RESEARCH **Open Access**

LDAGM: prediction lncRNA-disease asociations by graph convolutional auto-encoder and multilayer perceptron based on multi-view heterogeneous networks

Bing Zhang[1†], Haoyu Wang[1*†], Chao Ma[1], Hai Huang[1], Zhou Fang[2] and Jiaxing Qu[2]

†Bing Zhang and Haoyu Wang have contributed equally to this work.

*Correspondence: 13204556500@163.com

[1] Harbin University of Science and Technology, Harbin 150006, Heilongjiang province, China
[2] Cyberspace Research Center, Harbin 150001, Heilongjiang province, China

Abstract

Background: Long non-coding RNAs (lncRNAs) can prevent, diagnose, and treat a variety of complex human diseases, and it is crucial to establish a method to efficiently predict lncRNA-disease associations.

Results: In this paper, we propose a prediction method for the lncRNA-disease association relationship, named LDAGM, which is based on the Graph Convolutional Autoencoder and Multilayer Perceptron model. The method first extracts the functional similarity and Gaussian interaction profile kernel similarity of lncRNAs and miRNAs, as well as the semantic similarity and Gaussian interaction profile kernel similarity of diseases. It then constructs six homogeneous networks and deeply fuses them using a deep topology feature extraction method. The fused networks facilitate feature complementation and deep mining of the original association relationships, capturing the deep connections between nodes. Next, by combining the obtained deep topological features with the similarity network of lncRNA, disease, and miRNA interactions, we construct a multi-view heterogeneous network model. The Graph Convolutional Autoencoder is employed for nonlinear feature extraction. Finally, the extracted nonlinear features are combined with the deep topological features of the multi-view heterogeneous network to obtain the final feature representation of the lncRNA-disease pair. Prediction of the lncRNA-disease association relationship is performed using the Multilayer Perceptron model. To enhance the performance and stability of the Multilayer Perceptron model, we introduce a hidden layer called the aggregation layer in the Multilayer Perceptron model. Through a gate mechanism, it controls the flow of information between each hidden layer in the Multilayer Perceptron model, aiming to achieve optimal feature extraction from each hidden layer.

Conclusions: Parameter analysis, ablation studies, and comparison experiments verified the effectiveness of this method, and case studies verified the accuracy of this method in predicting lncRNA-disease association relationships.

Keywords: LncRNA-disease associations, Graph convolutional auto-encoder, Multilayer perceptron, Deep topological feature extraction, Feature aggregation

Background

Long non-coding RNA (lncRNA) is a class of RNA molecules that do not encode proteins but have a transcript length exceeding 200 nucleotides. They play crucial roles in regulating various biological processes, including acting as molecular scaffolds in the nucleus, facilitating alternative splicing, modulating chromosome structure, and regulating translation in the cytoplasm. Additionally, they can promote or inhibit mRNA degradation and sequester miRNAs, among other functions [1]. Numerous experiments have illustrated the close association between lncRNAs and the development of diverse diseases. For instance, lncRNA KTN1-AS1 promotes non-small cell carcinoma progression by sponging miR-130a-5p and activating PDPK1 [2]. Similarly, lncRNA LINC00460 suppresses ANXA3 expression by up-regulating miR-2-433p, thereby impeding colon cancer epithelial-mesenchymal transition [3]. lncRNA UCA1 regulates apoptosis in cancer cells by modulating EZH2 activation of the PI3K/AKT pathway, leading to cisplatin resistance. UCA1 emerges as a promising therapeutic target for managing gastric cancer [4]. Furthermore, the upregulation of lncRNA IFNG-AS1 expression correlates with an elevated risk of coronary artery disease [5].

In predicting lncRNA-disease associations, traditional experimental methods are often costly, labor-intensive, and prone to uncertainty. However, with the advancement of high-throughput sequencing technology, numerous databases focusing on lncRNA-disease associations have emerged, including LncRNADisease [6], Lnc2Cancer [7], NONCODE [8], and OMIM [9]. These databases house extensive information regarding disease semantics and lncRNA-disease associations. Therefore, employing machine learning algorithms to predict lncRNA-disease associations offers a valuable approach to analyzing these relationships more rapidly and comprehensively.

For predicting lncRNA-disease association relationships, existing computational methods can be categorized into matrix decomposition-based methods, network-based methods, random walk methods, machine learning, and deep learning methods [10–15]. In the realm of matrix decomposition-based approaches, Lu et al. [16] devised an inductive matrix completion model. This model constructs matrices by amalgamating lncRNA-disease, disease-gene, and gene-gene interactions, then extracts primary feature vectors to complete the correlation matrices. Fu et al. [17] introduced a matrix decomposition model known as MFLDA, which decomposes heterogeneous data sources via matrix triple decomposition of data matrices into low-rank matrices, thus exploring and leveraging their intrinsic structure. Xuan et al. [18] integrated three association networks: lncRNA-miRNA, miRNA-disease, and lncRNA-disease, to form a disease-weighted association network. They employed probabilistic matrix decomposition to infer potential lncRNA-disease associations.

In the domain of network-based approaches, Yang et al. [19] established a dichotomous network comprising coding, non-coding genes, and diseases. They leveraged known associations between diseases and causative genes and applied a propagation algorithm to uncover latent lncRNA-disease associations within this network. Li et al. [20] integrated lncRNA-disease association probability matrices with integrated disease and lncRNA similarities. They proposed a model called NCPLDA, based on network similarity projection, for predicting unknown lncRNA-disease associations. Sheng et al. [21] proposed a multi-task prediction graph comparison learning model for GCLMTP.

This model first constructs heterogeneous graphs of lncRNAs, miRNAs, and diseases, extracts potential topological features from them based on graph comparison learning, and then performs prediction of association relationships.

In the realm of random walk-based approaches, Chen et al. [22] introduced the IBWRLDA model, which integrates lncRNA expression similarity and disease semantic similarity to establish the initial probability vector of Random Walk with Restart (RWR). They employ an improved restart random walk algorithm for lncRNA-disease association prediction. Sun et al. [23] proposed the RWRlncD framework, a global network-based computational approach for inferring human lncRNA-disease associations. This method implements a random walk algorithm and restart method on the lncRNA functional similarity network. Yu et al. [24] developed the BRWLDA model for predicting associations between lncRNAs and complex diseases. BRWLDA utilizes the lncRNA functional similarity network, disease network, and available lncRNA-disease associations to construct a directed bi-relational network. Double random walks are then applied on this bi-relational network for association prediction. Sheng et al. [25] proposed the VADLP model, which first constructs a three-layer heterogeneous graph, extracts topological features using random walks, learns hidden topological relationships with a convolutional autoencoder, and models feature distribution with a variance autoencoder for association prediction.

In the realm of machine learning and deep learning methods, Wang et al. [26] employed an auto-encoder neural network to obtain optimal feature vectors for lncRNA-disease pairs. These vectors were then input into a deep random forest to predict potential lncRNA-disease associations. Yuan et al. [27] initially computed similarity matrices for lncRNAs, genes, and diseases, integrating them. They then utilized a neural network to learn the nonlinear features of the integrated network, extracting neighborhood information to derive similarity scores. These scores were ranked to predict lncRNA-disease associations. Lan et al. [28] introduced a computational model called LDICDL. This model first utilizes an auto-encoder to denoise lncRNA and disease feature information. Then, it employs matrix decomposition for association relationship prediction. To address the limitations of matrix decomposition, a hybrid model was utilized. Shi et al. [29] proposed an end-to-end model named VGAELDA, which integrates variational inference and a graph autoencoder for predicting lncRNA-disease associations. Xuan et al. [30] introduced GCNLDA, a graph convolutional network and convolutional neural network-based approach for inferring the association relationship of disease-associated lncRNA candidate genes. Lu et al. [31] proposed a computational framework for LDAEXC, which uses a deep autoencoder for feature extraction and inputs the extracted features into XGBoost for final prediction. Sheng et al. [32] proposed a multichannel attention self-encoder-based model to predict the association between lncRNA and disease. The model first constructs a lncRNA-miRNA-disease complex graph, then utilizes a graph self-encoder to learn multiple representations from it, and finally employs a Random Forest classifier to make predictions.

Although the methods mentioned above have achieved satisfactory results, they still have some shortcomings. Matrix decomposition and network-based methods often overlook the nonlinear features within the data. Random walk methods can capture nonlinear features, but they are susceptible to propagation errors and the influence of data noise. Machine learning methods sometimes struggle to uncover deeper information within the data and perform poorly with sparse or noisy data, similar to random walk algorithms. Deep learning methods, while effective, often require complex hidden layers to fit the data, resulting in poor performance with excessively sparse or noisy data. Additionally, deep learning methods entail significant computational requirements due to the complexity of hidden layers.

