OXFORD

Briefings in Bioinformatics, 2022, 1–12
https://doi.org/10.1093/bib/bbac370
Articles

LDAformer: predicting lncRNA-disease associations based on topological feature extraction and Transformer encoder

Yi Zhou, Xinyi Wang, Lin Yao and Min Zhu

Corresponding author. Min Zhu, College of Computer Science, Sichuan University, 1st Ring Road South 1 Section, 610065, Chengdu, China. Tel: 86-13980020500; Fax: 86-028-85469560; E-mail: zhumin@scu.edu.cn

Abstract

The identification of long noncoding RNA (lncRNA)-disease associations is of great value for disease diagnosis and treatment, and it is now commonly used to predict potential lncRNA-disease associations with computational methods. However, the existing methods do not sufficiently extract key features during data processing, and the learning model parts are either less powerful or overly complex. Therefore, there is still potential to achieve better predictive performance by improving these two aspects. In this work, we propose a novel lncRNA-disease association prediction method LDAformer based on topological feature extraction and Transformer encoder. We construct the heterogeneous network by integrating the associations between lncRNAs, diseases and micro RNAs (miRNAs). Intra-class similarities and inter-class associations are presented as the lncRNA-disease-miRNA weighted adjacency matrix to unify semantics. Next, we design a topological feature extraction process to further obtain multi-hop topological pathway features latent in the adjacency matrix. Finally, to capture the interdependencies between heterogeneous pathways, a Transformer encoder based on the global self-attention mechanism is employed to predict lncRNA-disease associations. The efficient feature extraction and the intuitive and powerful learning model lead to ideal performance. The results of computational experiments on two datasets show that our method outperforms the state-of-the-art baseline methods. Additionally, case studies further indicate its capability to discover new associations accurately.

Keywords: lncRNA-disease association, topological feature extraction, Transformer, global self-attention mechanism

Introduction

Long noncoding RNA (lncRNA) is a type of noncoding RNA with a length of more than 200 nucleotides [1]. It plays an important molecular regulatory role in biological life activities and is closely related to the occurrence and development of many diseases [2]. For example, AB007962 is downregulated in gastric cancer and associated with a poor prognosis [3]. AATBC regulates Pinin to promote metastasis in nasopharyngeal carcinoma [4]. 91H is associated with poor development in colorectal cancer by modifying HNRNPK expression [5].

So far, lots of lncRNA-disease associations (LDAs) have been validated through biological experiments, but due to the high cost of time and resources, the further advancement of which could be greatly limited. Based on known experimental data, an increasing number of computational methods are proposed to predict LDAs to address such shortcomings. Here, we summarize the existing computational methods into two categories: one is traditional machine learning-based, and the other is deep learning-based.

Chen *et al.* [6] proposed an essential assumption that similar diseases tend to be associated with functionally similar lncRNAs. Thus, the association and similarity information is commonly integrated and utilized. Traditional machine learning-based methods apply matrix operations, network propagation algorithms and classifiers. MFLDA [7] decomposed data matrices into low-rank matrices via matrix tri-factorization, optimized data integration weights and low-rank matrices, then reconstructed the LDA matrix as predictions. SIMCLDA [8] extracted lncRNA and disease features by similarity computation and principal component analysis, and then completed the association matrix based on the inductive matrix completion framework.

Network propagation algorithms can capture potential topology features, of which random walk is the most commonly used. RWRHLD [9] integrated an lncRNA-disease heterogeneous network, and then implemented the random walk with restart (RWR) algorithm to prioritize candidate LDAs. IRWRLDA [10] incorporated lncRNA expression similarity and disease semantic similarity to set the initial probability vector of RWR, thereby improving the LDA prioritization performance on the lncRNA-disease network. LDA-LNSUBRW [11] obtained the interaction possibility of unknown LDAs by pretreatment and predicted the potential associations based on linear neighborhood similarity and unbalanced bi-random walk. As a topological representation of the network, the adjacency matrix is critical. Ping *et al.* [12] constructed an lncRNA-disease bipartite network and calculated the dot product of the adjacency and similarity matrices as predictive values, which also inspired our topological feature extraction process.

Yi Zhou is a postgraduate student in the College of Computer Science, Sichuan University. Her research interests include deep learning, graph data analysis and bioinformatics.
Xinyi Wang is a postgraduate student in the College of Computer Science, Sichuan University. His research interests include deep learning and bioinformatics.
Lin Yao is a postgraduate student in the College of Computer Science, Sichuan University. His research interests include machine learning and bioinformatics.
Min Zhu is a professor in the College of Computer Science, Sichuan University. Her research interests include bioinformatics, visual analysis and image processing.
Received: May 25, 2022. **Revised:** July 27, 2022. **Accepted:** August 6, 2022

With the low-dimensional representations of lncRNAs and diseases, researchers utilize machine learning classifiers to determine if the given lncRNA–disease pairs are associated. LRLSLDA [6] developed a semi-supervised LDA prediction framework based on Laplacian regularized least squares. LDAP [13] fused similarity matrices of lncRNAs and diseases and predicted potential LDAs by bagging support vector machine. Among the traditional machine learning classifiers applied to LDA prediction, random forest (RF) performs well and is adopted by many methods such as DisLncRF [14], RFLDA [15] and IPCARF [16].

Deep learning has been widely applied in bioinformatics [17]. Compared with traditional machine learning algorithms, deep learning algorithms have stronger nonlinear fitting capabilities and are more flexible in applications, allowing for end-to-end low-dimensional representation extraction and classification prediction. Here, multilayer perceptron (MLP) and convolutional neural network (CNN) have shown promising performance. DMFLDA [18] designed a cascade of hidden layers to learn latent representations of lncRNAs and diseases, thus capturing the complex nonlinear LDA relationships. SDLDA [19] integrated features extracted through singular value decomposition and MLP, followed by LDA predictor of another perceptron. Xuan *et al.* [20–22] utilized the association and similarity information among lncRNAs, diseases and miRNAs, and further exploited the topological structures formed by them; such an idea of feature integration is now widely used in LDA predictions. They also proposed three CNN-based models: CNNLDA [20], CNNDLP [21] and LDAPred [22]. CNNLDA constructed a dual CNN-based model for learning the global and attention representations, CNNDLP made a combination of attention mechanism and CNN autoencoder and LDAPred designed a dual CNN predictive model for original and topological features.

Among the deep learning-based methods, graph neural network (GNN) algorithms take into account both nonlinear fitting and local topology learning to a certain extent. Due to the graph nature of association studies, in recent years, with the development of GNNs, they have been increasingly adopted in LDA predictions. GAMCLDA [23] utilized graph convolutional network (GCN) to encode node features and local graph structure, then reconstructed the LDA matrix by the inner product. VGAELDA [24] integrated variational inference and graph autoencoders for LDA prediction, where variational graph autoencoders (VGAE) inferred representations from features and graph autoencoders propagated labels via known LDAs. HGATLDA [25] incorporated node feature structural graphs and the lncRNA-disease topological structural graph, then developed a novel heterogeneous graph attention (HGAT) network framework based on meta-paths.

Moreover, several methods combined various traditional machine learning and deep learning algorithms. GCNLDA [26] processed node features with an attention mechanism, followed by prediction with CNN and GCN branches. GAERF [27] learned node embeddings via graph autoencoder and performed binary classification by RF. GCRFLDA [28] constructed an encoder consisting of GCN, MLP and conditional random field with an attention mechanism. VADLP [29] combined random walk, convolutional autoencoder, variance autoencoder and MLP. MGATE [30] utilized three graph autoencoders to learn multiple graph representations and merged them using an attention mechanism, followed by RF for prediction. GTAN [31] encoded the neighbor topologies with multiple graph attention neural networks and encoded the node attributes with attention mechanism, CNN and MLP. While improving prediction performances, the frameworks of the latest LDA prediction methods have become increasingly complex.

