JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Learning Association Characteristics by Dynamic Hypergraph and Gated Convolution Enhanced Pairwise Attributes for Prediction of Disease-Related lncRNAs

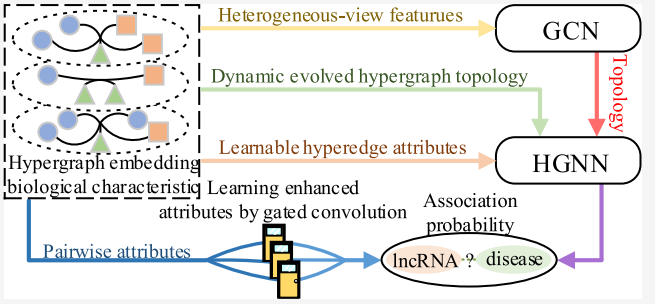
Ping Xuan, Siyuan Lu, Hui Cui, Shuai Wang, Toshiya Nakaguchi, and Tiangang Zhang*

Cite This: https://doi.org/10.1021/acs.jcim.4c00245 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via HEILONGJIANG UNIV on April 4, 2024 at 03:23:49 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: As the long non-coding RNAs (lncRNAs) play important roles during the incurrence and development of various human diseases, identifying disease-related lncRNAs can contribute to clarifying the pathogenesis of diseases. Most of the recent lncRNA-disease association prediction methods utilized the multi-source data about the lncRNAs and diseases. A single lncRNA may participate in multiple disease processes, and multiple lncRNAs usually are involved in the same disease process synergistically. However, the previous methods did not completely exploit the biological characteristics to construct the informative prediction models. We construct a prediction model based on adaptive hypergraph and gated convolution for lncRNA-disease association prediction (AGLDA), to embed and encode the biological characteristics about lncRNA−disease associations, the topological features from the entire heterogeneous graph perspective, and the gated enhanced pairwise features. First, the strategy for constructing hyperedges is designed to reflect the biological characteristic that multiple lncRNAs are involved in multiple disease processes. Furthermore, each hyperedge has its own biological perspective, and multiple hyperedges are beneficial for revealing the diverse relationships among multiple lncRNAs and diseases. Second, we encode the biological features of each lncRNA (disease) node using a strategy based on dynamic hypergraph convolutional networks. The strategy may adaptively learn the features of the hyperedges and formulate the dynamically evolved hypergraph topological structure. Third, a group convolutional network is established to integrate the entire heterogeneous topological structure and multiple types of node attributes within an lncRNA−disease−miRNA graph. Finally, a gated convolutional strategy is proposed to enhance the informative features of the lncRNA−disease node pairs. The comparison experiments indicate that AGLDA outperforms seven advanced prediction methods. The ablation studies confirm the effectiveness of major innovations, and the case studies validate AGLDA's ability in application for discovering potential disease-related lncRNA candidates.



INTRODUCTION

The long non-coding RNA (lncRNA) is an RNA molecule that exceeds 200 nucleotides in length and cannot be translated into a protein.[1−3] A growing body of research shows that the imbalance of lncRNA concentrations is closely linked with the development of various diseases, such as prostate cancer, Parkinson's disease, and diabetes mellitus.[3,4] Therefore, identifying lncRNAs associated with diseases is crucial for diagnosing and exploring disease pathogenesis.

The computational prediction method for a lncRNA-related disease has been shown to be effective in screening candidate lncRNAs associated with the disease.[5,6] Previous methods developed can be categorized into three groups. For the first group, the genomic location, tissue specificity, and expression profile use biological information related to lncRNA to predict the lncRNA−disease association. However, genomic location-based methods do not apply to lncRNAs without adjacent genes.[7,8] Gene-recorded diseases limit these approaches that require tissue specificity and expression profiles.[9,10]

The second group consists of machine learning methods. Chen and Yan[11] established a prediction model based on the hypothesis that similar diseases are related to lncRNAs with similar functions. A new similarity calculation method between lncRNAs was proposed to improve the association prediction performance. Zhao et al.[12] proposed a prediction model based on naïve Bayesian classifier. In addition, methods based on random walk (RW),[13,14] support vector machines (SVMs),[15,16] matrix factorization (MF),[17,18] matrix completion (MP),[19,20] and random forest[21,22] have been used to infer the tendency of the lncRNA−disease association. However, these methods do

Received: February 11, 2024
Revised: February 22, 2024
Accepted: February 27, 2024