To address the shortcomings of existing methods, we propose LDAGM, a method for predicting potential associations between lncRNA and disease. Firstly, we extract functional similarity and Gaussian interaction profile kernel similarity of lncRNA and miRNA, as well as semantic similarity and Gaussian interaction profile kernel similarity of diseases. This results in the construction of six homogeneous networks. Leveraging the method of deep topology feature extraction proposed in this paper, we fuse these six homogeneous networks to extract deep topological features, thereby expanding the data feature space and mitigating issues stemming from sparse network nodes. Subsequently, we integrate the fused network with similar networks of lncRNA, disease, and miRNA interactions to construct a multi-view heterogeneous network. Nonlinear features are then extracted from this network using the graph convolutional autoencoder (GCN-AE) model. Finally, based on positive and negative lncRNA-disease pairs, we obtain the final feature representation of the lncRNA-disease pairs by combining the deep topological features of the multi-view heterogeneous network with the nonlinear features extracted by the GCN-AE. These representations are subsequently inputted into a multilayer perceptron (MLP) model for predicting lncRNA-disease association relationships.

To enhance the performance and stability of the MLP model, we propose an aggregation layer within the MLP architecture to aggregate and regulate the flow of information between each hidden layer. Each hidden layer corresponds to an aggregation layer, with the latter receiving outputs from both the corresponding hidden layer and the preceding aggregation layer. Through the utilization of forgetting gates and input gates, the data received by the aggregation layer undergoes filtering and integration to generate the output of the current aggregation layer. Ultimately, the output of the final aggregation layer provides the association score of lncRNA-disease pairs. To verify the effectiveness of the proposed method, we conducted ablation experiments on the proposed deep topological feature extraction and aggregation layer. Additionally, comparison experiments were performed with four models proposed in recent years to evaluate the overall performance of the LDAGM model. The main innovations and contributions of this study are summarized as follows:

(1) This paper proposes a deep topological feature extraction method for integrating the functional similarity and Gaussian interaction profile kernel similarity of lncRNAs and miRNAs, as well as the semantic similarity and Gaussian interaction profile kernel similarity of diseases.

(2) A graph-convolutional auto-encoder model is employed to extract nonlinear features from the multi-view heterogeneous network. These nonlinear features are then combined with deep topological features to obtain the final feature representation of lncRNA-disease pairs.

(3) An aggregation layer is introduced within the multilayer perceptron model to aggregate and regulate the flow of information between each hidden layer. This enables each hidden layer to extract optimal features.

(4) Following a series of experimental comparisons, the effectiveness of the method described in this paper was confirmed.

Results and discussion

Experimental settings

In this section, a series of experiments are conducted using 5-fold cross-validation (5-CV) to compare and verify the superior performance of LDAGM. 5-CV splits the dataset into five disjoint subsets, with four subsets used for training the model and the remaining subset used for testing. All positive samples are selected and split according to 5-CV. To mitigate the negative effects caused by category imbalance, an equal number of negative samples as positive samples are randomly selected to train the prediction model. During testing, we randomly remove a portion of known associations from a positive sample of the test set and then test the trained model to evaluate the accuracy of LDAGM in mining potential associative relationships.

The experiments proceed as follows: firstly, evaluating the results of various hyperparameter settings on LDAGM; secondly, comparing LDAGM with four state-of-the-art algorithms from recent years; thirdly, conducting ablation experiments on LDAGM to validate the effectiveness of the proposed modules; and finally, analyzing case studies of five diseases: lung cancer, breast cancer, prostate cancer, hepatocellular carcinoma, and osteosarcoma by utilizing open-source bioinformatics databases and calibration to verify the accuracy of LDAGM.

In the experiments mentioned above, a learning rate of 1e-2 and a weight decay of 1e-5 were chosen for the training process. The model's effectiveness was evaluated using seven classical metrics: AUC, AUPR, MCC, ACC, Precision, Recall, and F1-Score.

Effect of parameters

This section evaluates the impact of four parameters on the experimental results: the number of neurons, the number of hidden layers, and the dropout rate in the MLP, along with the embedding dimension of the GCN-AE. These parameters aim to achieve optimal algorithm performance. The number of neurons and hidden layers in the MLP affect the model's fitting to the data. The Dropout rate influences the model's robustness, with an appropriate rate enhancing its generalization capability. The embedding dimension of the GCN-AE affects the representation of features and the complexity of learned features by the model.

We consider the number of neurons to range from {5, 10, 20, 30, 40}, the number of hidden layers to range from {1, 2, 4, 6, 8}, the dropout rate to vary from 0 to 1 with a step size of 0.1 for each change, and the GCN-AE embedding dimension to range from {16, 32, 64, 128, 256}. As depicted in Fig. 1 for Dataset 1, the optimal combination is achieved when the number of neurons is set to 40, the number of hidden layers is set to 2, the dropout rate is set to 0.1, and the GCN-AE embedding dimension is set to 128.

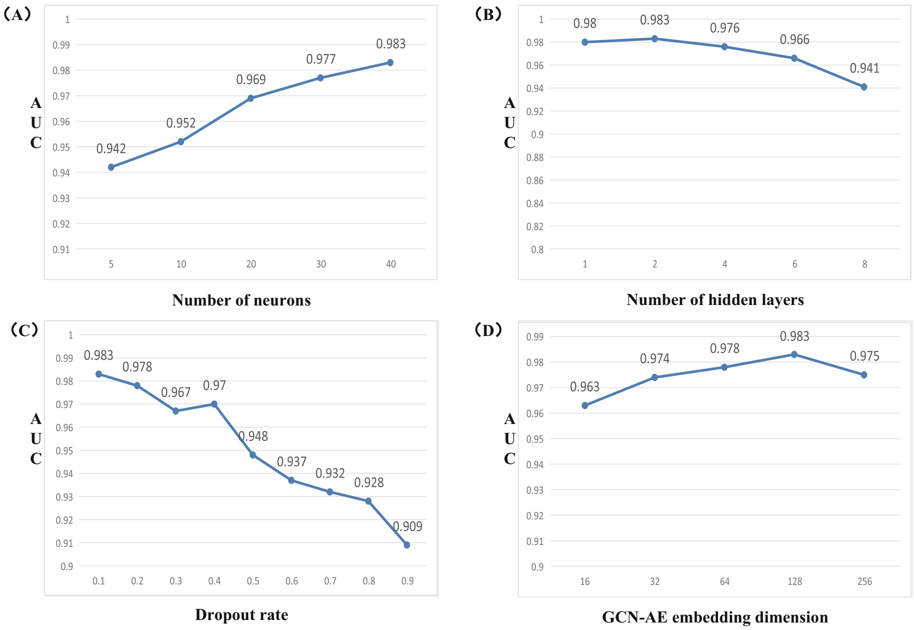

Fig. 1 Performance of LDAGM using different parameters. **A** Comparison of the AUC values under different numbers of neurons. **B** Comparison of the AUC values under different numbers of hidden layers. **C** Comparison of the AUC values under different Dropout rates. **D** Comparison of the AUC values under different GCN-AE embedding dimension

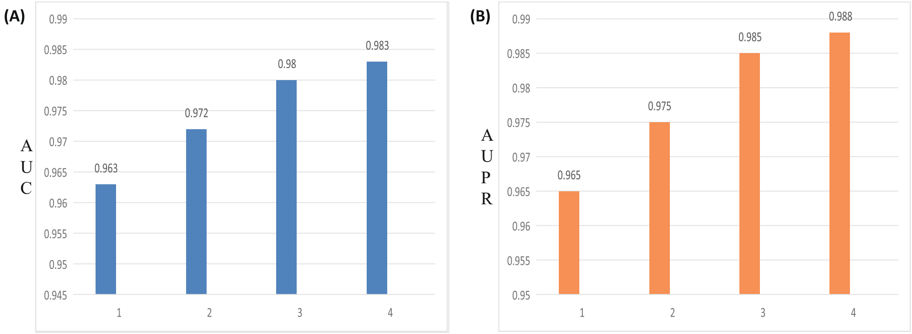

Fig. 2 Performance of LDAGM across different layers of deep topological feature extraction. **A** Comparison of AUC values across varying numbers of deep topological feature layers. **B** Comparison of AUPR values across varying numbers of deep topological feature layers

Optimal Number of Deep Topological Feature Layers

The number of deep topological feature layers has varying impacts on the results. To determine the optimal number of deep topological feature layers for achieving the best outcomes, we explore different values ranging from {1, 2, 3, 4} and validate their effects. As depicted in Fig. 2, the results reach their optimum when the number of deep topological feature layers is set to 2.