Although the methods above have achieved good performance, there is still space for improvement. On the one hand, it is inefficient to utilize the raw similarity and association information, and the extensive model computation without extracting key information in advance can blur the node embedding semantics and thus weaken the classification performance. On the other hand, the relatively limited performance gained from the complex model structures suggests that they do not fit well enough with the underlying assumptions of the LDA prediction.

To address such shortcomings, we propose an intuitive LDA prediction method LDAformer based on topological feature extraction and Transformer encoder [32]. Our main contributions are summarized as follows:

- We treat similarity and association information as the connectivity between nodes uniformly and construct the lncRNA-disease-miRNA weighted adjacency matrix.
- We design a multi-hop topological feature extraction process, obtaining the input features with explicit pathway semantics.
- We adopt the Transformer encoder as the LDA predictor; the powerful global self-attention mechanism could capture the interdependencies hidden in topological pathways.
- Experimental results indicate that LDAformer outperforms the state-of-the-art methods. Case studies further illustrate its promising predictive capability.

Materials

Datasets

In this study, to demonstrate the effectiveness of LDAformer, we evaluate it on two datasets:

- **Dataset1** from Fu's work [7] is widely referenced as a reliable benchmark dataset. It contains 240 lncRNAs, 412 diseases, 495 miRNAs and 2697 LDAs from Lnc2Cancer [33], LncRNADisease [34], GeneRIF [35], 1002 lncRNA-miRNA associations from starBase v2.0 [36], 13 562 miRNA-disease associations (MDAs) from HMDD v2.0 [37].
- **Dataset2** is integrated by ourselves, the data and detailed processing code are available at https://github.com/EchoChou990919/LDAformer. It contains 665 lncRNAs, 316 diseases, 295 miRNAs and 3833 LDAs from Lnc2Cancer v3.0 [38], LncRNADisease v2.0 [39], 2108 lncRNA-miRNA associations from starBase v2.0, 8540 MDAs from HMDD v3.2 [40]. Disease semantic information is obtained from the Disease Ontology[1] [41]. MeSH[2] and miRBase [42] assist in determining node names.

With the version update of original public databases, the experimental LDAs in dataset2 are more comprehensive. Therefore, it is more valuable to predict LDAs on dataset2 and at the same time more challenging due to the sparsity of associations.

Disease semantic similarity

Computational predictions of LDAs are generally based on the assumption that functionally similar lncRNAs tend to be associated with phenotypically similar diseases. Therefore, we obtain the disease semantic information from Disease Ontology, which represents the parent–child relationship between diseases in the data structure of a directed acyclic graph.

[1] https://disease-ontology.org/
[2] https://www.ncbi.nlm.nih.gov/mesh/

Then we calculate disease similarities using Wang's method [43]. For disease t_1, assuming $T1$ is the set involving t_1 and all of its ancestor terms, the semantic contribution values of all terms in $T1$ to $t1$ can be measured as follows:

$$S_{t_1}(t) = \begin{cases} 1, & t = t_1 \\ \max\limits_{t' \in children\ of\ t} \left(\frac{S_{t_1}(t')}{2}\right), & t \in T_1\ and\ t \neq t_1 \end{cases}. \quad (1)$$

For any other disease t_2, the semantic similarity between t_1 and t_2 is defined as

$$Sim(t_1, t_2) = \frac{\sum_{t \in T_1 \cap T_2} (S_{t_1}(t) + S_{t_2}(t))}{SV(t_1) + SV(t_2)}, \quad (2)$$

where $SV(t_1) = \sum_{t \in T_1} S_{t_1}(t)$ presents the summation of all the semantic contribution values of $T1$, and ditto for $SV(t_2)$. Semantic similarity describes the overlap ratio of their ancestor terms, with the closer the ancestor greater the weight.

LncRNA/miRNA functional similarity

Subsequently, lncRNA/miRNA functional similarities are measured according to Wang *et al.* [44] based on the calculated disease similarities and known LDAs.

For lncRNA/miRNA r_1 and r_2, assuming that there are n_1 diseases associated with r_1 and n_2 diseases associated with r_2, these diseases can be denoted as $t_{1i}, (1 \le i \le n_1)$ and $t_{2j}, (1 \le j \le n_2)$, thus the functional similarity between r_1 and r_2 is given by

$$Sim(r_1, r_2) = \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \max_{1 \le j \le n_2} (Sim(t_{1i}, t_{2j})) + \sum_{j=1}^{n_2} \max_{1 \le i \le n_1} (Sim(t_{2j}, t_{1i})) \right]. \quad (3)$$

LncRNA-disease-miRNA weighted adjacency matrix

For a dataset with l lncRNAs, m miRNAs and d diseases, the inter-class association matrices are defined as $\begin{cases} A_{LD} \in R^{l \times d} \\ A_{LM} \in R^{l \times m} \\ A_{DM} \in R^{d \times m} \end{cases}$, where the value at the corresponding position is 1 if there is an experimental association, 0 otherwise. And the intra-class similarity matrices are defined as $\begin{cases} S_L \in R^{l \times l} \\ S_D \in R^{d \times d} \\ S_M \in R^{m \times m} \end{cases}$, where the corresponding values are the calculated functional or semantic similarities. In particular, the self-similarities on the diagonal are all set to 0.

Viewing the dataset as a heterogeneous network, both the associations and similarities can be treated as the connectivities between nodes. So the above association and similarity matrices can be concatenated into a complete weighted adjacency matrix:

$$A = \begin{bmatrix} S_L & A_{LD} & A_{LM} \\ A_{LD}^T & S_D & A_{DM} \\ A_{LM}^T & A_{DM}^T & S_M \end{bmatrix}, \quad (4)$$

where A_{LD}^T, A_{LM}^T and A_{DM}^T denote the transpose of A_{LD}, A_{LM} and A_{DM}.

LDAformer

Based on the lncRNA-disease-miRNA weighted adjacency matrix, we propose an end-to-end LDA prediction method LDAformer, and the flowchart is shown in Figure 1. LDAformer consists of three parts: the topological feature extraction process, the Transformer encoder and the prediction layer.

Topological feature extraction process

According to the definition of matrix multiplication, the power of the adjacency matrix represents multi-hop connectivities between network nodes [45]. Thus, we propose a simple process to extract the multi-hop topology information:

$$A^{n_h} = \begin{cases} A, & n_h = 1 \\ norm(A^{n_h - 1}A), & n_h > 1 \end{cases}, \quad (5)$$

$$norm(A) = \frac{A_{diag0}}{max(A_{diag0})}, \quad (6)$$

where A_{diag0} denotes matrix A with the diagonal set to 0. For any two nodes, the value at the corresponding position of A^{n_h} is the normalized sum of the products of the n-hop weights between the nodes. For the i-th node lncRNA n_i and the j-th node disease n_j, the topological feature X is obtained as

$$X = \left[A_{*,i}; A_{*,j}; A_{*,i}^2; A_{*,j}^2; \cdots; A_{*,i}^{n_h}; A_{*,j}^{n_h} \right] W_{in}, \quad (7)$$

where $A_{*,i}^{n_h}$ and $A_{*,j}^{n_h}$ denote the i-th and j-th columns of A^{n_h}, and $W_{in} \in R^{2n_h \times d_{model}}$ is a learnable matrix to make a linear transformation. Here, $X \in R^{(l+d+m) \times d_{model}}$ has an explicit implication, in which the k-th row $X_{k,*} = \left[A_{k,i}; A_{k,j}; A_{k,i}^2; A_{k,j}^2; \cdots; A_{k,i}^{n_h}; A_{k,j}^{n_h} \right] W_{in} (1 \le k \le l + d + m)$ indicates the linearly transformed $2, 3, 4, \cdots, 2n_h$ hop pathways between lncRNA n_i and disease n_j mediated by the node n_k.

Transformer encoder

As a powerful deep learning model based on global self-attention, Transformer has shown excellent performance in understanding the long-range interdependencies of data. In LDAformer, we employ only the encoder part. In contrast to the graph attention neural network which learns neighborhood topology features, the Transformer encoder computes scaled dot product attention at the global scale, capturing the interdependency between any two topological pathways. And other common self-attention mechanisms are often used just for merging features and bridging algorithm modules, while the Transformer encoder is the only core learning module in our method.

