: lncRNA-Disease Association Prediction Based on Singular Value Decomposition and Deep Learning. *Methods* **2020**, *179*, 73−80.
(51) He, J.; Li, M.; Qiu, J.; Pu, X.; Guo, Y. HOPEXGB: A Consensual Model for Predicting miRNA/lncRNA-Disease Associations Using a Heterogeneous Disease-miRNA-lncRNA Information Network. *J. Chem. Inf. Model.* **2023**, *64*, 2863−2877.
(52) Sheng, N.; Wang, Y.; Huang, L.; Gao, L.; Cao, Y.; Xie, X.; Fu, Y. Multi-Task Prediction-Based Graph Contrastive Learning for Inferring the Relationship among lncRNAs, miRNAs and Diseases. *Briefings Bioinf.* **2023**, *24* (5), bbad276.
(53) Li, G.; Bai, P.; Liang, C.; Luo, J. Node-Adaptive Graph Transformer with Structural Encoding for Accurate and Robust lncRNA-Disease Association Prediction. *BMC Genomics* **2024**, *25* (1), 73.
(54) Wang, S.; Hui, C.; Zhang, T.; Wu, P.; Nakaguchi, T.; Xuan, P. Graph Reasoning Method Based on Affinity Identification and Representation Decoupling for Predicting lncRNA-Disease Associations. *J. Chem. Inf. Model.* **2023**, *63* (21), 6947−6958.
(55) Xuan, P.; Lu, S.; Cui, H.; Wang, S.; Nakaguchi, T.; Zhang, T. Learning Association Characteristics by Dynamic Hypergraph and Gated Convolution Enhanced Pairwise Attributes for Prediction of Disease-Related lncRNAs. *J. Chem. Inf. Model.* **2024**, *64*, 3569−3578.
(56) Shlens, J. A Tutorial on Principal Component Analysis. *arXiv* **2014**, arXiv:1404.1100v1.
(57) Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Networks* **2000**, *13* (4−5), 411−430.
(58) Vempala, S. S. *The Random Projection Method*; American Mathematical Soc., 2005.
(59) Lin, X.; Lu, Y.; Zhang, C.; Cui, Q.; Tang, Y.-D.; Ji, X.; Cui, C. LncRNADisease v3.0: An Updated Database of Long Non-Coding RNA-Associated Diseases. *Nucleic Acids Res.* **2024**, *52* (D1), D1365−D1369.
(60) Zhou, B.; Ji, B.; Shen, C.; Zhang, X.; Yu, X.; Huang, P.; Yu, R.; Zhang, H.; Dou, X.; Chen, Q.; Zeng, Q.; Wang, X.; Cao, Z.; Hu, G.; Xu, S.; Zhao, H.; Yang, Y.; Zhou, Y.; Wang, J. EVLncRNAs 3.0: An Updated Comprehensive Database for Manually Curated Functional Long Non-Coding RNAs Validated by Low-Throughput Experiments. *Nucleic Acids Res.* **2024**, *52* (D1), D98−D106.

https://doi.org/10.1021/acsomega.4c05365
ACS Omega 2024, 9, 35100−35112