Ablation studies

To verify the effectiveness of the deep topological feature extraction and aggregation layer, this section compares LDAGM with its three variants: (1) LDAGM-a, without the deep topological feature extraction and aggregation layer; (2) LDAGM-b, where deep

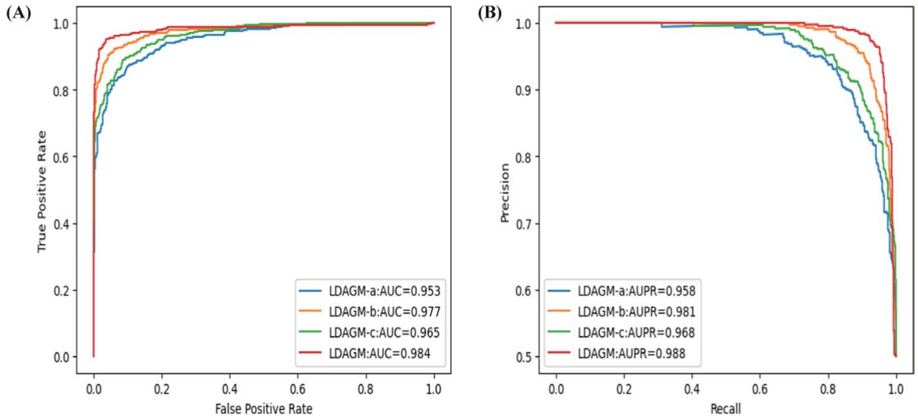

Fig. 3 Comparison of ROC curves (**A**) and PR curves (**B**) among LDAGM and its three variants

Table 1 Comparison of the evaluation metrics between LDAGM and its three variants, the results of LDAGM are optimal, as indicated in bold

Method	AUC	AUPR	MCC	ACC	Precision	Recall	F1-Score
LDAGM-a	0.952±0.0074	0.957±0.0032	0.881±0.0045	0.899±0.0078	0.859±0.0043	0.878±0.0021	0.763±0.0041
LDAGM-b	0.977±0.0062	0.981±0.0041	0.915±0.0052	0.920±0.0063	0.859±0.0074	0.910±0.0023	0.836±0.0043
LDAGM-c	0.965±0.0183	0.968±0.0145	0.896±0.0 0212	0.911±0.0171	0.877±0.0230	0.894±0.0244	0.793±0.0165
LDAGM	**0.983±0.0058**	**0.988±0.0047**	**0.930±0.0233**	**0.941±0.0122**	**0.983±0.0106**	**0.925±0.023**	**0.939±0.0131**

The bold number is the highest value of each column and its clarifes the superiority of our mode

topological feature extraction is utilized but not the aggregation layer; and (3) LDAGM-c, which employs the aggregation layer but not deep topological feature extraction.

As depicted in Fig. 3 and Table 1, LDAGM equipped with both deep topological feature extraction and the aggregation layer exhibits superior performance. Deep topological feature extraction delves deeply into the association relationships among nodes in heterogeneous networks, thereby enriching the feature representation of such networks. Meanwhile, the aggregation layer facilitates controlled information transfer between each hidden layer through three gate mechanisms: input gate, forget gate, and update gate. This enables each hidden layer to extract optimal features. The synergistic combination of these two components leads to an overall enhancement in model performance.

Comparison with other classifiers

We compared the performance of LDAGM with other classifiers, including Support Vector Machine (SVM), Random Forest (RF), Graph Attention Network (GAN), EXtreme Gradient Boosting (XGBoost). As shown in Table 2, the performance of LDAGM achieves the optimum.

Comparison of training set and test set with different ratios

To evaluate the performance of LDAGM on different proportions of training and testing sets, this section conducts training and testing by splitting the dataset into training and

Table 2 Comparison LDAGM with other classifiers, the results of LDAGM are optimal, as indicated in bold

Method	AUC	AUPR	MCC	ACC	Precision	Recall	F1-Score
SVM	0.930±0.0065	0.920±0.0136	0.862±0.0087	0.916±0.0063	0.924±0.0125	0.783±0.0036	0.856±0.0045
RF	0.864±0.0145	0.853±0.0365	0.804±0.0754	0.884±0.0452	0.836±0.0478	0.840±0.0136	0.812±0.0175
GAN	0.949±0.0025	0.952±0.0069	0.906±0.0078	0.891±0.0085	0.939±0.0069	0.848±0.0074	0.899±0.0085
XGBoost	0.934±0.0074	0.924±0.0051	0.883±0.0062	0.905±0.0084	0.925±0.0093	0.825±0.0065	0.901±0.0071
LDAGM	**0.983±0.0058**	**0.988±0.0047**	**0.930±0.0233**	**0.941±0.0122**	**0.983±0.0106**	**0.925±0.023**	**0.939±0.0131**

The bold number is the highest value of each column and its clarifes the superiority of our mode

Table 3 Comparison of training set and test set with different ratios

ratio	AUC	AUPR	MCC	ACC	Precision	Recall	F1-Score
1:2	0.983±0.0058	0.988±0.0047	0.930±0.0233	0.941±0.0122	0.983±0.0106	0.925±0.023	0.939±0.0131
1:4	0.982±0.0072	0.876±0.0045	0.819±0.0043	0.959±0.0147	0.752±0.0108	0.927±0.0086	0.817±0.0087
1:6	0.979±0.0053	0.770±0.0033	0.705±0.0076	0.943±0.0085	0.634±0.0139	0.894±0.0144	0.752±0.0085
1:8	0.981±0.0046	0.670±0.0021	0.600±0.0062	0.952±0.0011	0.509±0.0136	0.913±0.0153	0.664±0.0065
1:10	0.977±0.0056	0.640±0.0042	0.578±0.0041	0.934±0.0072	0.403±0.0192	0.883±0.0175	0.543±0.0055

testing sets at ratios of {1:2, 1:4, 1:6, 1:8, 1:10}. The corresponding metrics are presented in Table 3.

Comparison with other state-of-the-art methods

To assess the performance effectiveness of the LDAGM algorithm in predicting lncRNA-disease association relationships, we compared it with four classical algorithms: DMFLDA [14], SDLDA [33], GAERF [34], and MAGCNSE [35].

DMFLDA employs the MLP model to predict lncRNA-disease associations via deep matrix decomposition. It utilizes a series of nonlinear hidden layers to directly learn potential features from the lncRNA-disease interaction matrix, aiming for a more accurate feature representation. SDLDA utilizes singular value decomposition and a deep learning framework to extract both linear and nonlinear features of lncRNA and disease. By integrating these features, it obtains a more accurate representation of lncRNA and disease pairs, which is then used for association prediction. GAERF first extracts nonlinear features of lncRNA and disease using a graph self-encoder. Subsequently, it employs a random forest classifier to learn these features and predict lncRNA-disease associations. MAGCNSE extracts features from the multi-view matrix of lncRNA and disease using a graph convolutional attention network. These features are then inputted into a stacked integrated classifier composed of multiple traditional machine-learning classifiers to make the final association prediction.

As illustrated in Fig. 4 and Table 4, LDAGM achieves AUC scores of 0.983, 0.953 on Dataset 1and Dataset 2, respectively. Additionally, its AUPR scores are 0.988, 0.951, respectively, outperforming the other methods.

As shown in Tables 5 and 6, paired t-tests of model performance metrics for different datasets confirm that LDAGM is statistically significant when compared to other competing methods.