Encoder stacks

The encoder of Transformer is a stack of N identical layers, where N is set to 6 by default, and each contains two sub-layers: multi-head attention and feedforward network. For each sub-layer, a residual connection is employed, followed by a layer normalization. Such a process can be described mathematically as

$$H(X) = layerNorm(X + subLayer(X)), \quad (8)$$

where $subLayer(X)$ is the calculation implemented by multi-head attention or feedforward network, then $H(X)$ is the output of each sub-layer.

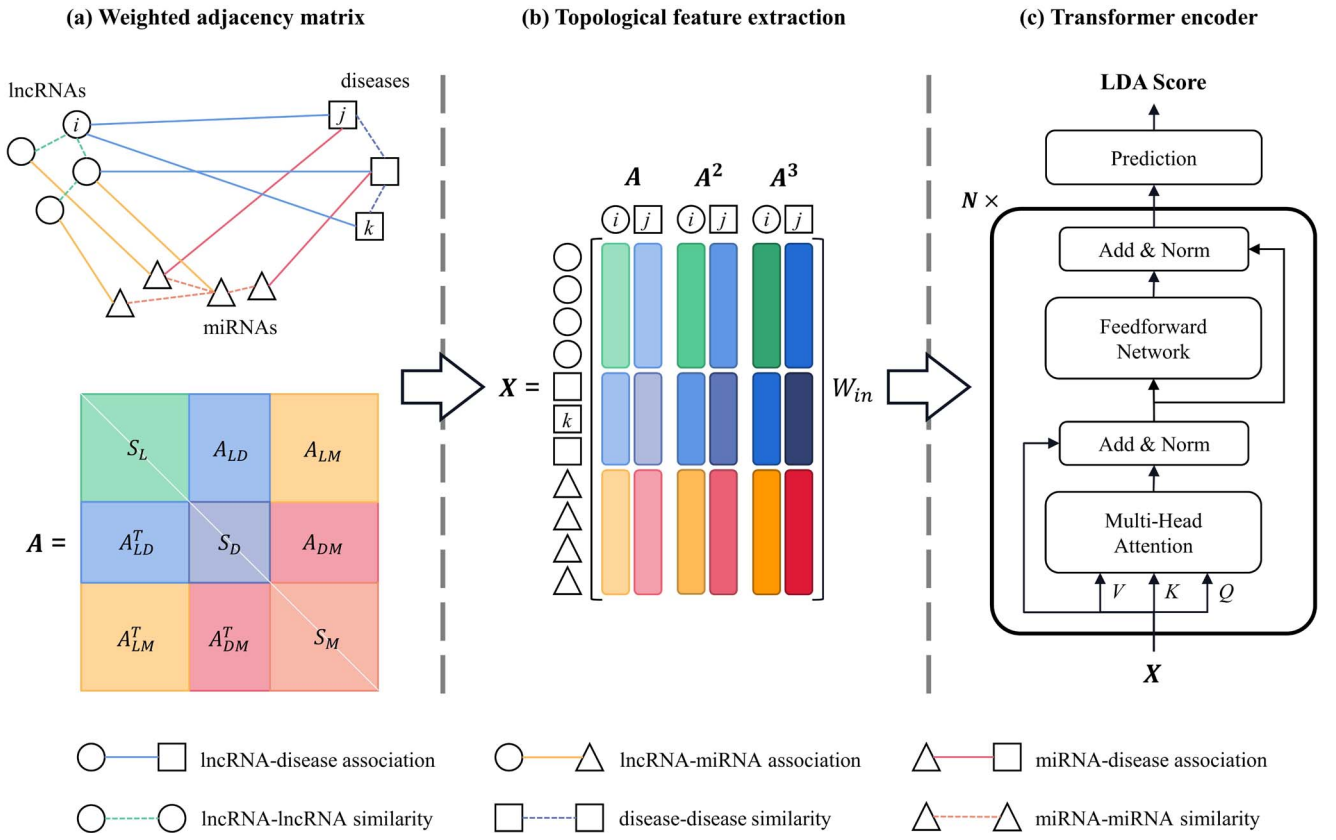

Figure 1. The flowchart of LDAformer. **(A)** The weighted adjacency matrix of lncRNA-disease-miRNA heterogeneous network. **(B)** The topological feature extraction process. It obtains model input features containing $2, 3, 4, \cdots, 2n_h$-hop pathways between lncRNA n_i and disease n_j. **(C)** The Transformer encoder. It captures the interdependencies between topological pathways and then outputs the predicted LDA score through a prediction layer.

Multi-head attention

Figure 2 shows the calculation process of multi-head attention. Scaled dot product attention has three inputs: Q (query), K (key), V (value), and d_K is the channel dimension of K. It is defined as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V, \tag{9}$$

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)}. \tag{10}$$

Multi-head attention splits the scaled dot product attention into n_{head} heads, where n_{head} is set to 2 by default. For each head, firstly three parallel linear transformations convert input X to Q, K and V:

$$Q_i = XW_i^Q, \tag{11}$$

$$K_i = XW_i^K, \tag{12}$$

$$V_i = XW_i^V, \tag{13}$$

where $1 \le i \le n_{head}$, $W_i^Q, W_i^K, W_i^V \in R^{d_{model} \times (d_{model}/n_{head})}$ are learnable parameters. Then the outputs of each head are aggregated together:

$$MHA(X) = concat(head_1, \cdots, head_{n_{head}})W^O, \tag{14}$$

$$head_i = Attention(Q_i, K_i, V_i), \tag{15}$$

where $W^O \in R^{(d_{model}/n_{head}) * n_{head} \times d_{model}}$ makes another linear transformation, and $MHA(X)$ is the overall output of multi-head attention.

Feedforward network

Feedforward network is a two-layer neural network separated by a ReLU activation:

$$FFN(X) = max(0, XW_1 + b_1)W_2 + b_2, \tag{16}$$

where $W_1 \in R^{d_{model} \times (d_{ff} * d_{model})}$, $W_2 \in R^{(d_{ff} * d_{model}) \times d_{model}}$, and d_{ff} is set to 2 by default.

Prediction layer

Prediction layer performs a simple aggregation on the flattened output of the self-attention layer, followed by a sigmoid activation to obtain the predicted LDA score p:

$$p = sigmoid(flatten(X_{encoded})W_{out} + b_{out}), \tag{17}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}, \tag{18}$$

where $W_{out} \in R^{(l+d+m) * d_{model} \times 1}$. Finally, binary cross-entropy is used as the loss function for optimizing LDAformer, which is defined as

$$Loss = -\sum[ylog(p) + (1 - y)log(1 - p)], \tag{19}$$

where y denotes the ground-truth label, $y = 1$ if there are experimental association records between lncRNA n_i and disease n_j included in the dataset, otherwise $y = 0$.

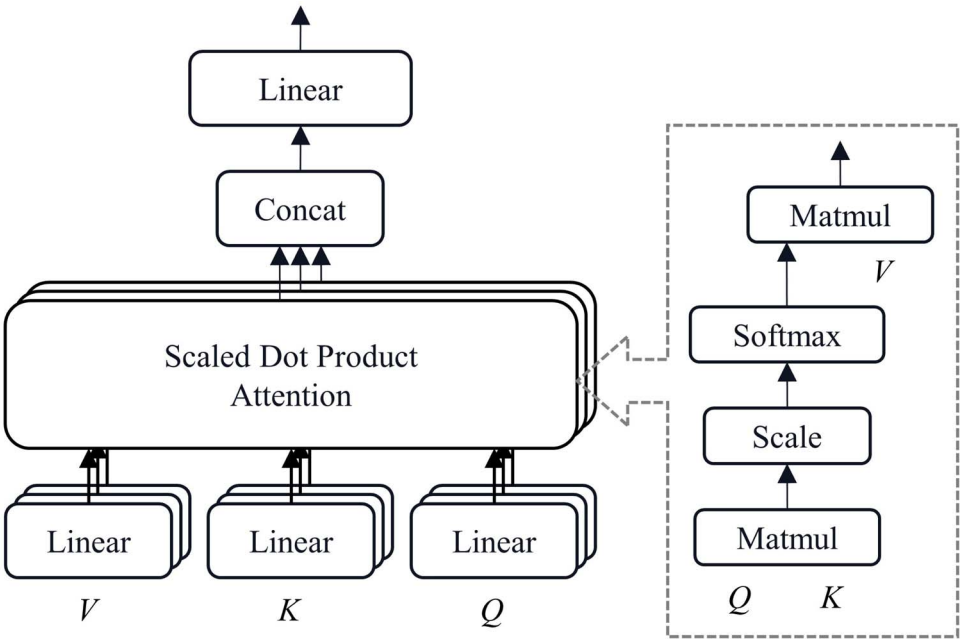

Figure 2. The calculation process of multi-head attention [32].

Experiments

Experimental environment and evaluation metrics

LDAformer is implemented in Python 3.8.8 and Pytorch 1.10.0; all experiments are conducted on an NVIDIA RTX 3090 GPU with 24GB memory. The codes are available at https://github.com/EchoChou990919/LDAformer, including the model itself, training and testing parts, default parameter settings and a demo of trained models. For dataset 1, training takes about 120 s, and testing takes about 30 s (0.35 ms/per lncRNA–disease pair). For dataset 2, training takes about 70 s, and testing takes about 85 s (0.423 ms/per lncRNA–disease pair).

In the experiments, 5-fold cross-validation is used to evaluate the performance of LDAformer. For each dataset, we take the known LDAs as positive samples and the unknown LDAs as negative samples. All positive samples are split into five subsets, four of which and equal-sized randomly selected negative samples used for training, and the rest one and all the remaining negative samples are used for testing. At the same time, the weighted adjacent matrix A is reconstructed: the corresponding values of the test positive samples in A_{LD} and A_{LD}^{T} are set to 0, and lncRNA functional similarities in S_L are also recalculated.

Two classic metrics are adopt for evaluation:

- **AUC** is the area under the Receiver Operating Characteristic (ROC) curve, which is typically used in binary classification to study the output of a classifier.
- **AUPR** is the area under the Precision-Recall (PR) curve, which is a useful measure of prediction when the classes are very imbalanced.

Effect of topological feature extraction

There are two significant operations in the topological feature extraction process of LDAformer: (1) For the lncRNA–disease pair to be predicted, the concatenation of the corresponding columns in $A, A^2, \cdots, A^{n_h}$. (2) A linear transformation that converts the dimensionality of the topological feature to d_{model}.

Setting other hyperparameters to default values, for n_h and d_{model}, we perform a grid search to analyze the effectiveness of the feature extraction processes. n_h is selected from {1, 2, 3, 4}, In particular, when $n_h = 1$, (1) does not work. And d_{model} is set from {4, 6, 8, 10, 12, 14, 16}, moreover, a *False* state is added to eliminate the effect of (2).

As shown in Figure 3, the two operations are effective individually and work best when properly combined. Operation (1) introduces the valuable multi-hop topological information, while operation (2) adjusts the dimensionality of the input feature to make it appropriate for the calculation of the Transformer encoder. More experimental results on both datasets are given in Supplementary Data 1. From the experimental results, there are several comparable settings. Considering the performance of our method on both metrics, n_h and d_{model} are set to 3 and 12 for dataset 1, and 3 and 6 for dataset 2.

Parameter analysis

In this section, we estimate the influence of three important hyperparameters with grid research: the number of attention heads n_{head} changed from {1, 2, 3, 4}, the number of encoder layers n_l selected from {2, 4, 6, 8} and the size ratio of hidden units in the feedforward network d_{ff} chosen from {0.5, 1, 2, 4}. There are multiple sets of hyperparameters that can achieve optimal results. Ultimately, n_h, n_l and d_{ff} are set to 1, 4 and 0.5 for dataset 1, 3, 2, and 1 for dataset 2.

As Figure 4 shows, we plot analysis histograms with the X-axis showing one unfixed hyperparameter and the Y-axis showing the AUC and AUPR values. It's apparent that our method is parameter insensitive, the evaluation metrics do not fluctuate significantly due to changes in a single parameter. For the full results of the grid search see Supplementary Data 2. Overall, for dataset 1, LDAformer with fewer hidden units, fewer heads and more layers performs better. For dataset 2, LDAformer suits with fewer hidden units, more heads and fewer layers. The situation varies in the two datasets, probably due to the differences in LDA sparsity. In the case of more sparse valid information, multiple heads jointly attend to information from different representation

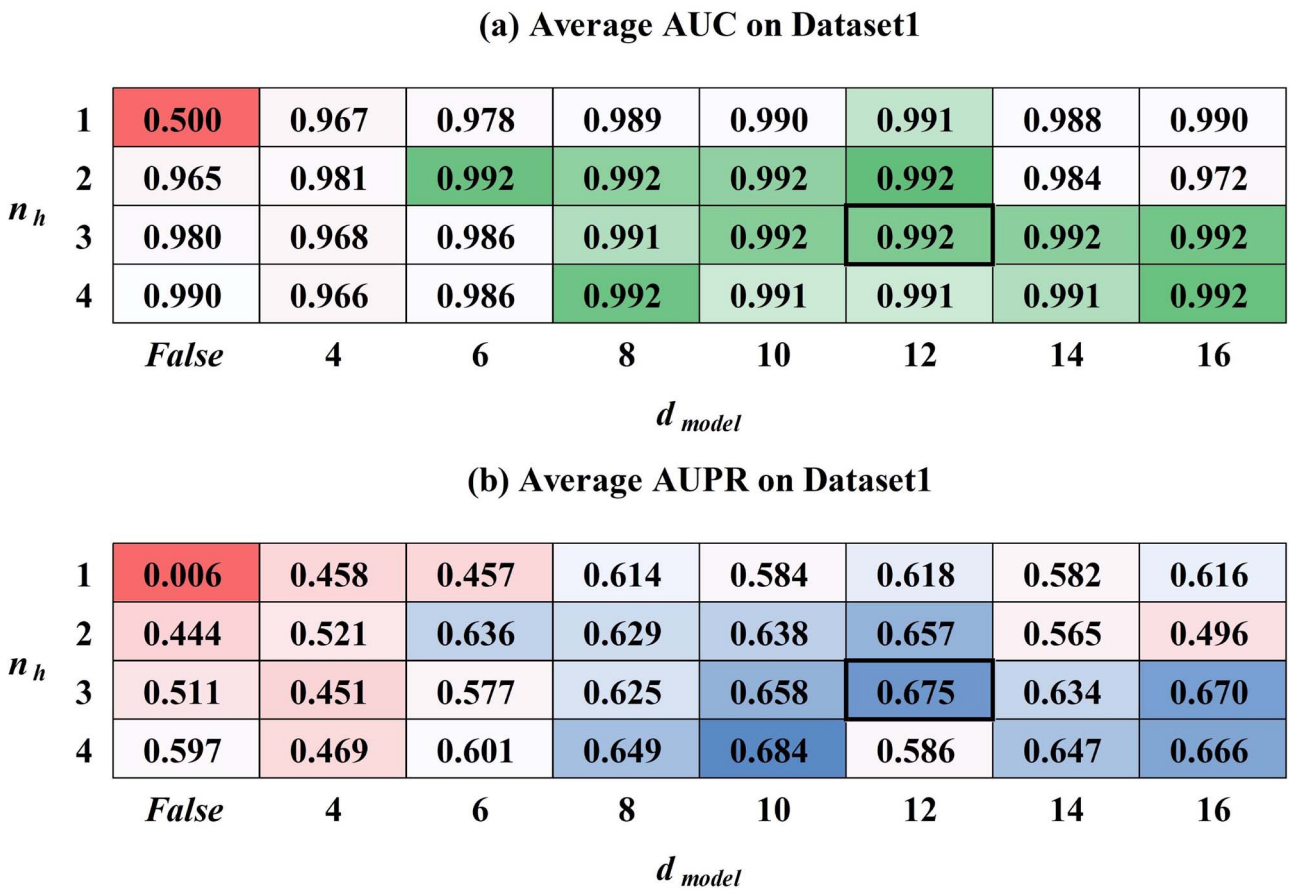

Figure 3. The effectiveness of our topological feature extraction process on dataset1. The n_h on the vertical axis indicates that the extracted features contain up to $2n_h$ hops of topological pathway information. The d_{model} on the horizontal axis represents the linear transformation converting the feature dimension from $2n_h$ to d_{model}. **(A)** Average AUC value with red–green color mapping, the higher the greener. **(B)** Average AUPR value with red–blue color mapping, the higher the bluer.

subspaces, enhancing the expressiveness of the model. And fewer layers may be able to moderate the overfitting problem caused by undersampling training.