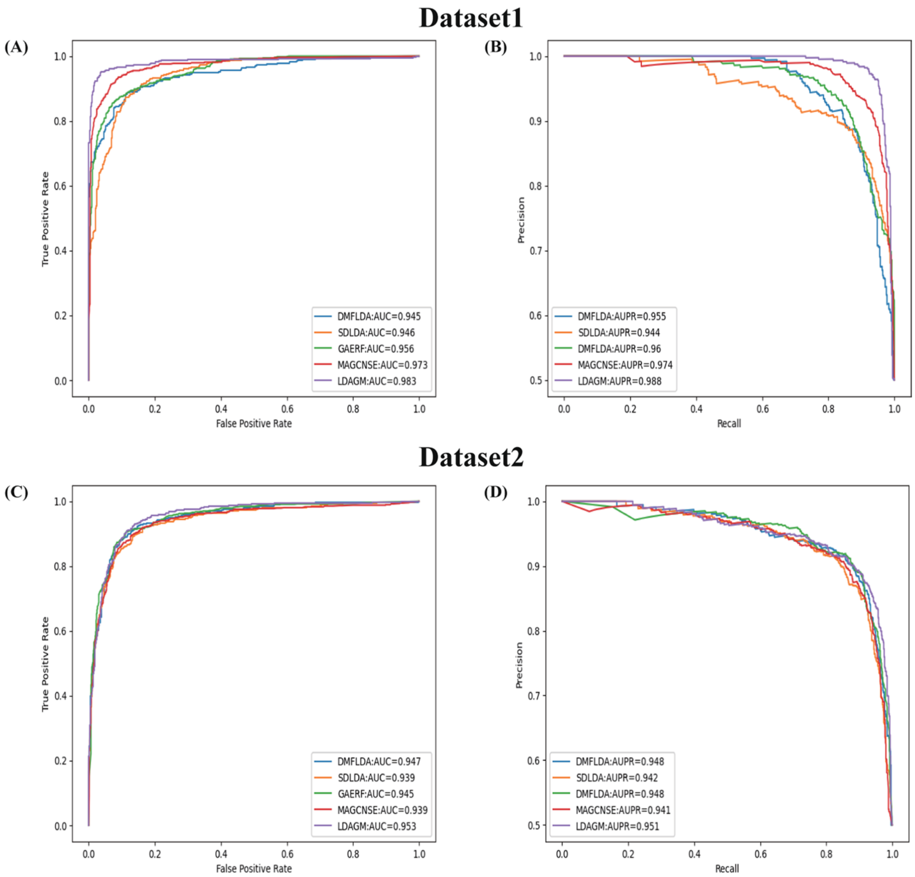

Fig. 4 Performance comparison between LDAGM and other state-of-the-art methods. **A**, **B** Comparison of ROC curves, AUC values, PR curves, and AUPR values on Dataset 1. **C**, **D** Comparison of ROC curves, AUC values, PR curves, and AUPR values on Dataset 2

Table 4 LDAGM metrics on two datasets

Dataset	AUC	AUPR	MCC	ACC	Precision	Recall	F1-Score
Dataset 1	0.983±0.0058	0.988±0.0047	0.930±0.0233	0.941±0.0122	0.983±0.0106	0.925±0.023	0.939±0.0131
Dataset 2	0.953±0.0053	0.951±0.0036	0.770±0.0087	0.883±0.0044	0.915±0.0054	0.846±0.0069	0.879±0.0046

Table 5 Paired t-test between the performances of LDAGM and the competing methods for Dataset 1

Methods	*p* value on Dataset 1
DMFLDA	0.0001485
SDLDA	0.0027289
GAERF	0.0025345
MAGCNSE	0.0867533

Case studies

To further validate LDAGM's performance in predicting lncRNA-disease associations, a specific case study is conducted as follows: Step 1, five target diseases: lung

Table 6 Paired t-test between the performances of LDAGM and the competing methods for Dataset 2

Methods	*p* value on Dataset 2
DMFLDA	0.0133635
SDLDA	0.0006963
GAERF	0.0053563
MAGCNSE	0.0003097

Table 7 The top 15 predicted lung cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	HOTAIR	LncRNADisease v3.0 Lnc2Cancer v3.0
2	MALAT1	LncRNADisease v3.0 Lnc2Cancer v3.0
3	NEAT1	LncRNADisease v3.0 Lnc2Cancer v3.0
4	MEG3	LncRNADisease v3.0 Lnc2Cancer v3.0
5	BANCR	LncRNADisease v3.0 Lnc2Cancer v3.0
6	HULC	LncRNADisease v3.0
7	H19	LncRNADisease v3.0 Lnc2Cancer v3.0
8	PVT1	LncRNADisease v3.0 Lnc2Cancer v3.0
9	KCNQ1OT1	LncRNADisease v3.0
10	LINC00663	Lnc2Cancer v3.0
11	AFAP1-AS1	LncRNADisease v3.0 Lnc2Cancer v3.0
12	GAS5	LncRNADisease v3.0 Lnc2Cancer v3.0
13	MIR17HG	Unconfirmed
14	HOTTIP	LncRNADisease v3.0 Lnc2Cancer v3.0
15	CCAT2	LncRNADisease v3.0 Lnc2Cancer v3.0

cancer, breast cancer, prostate cancer, hepatocellular carcinoma, and osteosarcoma are selected. Step 2, all known positive samples in the dataset are selected, along with an equal number of negative samples from unknown lncRNA-disease pairs unrelated to the target diseases, to construct the training dataset. Step 3, all associations between the lncRNA and the target diseases are selected to construct the testing dataset. Step 4, after training on the training set, the test set is used for testing, with output scores sorted in descending order. Step 5, validation is conducted using LncRNADisease v3.0 [36] (http://www.rnanut.net/lncrnadisease/index.php/home) and Lnc2Cancer v3.0 [37] (http://bio-bigdata.hrbmu.edu.cn/lnc2cancer/). If no evidence is found in both databases, consultation of PubMed literature is pursued.

Lung cancer, a leading cause of global mortality, validates 13 out of the top 15 lung cancer-related lncRNAs listed in Table 7. For instance, PVT1, directly regulated by the transcription factor YY1, plays a pivotal role in lung cancer by promoting its expression through transcription activation [38]. AFAP1-AS1, significantly up-regulated in lung cancer, contributes to the disease by regulating molecules associated with actin filament integrity [39].

Breast cancer, prevalent among women, confirms 12 out of the top 15 lncRNAs associated with the disease listed in Table 8. For instance, the expression of TUG1

Table 8 The top 15 predicted breast cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	TUG1	Lnc2Cancer v3.0
2	MEG3	LncRNADisease v3.0 Lnc2Cancer v3.0
3	HOTAIR	LncRNADisease v3.0 Lnc2Cancer v3.0
4	MALAT1	LncRNADisease v3.0 Lnc2Cancer v3.0
5	BANCR	LncRNADisease v3.0 Lnc2Cancer v3.0
6	HOTTIP	LncRNADisease v3.0 Lnc2Cancer v3.0
7	UCA1	LncRNADisease v3.0 Lnc2Cancer v3.0
8	HULC	LncRNADisease v3.0 Lnc2Cancer v3.0
9	MIR17HG	Unconfirmed
10	TUSC7	LncRNADisease v3.0 Lnc2Cancer v3.0
11	PVT1	LncRNADisease v3.0 Lnc2Cancer v3.0
12	GAS5	LncRNADisease v3.0 Lnc2Cancer v3.0
13	HCG4	Unconfirmed
14	EWSAT1	Unconfirmed
15	H19	LncRNADisease v3.0 Lnc2Cancer v3.0

Table 9 The top 15 predicted prostate cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	H19	Lnc2Cancer v3.0
2	MEG3	Lnc2Cancer v3.0
3	XIST	Lnc2Cancer v3.0
4	GAS5	LncRNADisease v3.0 Lnc2Cancer v3.0
5	MALAT1	Lnc2Cancer v3.0
6	PVT1	Lnc2Cancer v3.0
7	TUG1	Lnc2Cancer v3.0
8	CDKN2B-AS1	Lnc2Cancer v3.0
9	NEAT1	Lnc2Cancer v3.0
10	HOTAIR	Lnc2Cancer v3.0
11	AFAP1-AS1	Lnc2Cancer v3.0
12	BCYRN1	Unconfirmed
13	CCAT1	Lnc2Cancer v3.0
14	MIR17HG	Lnc2Cancer v3.0
15	UCA1	Lnc2Cancer v3.0

significantly decreases in breast cancer and regulates sensitivity to cisplatin chemotherapy, making it a potential treatment target [40]. uca1, notably up-regulated in breast cancer, plays a pivotal role in disease progression by modulating the EZH2/p21 axis and the PI3K/AKT signaling pathway. Silencing UCA1 increases drug-resistant cell sensitivity to tamoxifen, presenting it as a potential treatment target [41].

Prostate cancer, prevalent among men, validates 14 out of the top 15 lncRNAs associated with the disease listed in Table 9. For instance, MEG3 promotes H3K27 trimethylation of EN2 by binding to EZH2, inhibiting prostate cancer development and serving as a potential treatment target [42]. HOTAIR, overexpressed in prostate cancer, acts as a prognostic predictor and promotes cancer cell metastasis. It can induce cancer cell apoptosis by regulating miR-125a-5p to release caspase2, making it a potential treatment target [43].