Comparison with baseline methods

To demonstrate the superiority of LDAformer, we compared it with the following methods:

- **SIMCLDA** (2018) [8] is a classic LDA prediction method based on inductive matrix completion.
- **SDLDA** (2020) [19] is a hybrid computational framework, which combines linear and nonlinear features extracted by singular value decomposition and MLP.
- **DMFLDA** (2020) [18] is a deep matrix factorization model that learns nonlinear features through MLP.
- **GAMCLDA** (2020) [23] is a computational framework based on graph autoencoder matrix completion. GCN is utilized as the encoder, and the inner product of embedding is used as the decoder to reconstruct the association matrix.
- **VGAELDA** (2021) [24] is an end-to-end model based on variational inference and graph autoencoder, where variational graph autoencoders infer node representations, while graph autoencoders propagate labels.
- **LR-GNN** (2022) [46] is a GNN based on link representation to identify potential molecular associations. Here we adopt it for LDA prediction.

These baselines are based on traditional machine learning and deep learning algorithms. In particular, GAMCLDA, VGAELDA and LR-GNN utilized GNNs. We run these methods with their default parameters, and average ROC and PR curves of 5-fold cross-validation are shown in Figure 5. Additionally, the significance of the differences between our method and these baseline methods in AUC and AUPR values is verified using the two-tailed paired-sample t-test on both datasets. As shown in Supplementary Data 3, LDAformer performs significantly better than these methods with P-values less than 0.05.

As a widely used benchmark dataset, more comparisons can be performed on dataset 1. We further consider seven state-of-the-art methods: GCNLDA [26], CNNDLP [21], VADLP [29], GCRFLDA [28], GAERF [27], GTAN [31] and MGATE [30]. They have shown promising performance with model structures combining GNN and other deep learning methods. See Table 1, LDAformer achieves the highest AUC of 0.994, 1.12% higher than the second-best GATN. And it performs superior well on AUPR, compared with other top methods with the level of 0.4–0.5, LDAformer achieves the highest 0.709, 44.40% higher than the second-highest GAERF. In the dataset with a positive-to-negative ratio lower than 1:35, such results indicate a very high quality of the predictions.

Still seen in Figure 5, for dataset 2, LDAformer consistently outperforms the baselines. It achieves the best AUC of 0.941, which is 1.18% higher than the second-best LR-GNN. And it is still

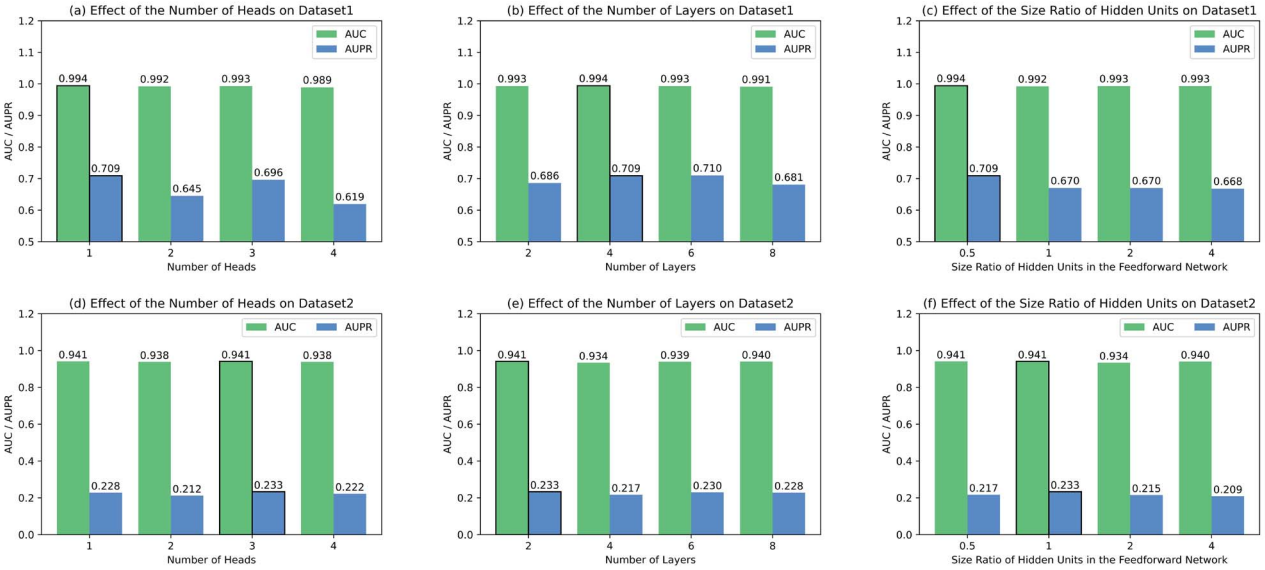

Figure 4. Parameter analysis results. On both datasets, comparisons of AUC and AUPR values on the number of heads h_{head}, the number of layers n_l and the size ratio of hidden units in the feedforward network d_{ff}. The selected ones are framed by black lines.