Detailed predictive scores with lncRNAs for the mentioned diseases are provided in Additional File 1: Table 1, Additional File 1: Table 2, and Additional File 1: Table 3. Predictive information concerning hepatocellular carcinoma is available in Additional File 1: Table 4, while information regarding osteosarcoma is presented in Additional File 1: Table 5.

Conclusions

The analysis of potential lncRNA-disease associations through computational methods aids in identifying disease biomarkers and enhancing preventative measures, thereby reducing labor costs and improving efficiency. This paper introduces a novel approach for predicting lncRNA-disease associations, named LDAGM. Initially, functional similarity and Gaussian spectral kernel similarity of lncRNA, miRNA, and semantic similarity and Gaussian spectral kernel similarity of disease are extracted. These six homogeneous networks are fused using deep topological feature extraction to achieve feature complementation. Subsequently, the fused homogeneous network is integrated with similar networks of lncRNA, disease, and miRNA interactions to construct a multi-view heterogeneous network. This network is then inputted into a graph convolutional autoencoder for nonlinear feature extraction. Nonlinear features are combined with deep topological features of the multi-view heterogeneous network to construct the final feature representation of lncRNA-disease pairs. The pairs are inputted into an MLP model for predicting lncRNA-disease associations. To enhance MLP model performance, an aggregation layer is added to aggregate and control information flow in each hidden layer, enabling optimal feature fitting to each layer. Experimental results demonstrate that deep topological feature extraction and the aggregation layer enhance overall model performance, with AUC scores of 0.983, 0.953, and AUPR scores of 0.988, 0.951, outperforming other methods on Dataset1 and Dataset2. Model accuracy in predicting lncRNA-disease associations is validated through case studies.

In future work, improvements will be made to both association relationship analysis and deep learning model construction. The method currently utilizes only lncRNA, disease, and miRNA association relationships, without comprehensive consideration of biological information such as protein information, RNA sequence information, and drug targeting effects. Additionally, enhancements can be made to the deep learning method for integrating and mining lncRNA-disease association relationships by introducing attention mechanisms and adversarial training on data.

Methods

Datasets

To validate the effectiveness of LDAGM, we evaluated it on two datasets:

(1) Dataset 1: We utilized the widely cited benchmark dataset introduced by Fu et al. [17], comprising 240 lncRNAs, 412 diseases, and 495 miRNAs. This dataset includes 2,697 lncRNA-disease association nodes sourced from LncRNADisease

[6], Lnc2Cancer [7], and GeneRIF [44]. Additionally, it incorporates 1002 lncRNA-miRNA association nodes from startBase v2.0 [45] and another 1002 lncRNA-miRNA association nodes from HMDD v2.0 [46]. Furthermore, it comprises 13562 miRNA-disease association nodes from HMDD v2.0 [46].

(2) Dataset 2: We employed a dataset generated by Zhou et al.[47], compassing 665 lncRNAs, 316 diseases, and 295 miRNAs. This dataset comprises 3833 lncRNA-disease association nodes sourced from Lnc2Cancer v3.0 [37] and LncRNADisease v2.0 [36]. Additionally, it includes 2108 lncRNA-miRNA association nodes from starBase v2.0 [45] and 8540 miRNA-disease association nodes from HMDD v3.0 [48].

Disease semantic similarity

From the Disease Ontology database DO [49], we obtained the semantic information of diseases, represented by a directed acyclic graph illustrating the parent–child relationship of diseases. This approach is based on the method proposed by Wang et al. [50] to compute disease similarity. For a disease d_i, let D_i denote the set containing disease d_i and all its ancestor terms. The contribution $W_{d_i}(d)$ of other disease nodes d to d_i in this set can be expressed by the following equation:

$$W_{d_i} = \begin{cases} 1 d_i = d \\ \max\limits_{d' \in \text{children of} d} \left(\Delta \times W_{d_i}\left(d'\right)\right) d_i \neq d \end{cases} \tag{1}$$

The symbol Δ represents the semantic contribution attenuation factor, which is set to 0.5 here. The contribution of disease d_i to itself has a value of 1, and the contribution of disease d to d_i decreases as the number of disease nodes spaced between them increases. As a result, the semantic value of disease d_i is represented by the following equation:

$$DS(d_i) = \sum_{d \in D_i} W_{d_i}(d) \tag{2}$$

According to the assumption that if two diseases have more intersecting nodes in the same set, then the stronger their correlation is, the semantic similarity of diseases d_i and d_j can be expressed as follows:

$$DSim(d_i, d_j) = \frac{\sum_{d \in D_i \cap D_j} \left(W_{d_i}(d) + W_{d_j}(d)\right)}{DS(d_i) + DS(d_j)} \tag{3}$$

LncRNA and miRNA functional similarity

Based on the hypothesis of Wang et al. [51] that diseases possessing similar phenotypes are more likely to be associated with functionally similar lncRNAs and miRNAs, the semantic similarity of diseases calculated above can be combined with the association between lncRNAs and miRNAs and diseases to calculate the functional similarity of lncRNAs and miRNAs with the following formula:

$$FSim(R_i, R_j) = \frac{\left[\sum_{i=1}^{n_1} \max_{1\leq j\leq n_2} (DSim(d_{1i}, d_{2j})) + \sum_{j=1}^{n_2} \max_{1\leq i\leq n_1} (DSim(d_{2j}, d_{1i}))\right]}{n_1 + n_2} \tag{4}$$

Where R_1 and R_2 are the two lncRNA or miRNA nodes for which similarity is to be calculated, and n_1 and n_2 are the number of disease nodes associated with R_i and R_j.

Gaussian interaction profile kernel similarity for LncRNA, MiRNA, and Disease

Similar lncRNAs or miRNAs are more likely to be associated with similar diseases, based on the method proposed by Chen et al [52], the Gaussian spectral nuclear similarity LGSim formula for lncRNAs is as follows:

$$LGSim = exp(-\gamma_l ||LD(i, :) - LD(j, :)||^2) \tag{5}$$

$$\gamma_l = 1/\left(\frac{1}{n_l}\sum_{i=1}^{n_l} ||LD(i, :)||^2\right) \tag{6}$$

where γ_l denotes the standardized core bandwidth for lncRNA similarity calculation which is generally set to 1, and n_l denotes the number of lncRNAs. Similarity for disease DGSim is computed as follows:

$$DGSim = exp(-\gamma_d ||LD(:, i) - LD(:, j)||^2) \tag{7}$$

$$\gamma_d = 1/\left(\frac{1}{n_d}\sum_{i=1}^{n_d} ||LD(:, i)||^2\right) \tag{8}$$

where γ_d denotes the standardized core bandwidth for disease similarity calculation, and n_d denotes the number of disease. Similarity for miRNAs MGSim is computed as follows:

$$MGSim = exp(-\gamma_m ||MD(i, :) - MD(j, :)||^2) \tag{9}$$

$$\gamma_m = 1/\left(\frac{1}{n_m}\sum_{i=1}^{n_m} ||LD(i :,)||^2\right) \tag{10}$$

where γ_m denotes the standardized core bandwidth for miRNA similarity calculation, and n_m denotes the number of disease.