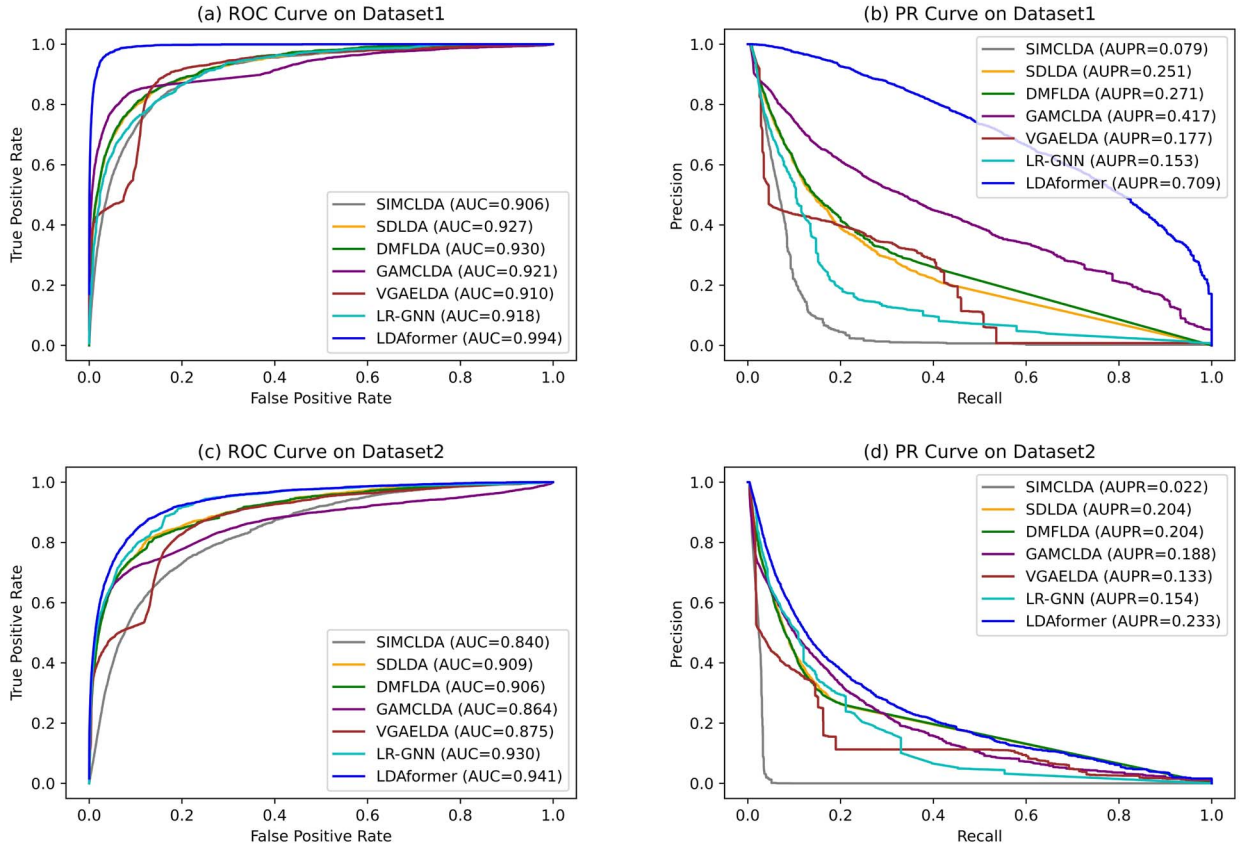

Figure 5. Performance comparison between LDAformer and other baseline methods. **(A)–(B)** Comparisons of ROC curve, AUC value, PR curve and AUPR value on dataset 1. **(C)–(D)** Comparisons of ROC curve, AUC value, PR curve and AUPR value on dataset 2.

outstanding on AUPR, achieving the highest 0.233, which is 9.31% better than the second-best DMFLDA or SDLDA. On further analysis, the big difference between the results on the two datasets is caused by innate differences in the construction of the datasets. Concretely, an example of a visual analysis we do for dataset 1 is shown at https://github.com/EchoChou990919/LDAformer/blob/main/files/vis4dataset1.md. We find that the structure of dataset 1 is probably more consistent with the underlying assumption of LDA prediction that similar lncRNAs tend to be associated with similar diseases. Therefore prediction methods generally have

Table 1. Further comparisions on dataset 1

Method	Average AUC	Average AUPR
GCNLDA (2019)	0.960	0.223
CNNDLP (2019)	0.969	0.286
VADLP (2021)	0.956	0.449
GCRFLDA (2021)	0.959	0.405
GAERF (2021)	0.980	0.491
MGATE (2022)	0.964	0.413
GTAN (2022)	0.983	0.454
LDAformer	**0.994**	**0.709**

LDAformer outperforms other state-of-the-art methods.

higher AUC and AUPR values on dataset 1 than on dataset 2. And focusing on the comparison between methods, LDAformer is optimal.

Adjustments to the self-attention encoder

The classic structure of the Transformer encoder contains multi-head attention, residual connection, layer normalization and feedforward network. In addition to the core multi-head attention, we perform ablation analysis on the other three parts.

As shown in Table 2, experimental results prove that it is better to keep the original structure. The removal of residual connection causes the worst performance drop, because, without residual connection, the problem of degradation plagues deep networks. And it is interesting to discuss the role of layer normalization. On the one hand, layer normalization is conducive to model training and alleviates the problem of gradient explosion. On the other hand, the principle of LDA prediction is the multi-hop accessible pathways between lncRNA–disease pairs. If the lncRNA-disease-miRNA network is too sparse, layer normalization will damage the salience of feature embeddings inside the model. So on dataset2, the offset of these two influences makes the effect of the removal of layer normalization on the evaluation metrics minimal. And it is conceivable that in the far more sparse cases, we can even try to adjust the layer normalization part for better performance.

Ablation study

LDAformer predicts LDAs based on topological feature extraction and Transformer encoder. It has been experimentally demonstrated that our topological feature extraction process is effective, and it is optimal to retain the original structure of the Transformer encoder. In this section, we conduct further ablation experiments, replacing the Transformer encoder with more commonly used deep learning algorithms CNN or MLP. See Figure 6 and Supplementary Data 4, with well-designed structures, our MLP and CNN ablation models also achieve the ideal performance, comparable to or even slightly better than the baseline methods utilized the same algorithms (CNN: CNNDLP, MLP: DMFLDA, SDLDA). Such results are likely to benefit from our unique topological feature extraction, which introduces valuable information on multi-hop topological pathways. But they are still worse than the complete LDAformer with a Transformer encoder, which proves the irreplaceable powerful learning ability of the global self-attention, capturing the interdependency between any two topological pathways at the global scale.

Case study

In order to further verify the prediction capability of LDAformer in practical situations, we conduct case studies on both datasets separately. We take all known LDAs and equal-sized randomly selected unknown samples into training, and all unknown lncRNA–disease pairs for prediction. The prediction results are available at https://github.com/EchoChou990919/LDAformer and analyses are as follows.

For dataset1, by collecting experimental evidence in Lnc2Cancer v3.0 and lncRNADisease v2.0, we find there are 1258 experimentally verified LDAs that were not previously collected. Without being included in the learning process, 280 of them are still successfully predicted as associated (LDA score $>$ 0.5), which fully demonstrates the effectiveness of our method. As seen in Table 3, we investigate case studies on colon cancer (CC, DOID: 219), osteosarcoma (OS, DOID: 3347) and esophageal squamous cell carcinoma (ESCC, DOID: 3748).

CC is a type of colorectal cancer located in the colon, most of which are adenocarcinomas. It stands as a paradigm for our understanding of the molecular basis of human cancer [47]. Sorting by predicted scores in descending order, 9 of the top 10 candidate lncRNAs are confirmed by literature. For example, XIST expression level was upregulated in CC tissues and cell lines, and the growth rate of cells transfected with si-XIST was significantly decreased compared with that with si-NC, which was reversed by miR-34a targeted with 3'-UTR [48].

OS is a cancerous tumor in a bone. It is an aggressive malignant neoplasm that arises from primitive transformed cells of mesenchymal origin and that exhibits osteoblastic differentiation and produces malignant osteoid [49]. Here 7 of the top 10 candidate lncRNAs are confirmed by literature evidence. For example, increased EWSAT1 expression was associated with poor outcomes in osteosarcoma patients, and EWSAT1 could serve as a potential unfavorable prognostic biomarker [50].

ESCC is an esophageal carcinoma that derives from epithelial squamous cells located in the esophagus. It accounts for about 90% of the 456 000 incident esophageal cancers each year [51]. It is

Table 2. Adjustments to the self-attention encoder

Dataset	Add[1]	Norm[2]	FF[3]	Avg AUC	Avg AUPR
Dataset1	×	✓	✓	0.849	0.189
	✓	×	✓	0.988	0.589
	✓	✓	×	0.985	0.615
	✓	✓	✓	**0.994**	**0.709**
Dataset2	×	✓	✓	0.918	0.130
	✓	×	✓	0.940	0.231
	✓	✓	×	0.932	0.212
	✓	✓	✓	**0.941**	**0.233**

× or ✓ indicate the removal or retention of corresponding parts. [1]Abbreviation for residual connection. [2]Abbreviation for layer normalization. [3]Abbreviation for feedforward network.