LDAGM

Deep topology feature extraction

To cope with the problem of sparse network structure, we propose a multi-similarity network fusion method for deep topological feature extraction based on the already computed functional similarity and Gaussian interaction profile kernel similarity of lncRNA and miRNA and the semantic similarity and Gaussian interaction profile kernel similarity of disease, to realize the complementation of the network features, and

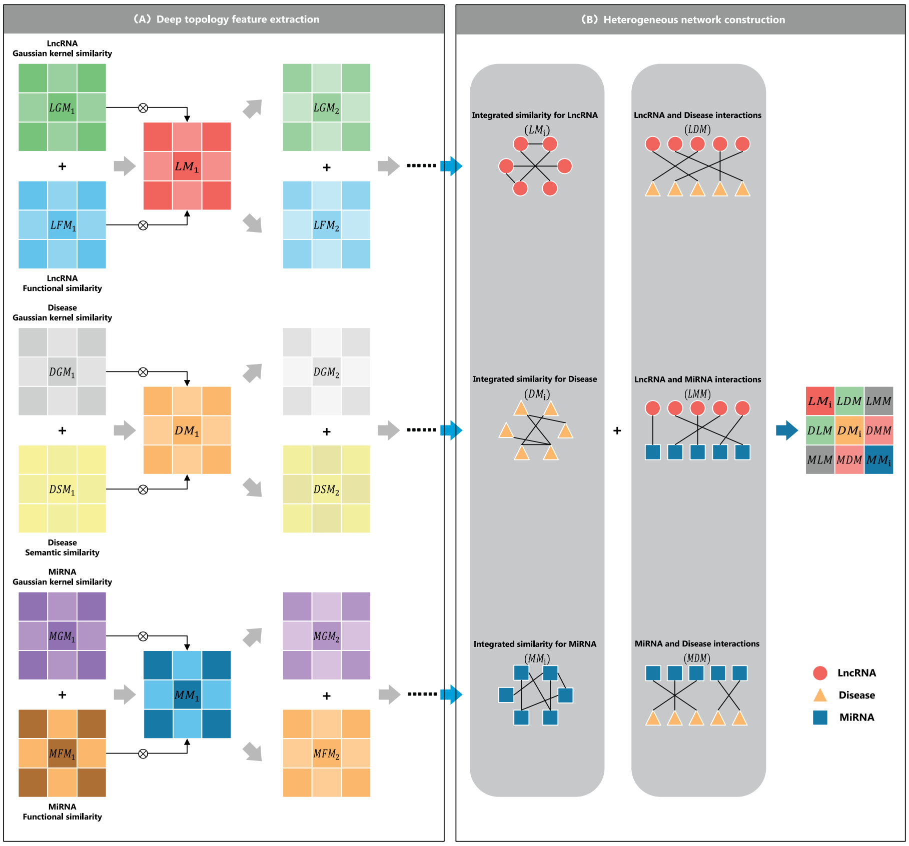
(A) Deep topology feature extraction
LncRNA Gaussian kernel similarity
LGM_1
+
LFM_1
LncRNA Functional similarity
LM_1
LGM_2
+
LFM_2
Disease Gaussian kernel similarity
DGM_1
+
DSM_1
Disease Semantic similarity
DM_1
DGM_2
+
DSM_2
MiRNA Gaussian kernel similarity
MGM_1
+
MFM_1
MiRNA Functional similarity
MM_1
MGM_2
+
MFM_2
(B) Heterogeneous network construction
Integrated similarity for LncRNA
(LM_1)
LncRNA and Disease interactions
(LDM)
Integrated similarity for Disease
(DM_1)
+
LncRNA and MiRNA interactions
(LMM)
Integrated similarity for MiRNA
(MM_1)
MiRNA and Disease interactions
(MDM)
LM_1 LDM LMM
DLM DM_1 DMM
MLM MDM MM_1
● LncRNA
▲ Disease
■ MiRNA

Fig. 5 Deep topological feature extraction and multi-view heterogeneous network construction. **A** Deep topological feature extraction based on fusion of lncRNA, miRNA functional similarity, Gaussian interaction profile kernel similarity and disease semantic similarity, Gaussian interaction profile kernel similarity. **B** Integration of the fused lncRNA, disease, and miRNA homogeneous networks with the similar networks of the three interactions to construct a multi-view heterogeneous network

to combine the fused deep homogeneous networks, namely, the three deep homogeneous networks, namely, lncRNA, disease, and miRNA, to obtain the multi-similar network. networks and the three interacting similarity networks are combined to obtain a multi-view heterogeneous network, as shown in Fig. 5.

Let lncRNA functional similarity adjacency matrix be LFM_1, Gaussian interaction profile kernel similarity adjacency matrix be LGM_1, and fused similarity adjacency matrix be LM_i. miRNA functional similarity adjacency matrix be MFM_1, Gaussian interaction profile kernel similarity adjacency matrix be MGM_1, and fused similarity adjacency matrix be MM_i, and disease semantic similarity adjacency matrix be DSM_1, Gaussian interaction profile kernel similarity adjacency matrix be DGM_1, and fused similarity adjacency matrix be DM_i, then deep topological feature extraction formula is as follows.

$$LM_1 = \frac{(LFM_1 + LGM_1)}{\max(LFM_1 + LGM_1)} \tag{11}$$

$$MM_1 = \frac{(MFM_1 + MGM_1)}{\max(MFM_1 + MGM_1)} \tag{12}$$

$$DM_1 = \frac{(DSM_1 + DGM_1)}{\max(DSM_1 + DGM_1)} \tag{13}$$

where $\max(\cdot)$ is the maximum operation. After the first layer of topological features is extracted, the functional similarity and Gaussian interaction profile kernel similarity neighboring matrices of lncRNA, miRNA and the semantic similarity and Gaussian interaction profile kernel similarity neighboring matrices of disease are updated with the following equations:

$$LFM_2 = LM_1 \otimes LFM_1 \tag{14}$$

$$LGM_2 = LM_1 \otimes LGM_1 \tag{15}$$

$$MFM_2 = MM_1 \otimes MFM_1 \tag{16}$$

$$MGM_2 = MM_1 \otimes MGM_1 \tag{17}$$

$$DSM_2 = DM_1 \otimes DSM_1 \tag{18}$$

$$DGM_2 = DM_1 \otimes DGM_1 \tag{19}$$

Where $\otimes$ is the matrix dot product operation, repeat the operation of Eq. 11, Eq. 12 and Eq. 13 to extract the second layer of topological features, and then continue to update the functional similarity, Gaussian interaction profile kernel similarity and semantic similarity of disease of lncRNA, miRNA, Gaussian interaction profile kernel similarity adjacency matrix, and keep repeating, to extract the deep topological features.

After extracting the deep topological features, they are integrated with similar networks of lncRNA, disease, and miRNA interactions in order to construct a multi-view heterogeneous network. The multi-view heterogeneous network is represented by the form of a neighbor-joining matrix with the following structure:

$$A_i = \begin{bmatrix} LM_i & LD & LM \\ DL & DM_i & DM \\ ML & MD & MM_i \end{bmatrix} \tag{20}$$

Where, LM_i, DM_i, MM_i are the topological features of lncRNA, disease and miRNA at layer i; LD is the lncRNA-disease association matrix; LM is the lncRNA-miRNA association matrix; DM is the disease-miRNA association matrix; and DL, ML, and MD are the transpositions of LD, LM and DM.

GCN-AE

After integrating the multi-view heterogeneous networks, each view heterogeneous network is sequentially fed into the GCN-AE for nonlinear feature extraction, which

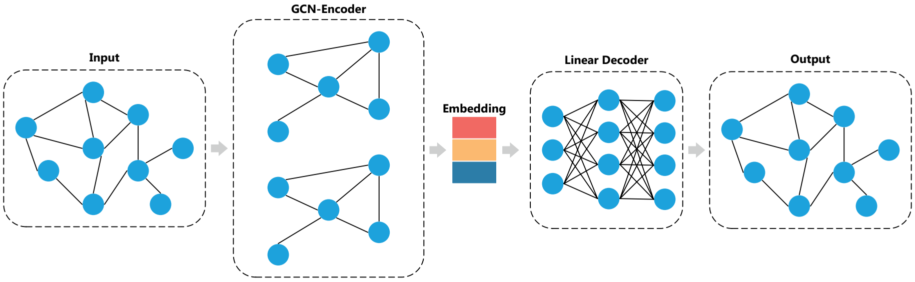

Fig. 6 Nonlinear feature extraction. The multi-view heterogeneous network is fed into the encoder, which undergoes a convolution operation to obtain the embedded form of the input data and undergoes a bilinear decoding layer to decode the output of the encoder for reduction

ensures that the information at each layer is fully learned and represented. The GCN-AE is divided into an encoder and a decoder, the encoder gets the low-dimensional embedded form of the input data, which can reflect the nonlinear relationships in the input data, and the decoder decodes the output of the encoder to restore the data, and the process is shown in Fig. 6.

At the encoder layer, the input data is first Laplace normalized to reduce the noise in the data with the following equation:

$$L = D^{-\frac{1}{2}} A_i D^{-\frac{1}{2}} \tag{21}$$

where D is the diagonal matrix consisting of the degrees of each row of A_i. After calculating the Laplace normalized matrix of A_i, a convolution operation is performed on it and the result of the convolution is linearly transformed to obtain the output of the encoder with the following equation:

$$R_e = ReLU[(A_i \times L)W + b] \tag{22}$$

where W is a learnable weight matrix, b is a learnable bias term, and $ReLU(\cdot)$ is a nonlinear activation function. After obtaining the output of the encoder, it is fed into the decoder, and the output of the encoder is decoded through the bilinear layer with the following equation:

$$R_d = [ReLU(R_e W + b)]W + b \tag{23}$$

After obtaining the output of the decoder, the loss between the output of the decoder and A_i is measured using the mean square error loss function, and the loss is reduced by continuous iterative training, and finally, more accurate encoder embedding features can be obtained, and the formula for the mean square error loss function is as follows:

$$Loss = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - y_i'\right)^2 \tag{24}$$

MLP

The nonlinear features are integrated with the deep topological features of the multi-view heterogeneous network to obtain the final feature representation of lncRNA-disease

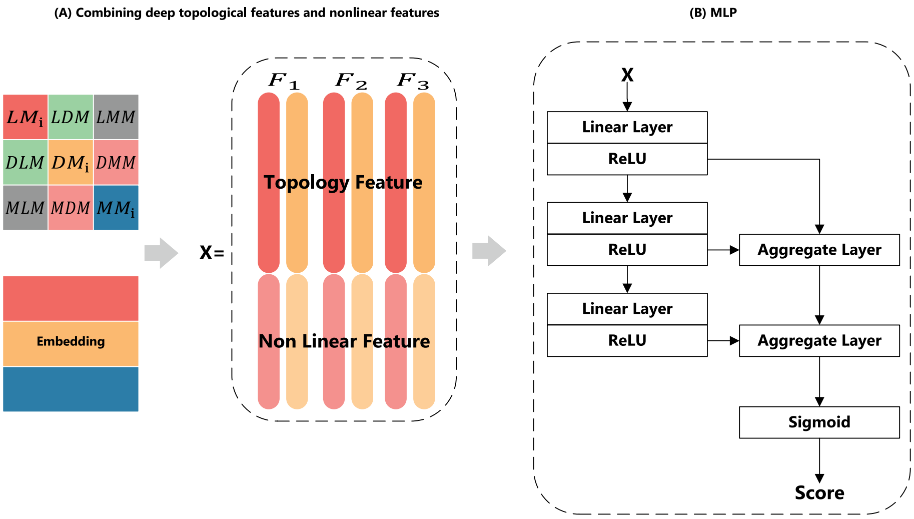

Fig. 7 Deep topological and nonlinear feature integration and MLP training. **A** Deep topological features and nonlinear features are integrated to obtain the final feature representation of lncRNA-disease pairs. **B** The integrated feature representations are input into the MLP model, and the final scores are obtained after a series of hidden and aggregated layers to fit the feature representations, and after a Sigmoid layer

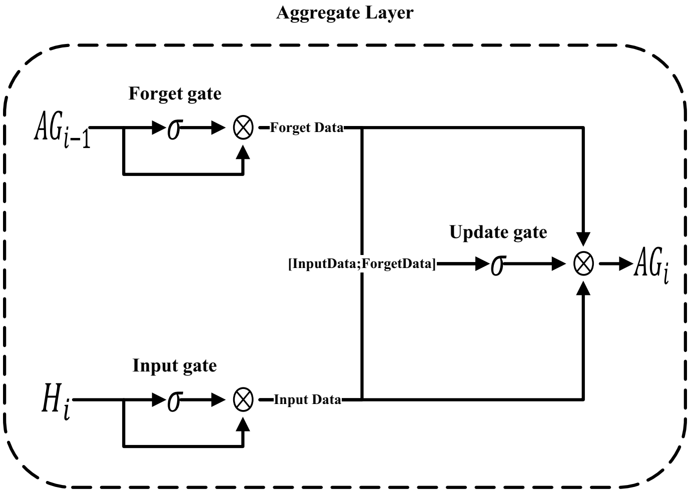

Fig. 8 Aggregate Layer. The output of the previous aggregate layer, AG_{i-1}, and the output of the current hidden layer, H_i, are passed through the forgetting gate and the input gate, where the features are filtered and the important features are retained. The update gate integrates the results of the forgetting gate and input gate outputs to get the output of the current aggregate layer

pairs, which are input into the MLP model for the prediction of lncRNA-disease association relationship. In order to improve the performance and stability of the MLP model, this paper proposes an aggregation layer in the MLP model, which is used to control the flow of information between each hidden layer, so that each hidden layer extracts the optimal features. Each hidden layer corresponds to an aggregation layer, and the aggregation layer receives inputs from the previous aggregation layer while receiving inputs from the current hidden layer, and the flow is shown in Fig. 7.

The MLP model consists of multiple hidden layers, each of which receives the output of the previous hidden layer as input and linearly transforms it to fit the data, and the formula for the linear transformation made by each hidden layer is as follows:

$$H_i = WH_{i-1} + b \tag{25}$$

where H_{i-1} is the output of the previous hidden layer, H_i is the output of the current hidden layer, and both W and b are learnable weight matrices.The aggregation layer consists of three gates, namely input gate, forget gate, and update gate, and the flow is shown in Fig. 8.

The forgetting gate is used to control the inflow of information from the previous aggregate layer by keeping the important features and discarding the unimportant ones with the following formula:

$$FW_i = Sigmoid(AG_{i-1}W + b) \tag{26}$$

$$FD_i = FW_i \otimes AG_{i-1} \tag{27}$$

where AG_{i-1} is the output of the previous aggregate layer, $Sigmoid(\cdot)$ is a nonlinear activation function that maps the values of the input into the interval [0, 1], FW_i stands for the weight of the forgetting gate, and $\otimes$ stands for the dot product operation, FD_i is the output of the forgetting gate.

Input gates are used to control the inflow of information into the current hidden layer, retaining important features and discarding unimportant features with the following formula:

$$IW_i = Sigmoid(H_iW + b) \tag{28}$$

$$ID_i = IW_i \otimes H_i \tag{29}$$

where H_i is the output of the current hidden layer and IW_i is the weight of the input gate, ID_i is the output of the input gate.

The update gate integrates the data that passes through the forgetting gate and the data that passes through the input gate to get the output of the current aggregate layer with the following formula:

$$UW_i = Sigmoid([ID_i : FD_i]W + b) \tag{30}$$

$$AG_i = UW_i \otimes ID_i + (1 - UW_i) \otimes FD_i \tag{31}$$

Where $[ID_i : FD_i]$ represents the splicing operation of ID_i and FD_i along the last dimension, and UW_i is the weight of the update gate, AG_i is the output of the current aggregate layer. Stacking multiple aggregate layers, after considering the output of each hidden layer globally, by updating the gates, allows the model to dynamically learn which traits should be retained for each hidden layer.

The output of the last aggregate layer is passed through the Sigmoid activation function and mapped to the interval [0,1], and the loss measure is performed using the binary cross entropy loss function with the following formula:

$$Loss = -\sum [ylog(p) + (1-y)log(1-p)] \tag{32}$$

where y denotes the label and is 1 if this lncRNA-disease is associated and 0 otherwise, and p represents the probability that the sample is predicted to be a positive case.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05950-z.

Supplementary file 1.

Supplementary file 2.

Acknowledgements

The authors express their gratitude to the editor and anonymous reviewers for their invaluable feedback and suggestions, which significantly contributed to enhancing the quality of our manuscript.

Author Contributions

HW, and BZ designed the methodology. HW, CM and HH developed and implemented the method. HW and ZF performed the experiments and analyzed the results. HW and JQ wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation, China (No. 62172123), the Key Research and Development Program of Heilongjiang (Grant No. 2022ZX01A36), the Special projects for the central government to guide the development of local science and technology, China (No. ZY20B11) and the Harbin Manufacturing Technology Innovation Talent Project (No. CXRC20221104236).

Availability of data and materials

The source code and dataset analyzed in the current study are available at https://github.com/nofou/LDAGM.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no Conflict of interest.