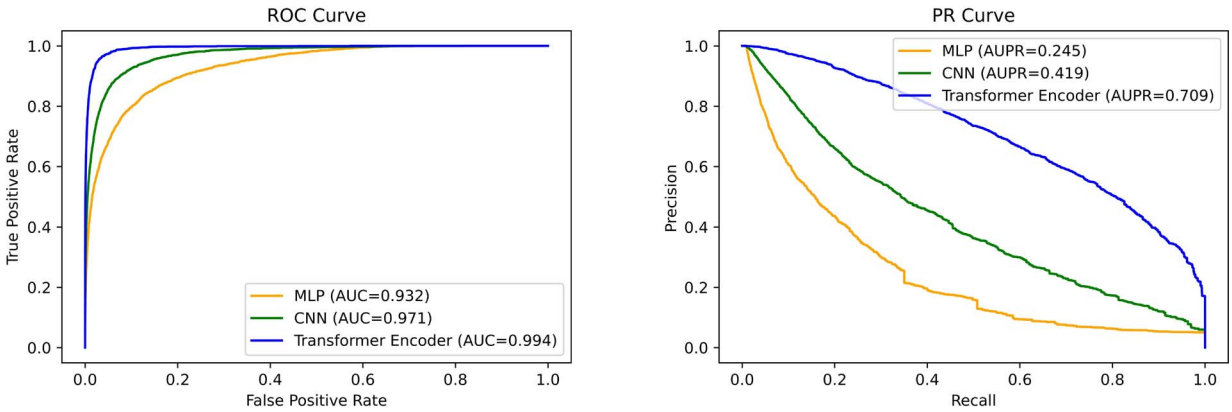

Figure 6. Replace the Transformer encoder with CNN or MLP. Comparisons of ROC curve, AUC value, PR curve and AUPR value on dataset1.

Table 3. The top 10 CC, OS or ESCC associated candidate lncRNAs

Disease	Rank	lncRNA	Literature Evidence (PMID)
Colon cancer (CC)	1	XIST	29679755*
	2	GAS5	28722800*
	3	PVT1	30504754*, 29552759*
	4	KCNQ1OT1	31040703*
	5	TUG1	31697952*, 27634385*
	6	NEAT1	32077635*, 31173354*
	7	UCA1	30652355*, 26885155*
	8	BANCR	28979803+
	9	HOTTIP	29274585+
	10	NPTN-IT1	Unconfirmed
Osteosarcoma (OS)	1	EWSAT1	29243774*, 27860482*
	2	PVT1	32021563*, 31699956*
	3	GAS5	30013899*, 31337976*
	4	H19	27186302*, 29568924*
	5	PCA3	Unconfirmed
	6	RMST	Unconfirmed
	7	DANCR	31918278*, 29753317*
	8	MIR155HG	Unconfirmed
	9	HCP5	30554864*
	10	CDKN2B-AS1	31724892*, 31950433*
Esophageal squamous cell carcinoma (ESCC)	1	HNF1A-AS1	25608466*
	2	MALAT1	31938345*, 25613496*
	3	MEG3	30990378*, 27778235*
	4	UCA1	25550835*, 30002691*
	5	BCYRN1	27143917+
	6	TINCR	26833746*
	7	CCAT2	25919911*, 25677908*
	8	HOTTIP	28534516*, 27806322*
	9	ZEB1-AS1	31638344*, 26617942*
	10	CBR3-AS1	24337686+

*Contained in Lnc2Cancer v3.0 or lncRNADisease v2.0. + Searched by ourselves.

impressive that all of the top 10 candidate lncRNAs are confirmed by literature evidence. For example, plasma levels of HNF1A-AS1 were significantly higher in ESCC patients compared with normal controls [52].

For dataset2, all unknown lncRNA–disease pairs are ranked by predicted LDA score. As shown in Table 4, we investigate the top 15 potential lncRNA–disease pairs and are able to find literature evidence for 12 of them. It further demonstrates the promising predictive capability of LDAformer. For example, PVT1 promotes invasive growth of lung adenocarcinoma cells by targeting miR-378c to regulate SLC2A1 expression [53]. PVT1/EZH2/LATS2 interactions might serve as targets for lung adenocarcinoma diagnosis and therapy [54]. Here, the confirmed LDAs are available for the diagnosis and treatment of diseases, and unconfirmed lncRNA–disease pairs might be able to guide biological experiments.

Conclusion

Biomedical research has revealed the crucial role of lncRNAs in many biological processes involved in diseases. And computational methods are increasingly proposed to predict potential LDAs, by which researchers obtain reliable LDAs at low cost and then contribute to the diagnosis and treatment of

Table 4. The top 15 potential lncRNA–disease pairs on Dataset2

Rank	LncRNA	Disease	Literature Evidence (PMID)
1	CDKN2B-AS1	stomach cancer	32767927
2	CDKN2B-AS1	lung non-small cell carcinoma	31775885
3	PVT1	lung adenocarcinoma	26908628, 32960438
4	CDKN2B-AS1	lung adenocarcinoma	Unconfirmed
5	TUG1	lung cancer	24853421, 28069000
6	XIST	lung cancer	27501756
7	CDKN2B-AS1	glioblastoma	Unconfirmed
8	PVT1	glioblastoma	34938610
9	DLEU2	colorectal cancer	33391439
10	HCP5	hepatocellular carcinoma	34148029
11	CDKN2B-AS1	renal cell carcinoma	32814766
12	CDKN2B-AS1	thyroid gland papillary carcinoma	Unconfirmed
13	CDKN2B-AS1	urinary bladder cancer	33182065
14	CASC15	colorectal cancer	33395735
15	LOXL1-AS1	colorectal cancer	32821123

Literature evidence is searched by ourselves.

diseases. In this work, we proposed LDAformer, an LDA prediction method based on topological feature extraction and self-attention mechanism. The associations among lncRNA, disease and miRNA are integrated from public databases. After the similarity calculation, to unify semantic information, inter-class associations and intra-class similarities are concatenated into the lncRNA-disease-miRNA weighted adjacency matrix. And based on it, we designed a topological feature extraction process to capture multi-hop pathway information. Then, we adopt a predictor based on the self-attention encoder to learn the interdependencies between pathways globally. Experimental results indicate that LDAformer performs better than baseline methods, and can accurately discover potential lncRNA–disease pairs in practical cases.

However, there are still some limitations. First, the basic principle of our method is the assumption that similar diseases tend to be associated with similar lncRNAs, which leads to neglect of lncRNAs or diseases with no known associations. Then, compared with the real world, only a small subset of lncRNAs and diseases are contained in the datasets, so the trend is to expand the magnitude of data. Transformer has the disadvantage of being computationally intensive and may have difficulty adapting to this trend. In the future, we will consider integrating more auxiliary biological associations, introducing more information such as RNA sequences, and reducing the computational effort.

Key Points

- We integrate heterogeneous lncRNA-disease-miRNA association data from several latest versions of public databases.
- A topology feature extraction process is designed that unifies similarity and association information and represents multi-hop pathways.
- Based on the powerful Transformer encoder, our method learns interdependencies between pathways and calculates the association scores of lncRNA–disease pairs.
- The computational experimental results on benchmark datasets show that our method outperforms other state-of-the-art methods, and case studies further demonstrate the capability to discover potential associations.

Data availability

The datasets and source code are available at https://github.com/EchoChou990919/LDAformer.

Supplementary data

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos.62172289) and Chengdu Science and Technology Project (2021-YF05-02071-SN).

References

1. Wapinski O, Chang HY. Long noncoding rnas and human disease. *Trends Cell Biol* 2011;**21**(6):354–61.
2. Fernandes JCR, Acuña SM, Aoki JI, *et al.* Long non-coding rnas in the regulation of gene expression: physiology and disease. *Noncoding RNA* 2019;**5**(1):17.
3. Wang J-J, Yang Y-C, Song Y-X, *et al.* Long non-coding rna ab007962 is downregulated in gastric cancer and associated with poor prognosis. *Oncol Lett* 2018;**16**(4):4621–7.
4. Tang T, Yang L, Cao Y, *et al.* Lncrna aatbc regulates pinin to promote metastasis in nasopharyngeal carcinoma. *Mol Oncol* 2020;**14**(9):2251–70.
5. Gao T, Liu X, He B, *et al.* Exosomal lncrna 91h is associated with poor development in colorectal cancer by modifying hnrnpk expression. *Cancer Cell Int* 2018;**18**(1):1–10.
6. Chen X, Yan G-Y. Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* 2013;**29**(20):2617–24.
7. Guangyuan F, Wang J, Domeniconi C, *et al.* Matrix factorization-based data fusion for the prediction of lncrna–disease associations. *Bioinformatics* 2018;**34**(9):1529–37.
8. Chengqian L, Yang M, Luo F, *et al.* Prediction of lncrna–disease associations based on inductive matrix completion. *Bioinformatics* 2018;**34**(19):3357–64.