Received: 1 April 2024 Accepted: 1 October 2024
Published online: 15 October 2024

References

1. Graf J, Kretz M. From structure to function: route to understanding lncrna mechanism. BioEssays. 2020;42(12):2000027.
2. Li C, Zhao W, Pan X, Li X, Yan F, Liu S, Feng J, Lu J. LncRNA KTN1-AS1 promotes the progression of non-small cell lung cancer via sponging of miR-130a-5p and activation of PDPK1. Oncogene. 2020;39(39):6157–71.
3. Hong W, Ying H, Lin F, Ding R, Wang W, Zhang M. lncRNA LINC00460 silencing represses EMT in colon cancer through downregulation of ANXA2 via upregulating miR-433-3p. Mol Therapy-Nucl Acids. 2020;19:1209–18.
4. Dai Q, Zhang T, Pan J, Li C. LncRNA UCA1 promotes cisplatin resistance in gastric cancer via recruiting EZH2 and activating PI3K/AKT pathway. J Cancer. 2020;11(13):3882.
5. Xu Y, Shao B. Circulating lncRNA IFNG-AS1 expression correlates with increased disease risk, higher disease severity and elevated inflammation in patients with coronary artery disease. J Clin Labor Anal. 2018;32(7):22452.
6. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNA disease: a database for long-non-coding RNA-associated diseases. Nucl Acids Res. 2012;41(D1):983–6.
7. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M, Wang L, et al. Lnc2cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucl Acids Res. 2016;44(D1):980–5.

8. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R. Noncode: an integrated knowledge database of non-coding RNAs. Nucl Acids Res. 2005;33:112–5.
9. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. Omim. org: online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. Nucl Acids Res. 2015;43:789–98.
10. Lin L, Chen R, Zhu Y, Jing H, Chen L. SCCPMD: probability matrix decomposition method subject to corrected similarity constraints for inferring long non-coding RNA-disease associations. Front Microbiol. 2023;13:1093615.
11. Zhu Q, Fan Y, Pan X. Fusing multiple biological networks to effectively predict miRNA-disease associations. Curr Bioinfo. 2021;16(3):371–84.
12. Yao Y, Ji B, Lv Y, Li L, Xiang J, Liao B, Gao W. Predicting lncRNA-disease association by a random walk with restart on multiplex and heterogeneous networks. Front Genet. 2021;12:712170.
13. Bai Z, Lu J, Chen A, Zheng X, Wu M, Tan Z, Xie J. Identification and validation of cuproptosis-related lncRNA signatures in the prognosis and immunotherapy of clear cell renal cell carcinoma using machine learning. Biomolecules. 2022;12(12):1890.
14. Zeng M, Lu C, Fei Z, Wu F-X, Li Y, Wang J, Li M. DMFLDA: a deep learning framework for predicting lncRNA-disease associations. IEEE/ACM Trans Comput Biol Bioinfo. 2020;18(6):2353–63.
15. Sheng N, Huang L, Lu Y, Wang H, Yang L, Gao L, Xie X, Fu Y, Wang Y. Data resources and computational methods for lncRNA-disease association prediction. Comput Biol Med. 2023;153:106527.
16. 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):3357–64.
17. Fu G, Wang J, Domeniconi C, Yu G. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. Bioinformatics. 2018;34(9):1529–37.
18. Xuan Z, Li J, Yu J, Feng X, Zhao B, Wang L. A probabilistic matrix factorization method for identifying lncRNA-disease associations. Genes. 2019;10(2):126.
19. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, Wang B. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. PloS one. 2014;9(1):87797.
20. Li G, Luo J, Liang C, Xiao Q, Ding P, Zhang Y. Prediction of lncRNA-disease associations based on network consistency projection. IEEE Access. 2019;7:58849–56.
21. 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. Brief Bioinfo. 2023;24(5):276.
22. Chen X, You Z-H, Yan G-Y, Gong D-W. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget. 2016;7(36):57919.
23. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Molecular BioSyst. 2014;10(8):2074–81.
24. Yu G, Fu G, Lu C, Ren Y, Wang J. BRWLDA: bi-random walks for predicting lncRNA-disease associations. Oncotarget. 2017;8(36):60429.
25. Sheng N, Cui H, Zhang T, Xuan P. Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA-disease association prediction. Brief Bioinfo. 2021;22(3):067.
26. Wang W, Guan X, Khan MT, Xiong Y, Wei D-Q. LMI-DForest: a deep forest model towards the prediction of lncRNA-miRNA interactions. Comput Biol Chem. 2020;89:107406.
27. Yuan L, Zhao J, Sun T, Shen Z. A machine learning framework that integrates multi-omics data predicts cancer-related lncRNAs. BMC Bioinfo. 2021;22(1):332.
28. Lan W, Lai D, Chen Q, Wu X, Chen B, Liu J, Wang J, Chen Y-PP. LDICDL: LncRNA-disease association identification based on collaborative deep learning. IEEE/ACM Trans Comput Biol Bioinfo. 2020;19(3):1715–23.
29. Shi Z, Zhang H, Jin C, Quan X, Yin Y. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. BMC Bioinfo. 2021;22:1–20.
30. Xuan P, Pan S, Zhang T, Liu Y, Sun H. Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. Cells. 2019;8(9):1012.
31. Lu C, Xie M. lncRNA: lncrna-disease associations prediction with deep autoencoder and xgboost classifier. Interdiscipl Sci Comput Life Sci. 2023;15(3):439–51.
32. Sheng N, Huang L, Wang Y, Zhao J, Xuan P, Gao L, Cao Y. Multi-channel graph attention autoencoders for disease-related lncRNAs prediction. Brief Bioinfo. 2022;23(2):604.
33. 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.
34. Wu Q-W, Xia J-F, Ni J-C, Zheng C-H. GAERF: predicting lncRNA-disease associations by graph auto-encoder and random forest. Brief Bioinfo. 2021;22(5):391.
35. Liang Y, Zhang Z-Q, Liu N-N, Wu Y-N, Gu C-L, Wang Y-L. MAGNSE: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. BMC Bioinfo. 2022;23(1):189.
36. Lin X, Lu Y, Zhang C, Cui Q, Tang Y-D, Ji X, Cui C. LncRNA disease v3.0: an updated database of long non-coding RNA-associated diseases. Nucl Acids Res. 2024;52(D1):1365–9.
37. Gao Y, Shang S, Guo S, Li X, Zhou H, Liu H, Sun Y, Wang J, Wang P, Zhi H, et al. Lnc2Cancer 30: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. Nucleic acids research. 2021;49(D1):1251–8.
38. Huang T, Wang G, Yang L, Peng B, Wen Y, Ding G, Wang Z. Transcription factor YY1 modulates lung cancer progression by activating lncRNA-PVT1. DNA Cell Biol. 2017;36(11):947–58.
39. Zeng Z, Bo H, Gong Z, Lian Y, Li X, Li X, Zhang W, Deng H, Zhou M, Peng S, et al. AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis. Tumor Biol. 2016;37:729–37.
40. Tang T, Cheng Y, She Q, Jiang Y, Chen Y, Yang W, Li Y. Long non-coding RNA tug1 sponges MIR-197 to enhance cisplatin sensitivity in triple negative breast cancer. Biomed Pharmacother. 2018;107:338–46.
41. Li Z, Yu D, Li H, Lv Y, Li S. Long non-coding RNA UCA1 confers tamoxifen resistance in breast cancer endocrinotherapy through regulation of the EZH2/p21 axis and the PI3K/AKT signaling pathway. Int J Oncol. 2019;54(3):1033–42.

42. Zhou Y, Yang H, Xia W, Cui L, Xu R, Lu H, Xue D, Tian Z, Ding T, Cao Y, et al. LncRNA MEG3 inhibits the progression of prostate cancer by facilitating H3K27 trimethylation of EN2 through binding to EZH2. J Biochem. 2020;167(3):295–301.
43. Tang L, Shen H, Li X, Li Z, Liu Z, Xu J, Ma S, Zhao X, Bai X, Li M, et al. MIR-125a-5p decreases after long non-coding RNA HOTAIR knockdown to promote cancer cell apoptosis by releasing caspase 2. Cell Death Dis. 2016;7(3):2137–2137.
44. Lu, Z., Bretonnel Cohen, K., Hunter, L.: Generif quality assurance as summary revision. In: Biocomputing, 2007; pp. 269–280. World Scientific
45. 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. Nucl Acids Res. 2014;42(D1):92–7.
46. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucl Acids Res. 2014;42(D1):1070–4.
47. Zhou Y, Wang X, Yao L, Zhu M. LDAformer: predicting lncRNA-disease associations based on topological feature extraction and transformer encoder. Brief Bioinfo. 2022;23(6):370.
48. 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. Nucl Acids Res. 2019;47:1013–7.
49. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, et al. Human disease ontology 2018 update: classification, content and workflow expansion. Nucl Acids Res. 2019;47(D1):955–62.
50. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of go terms. Bioinformatics. 2007;23(10):1274–81.
51. 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):1644–50.
52. Chen X, Yan CC, Zhang X, You Z-H, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: within and between score for miRNA-disease association prediction. Sci Rep. 2016;6(1):21106.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.