9. Zhou M, Wang X, Li J, *et al.* Prioritizing candidate disease-related long non-coding rnas by walking on the heterogeneous lncrna and disease network. *Mol Biosyst* 2015;**11**(3):760–9.
10. Chen X, You Z-H, Yan G-Y, *et al.* Irwrlda: improved random walk with restart for lncrna-disease association prediction. *Oncotarget* 2016;**7**(36):57919–31.
11. 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 Bioinform* 2020;**19**(2):1–997.
12. Ping P, Wang L, Kuang L, *et al.* A novel method for lncrna-disease association prediction based on an lncrna-disease association network. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**(2):688–93.
13. Lan W, Li M, Zhao K, *et al.* Ldap: a web server for lncrna-disease association prediction. *Bioinformatics* 2017;**33**(3):458–60.
14. Pan X, Jensen LJ, Gorodkin J. Inferring disease-associated long non-coding rnas using genome-wide tissue expression profiles. *Bioinformatics* 2019;**35**(9):1494–502.
15. Yao D, Zhan X, Zhan X, *et al.* A random forest based computational model for predicting novel lncrna-disease associations. *BMC bioinformatics* 2020;**21**(1):1–18.
16. Zhu R, Wang Y, Liu J-X, *et al.* Ipcarf: improving lncrna-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. *BMC bioinformatics* 2021;**22**(1):1–17.
17. Zhang Y, Yan J, Chen S, *et al.* Review of the applications of deep learning in bioinformatics. *Current Bioinformatics* 2020;**15**(8):898–911.
18. Zeng M, Chengqian L, Fei Z, *et al.* Dmflda: a deep learning framework for predicting incrna–disease associations. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**18**:2353–63.
19. Zeng M, Chengqian L, Zhang F, *et al.* Sdlda: lncrna-disease association prediction based on singular value decomposition and deep learning. *Methods* 2020;**179**:73–80.
20. Xuan P, Cao Y, Zhang T, *et al.* Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncrna genes. *Front Genet* 2019;**10**:416.
21. Xuan P, Sheng N, Zhang T, *et al.* Cnndlp: a method based on convolutional autoencoder and convolutional neural network with adjacent edge attention for predicting lncrna–disease associations. *Int J Mol Sci* 2019;**20**(17):4260.
22. Xuan P, Jia L, Zhang T, *et al.* Ldapred: a method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncrnas. *Int J Mol Sci* 2019;**20**(18):4458.
23. Ximin W, Lan W, Chen Q, *et al.* Inferring lncrna-disease associations based on graph autoencoder matrix completion. *Comput Biol Chem* 2020;**87**:107282.
24. Shi Z, Zhang H, Jin C, *et al.* A representation learning model based on variational inference and graph autoencoder for predicting lncrna-disease associations. *BMC bioinformatics* 2021;**22**(1):1–20.
25. Zhao X, Zhao X, Yin M. Heterogeneous graph attention network based on meta-paths for lncrna–disease association prediction. *Brief Bioinform* 2022;**23**(1):bbab407.
26. Xuan P, Pan S, Zhang T, *et al.* Graph convolutional network and convolutional neural network based method for predicting lncrna-disease associations. *Cell* 2019;**8**(9):1012.
27. Wu Q-W, Xia J-F, Ni J-C, *et al.* Gaerf: predicting lncrna-disease associations by graph auto-encoder and random forest. *Brief Bioinform* 2021;**22**(5):bbaa391.
28. Fan Y, Chen M, Pan X. Gcrflda: scoring lncrna-disease associations using graph convolution matrix completion with conditional random field. *Brief Bioinform* 2022;**23**(1):bbab361.
29. Sheng N, Cui H, Zhang T, *et al.* Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncrna–disease association prediction. *Brief Bioinform* 2021;**22**(3):bbaa067.
30. Sheng N, Huang L, Wang Y, *et al.* Multi-channel graph attention autoencoders for disease-related lncrnas prediction. *Brief Bioinform* 2022;**23**(2):bbab604.
31. Xuan P, Zhan L, Cui H, *et al.* Graph triple-attention network for disease-related lncrna prediction. *IEEE Journal of Biomedical and Health Informatics* 2022;**26**(6):2839–49.
32. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Advances in neural information processing systems* 2017;**30**.
33. Ning S, Zhang J, Wang P, *et al.* Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic Acids Res* 2016;**44**(D1):D980–5.
34. Chen G, Wang Z, Wang D, *et al.* Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic Acids Res* 2012;**41**(D1):D983–6.
35. Lu Z, Cohen KB, Hunter L. Generif quality assurance as summary revision. *Pacific Symposium on Biocomputing* 2007;269–80.
36. Li J-H, Liu S, Zhou H, *et al.* Starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic Acids Res* 2014;**42**(D1):D92–7.
37. Yang L, Qiu C, Jian T, *et al.* Hmdd v2.0: a database for experimentally supported human microrna and disease associations. *Nucleic Acids Res* 2014;**42**(D1):D1070–4.
38. Gao Y, Shang S, Guo S, *et al.* 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):D1251–8.
39. Bao Z, Yang Z, Huang Z, *et al.* Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases. *Nucleic Acids Res* 2019;**47**(D1):D1034–7.
40. Huang Z, Shi J, Gao Y, *et al.* Hmdd v3.0: a database for experimentally supported human microrna–disease associations. *Nucleic Acids Res* 2019;**47**(D1):D1013–7.
41. Schriml LM, Mitraka E, Munro J, *et al.* (eds). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 2019;**47**(D1):D955–62.
42. Kozomara A, Birgaoanu M, Griffiths-Jones S. Mirbase: from microrna sequences to function. *Nucleic Acids Res* 2019;**47**(D1):D155–62.
43. Wang JZ, Zhidian D, Payattakool R, *et al.* A new method to measure the semantic similarity of go terms. *Bioinformatics* 2007;**23**(10):1274–81.
44. Wang D, Wang J, Ming L, *et al.* Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* 2010;**26**(13):1644–50.
45. Duncan A. Powers of the adjacency matrix and the walk matrix. *The Collection* 2004;**9**:4–11.
46. Kang C, Zhang H, Liu Z, *et al.* Lr-gnn: a graph neural network based on link representation for predicting molecular associations. *Brief Bioinform* 2022;**23**(1):bbab513.
47. Markowitz SD, Dawson DM, Willis J, *et al.* Focus on colon cancer. *Cancer Cell* 2002;**1**(3):233–6.
48. Sun N, Zhang G, Liu Y. Long non-coding RNA XIST sponges miR-34a to promotes colon cancer progression via Wnt/β-catenin signaling pathway. *Gene* 2018;**665**:141–8.
49. Luetke A, Meyers PA, Lewis I, *et al.* Osteosarcoma treatment–where do we stand? A state of the art review. *Cancer Treat Rev* 2014;**40**(4):523–32.

50. Zhang GY, Zhang JF, Hu XM, *et al*. Clinical significance of long non-coding rna ewsat1 as a novel prognostic biomarker in osteosarcoma. *Eur Rev Med Pharmacol Sci* 2017;**21**(23): 5337–41.
51. Abnet CC, Arnold M, Wei W-Q. Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* 2018;**154**(2): 360–73.
52. Tong Y-S, Wang X-W, Zhou X-L, *et al*. Identification of the long non-coding rna pou3f3 in plasma as a novel biomarker for diagnosis of esophageal squamous cell carcinoma. *Mol Cancer* 2015;**14**(1):1–13.
53. Xia H, Zhang Z, Yuan J, *et al*. The lncrna pvt1 promotes invasive growth of lung adenocarcinoma cells by targeting mir-378c to regulate slc2a1 expression. *Hum Cell* 2021;**34**(1):201–10.
54. Wan L, Sun M, Liu G-J, *et al*. Long noncoding rna pvt1 promotes non–small cell lung cancer cell proliferation through epigenetically regulating lats2 expression. *Mol Cancer Ther* 2016;**15**(5):1082–94.