© XXXX American Chemical Society

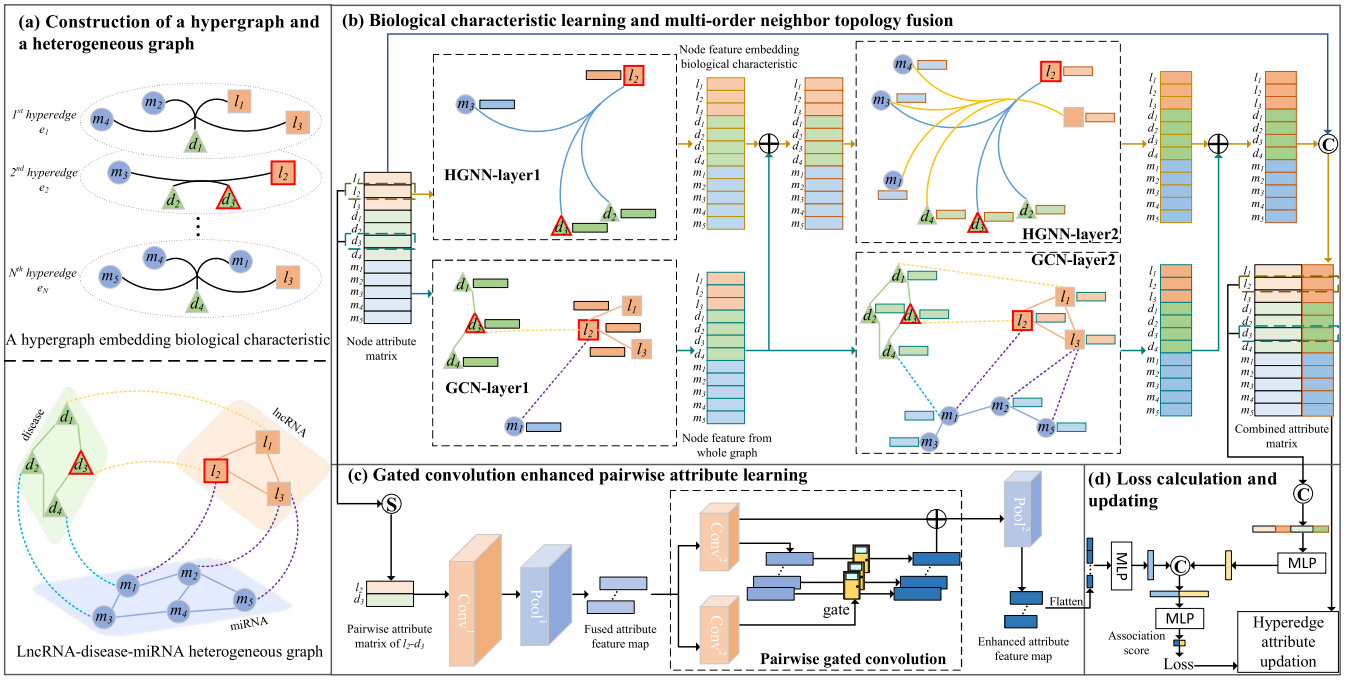


Figure 1. Overall framework of the AGLDA model. (a) Construction of a hypergraph embedding the association characteristic and a heterogeneous graph with multiple kinds of nodes. (b) Learning biological association characteristic of nodes and fusing multi-order neighbor topologies. (c) Learning the pairwise attributes with gated convolution enhanced strategy. (d) Calculation of the association estimation loss and updating the attributes of each hyperedge.

not utilize miRNA or protein information related to lncRNAs and diseases.

Several recent methods were proposed based on deep learning, and they are utilized for predicting the circRNA−disease associations,[23,24] the miRNA−disease associations,[25,26] the disease-related genes,[27,28] and the disease-related lncRNAs. These methods achieved decent prediction performances, which indicated that learning the deep and representative features is important for the improved prediction performance. One model was constructed using a dual convolutional neural network to predict lncRNA-associated diseases.[29] Another model composed of a convolutional autoencoder and variance autoencoder was proposed to learn node pair attributes and distributions.[30] Meta-paths, graph neural networks, and recurrent neural networks have also been used for designing prediction models.[31−35] Despite these previous methods, they overlooked the complex connections between multiple lncRNAs (miRNAs) and multiple diseases, representing the key biological characteristics between lncRNA (miRNA) and diseases.

In this article, we present a novel association prediction method based on the **a**daptive dynamic hyper**g**raph and **g**ated convolution for **l**ncRNA-**d**isease **a**ssociation prediction (AGLDA), for adaptively learning the characteristics from multiple biological views, the topological features from a heterogeneous network, and the pairwise attributes. The contributions provided by AGLDA include the following.

First, an lncRNA is often involved in multiple disease processes, and several lncRNA and miRNAs may synergistically join within the same disease process. A hyperedge can denote the complex relationships among multiple lncRNAs and diseases. Furthermore, multiple hyperedges reflect diverse relationships from multiple views.

Second, we design a strategy to leverage adaptive learning of the hyperedge features, which is beneficial for forming the dynamic hypergraph topological structure. A dynamic hypergraph neural network (DHNN) module is constructed to encode the evolutive hypergraph topological structures.

Third, a heterogeneous graph formed with the lncRNA, disease, and miRNA nodes is established with various connections that formulate a complicated heterogeneous topology. We construct a module based on heterogeneous graph convolutional networks (MHGCN) to fuse attributes from multiple node types and their topology from the entire heterogeneous graph perspective.

Finally, a novel strategy based on gated convolutional neural networks (GCNN) is presented to enhance the more informative attributes of a pair of lncRNA−disease nodes. The improved prediction performance and the effectiveness of our proposed innovations are demonstrated by comparing our method with seven advanced prediction methods and performing ablation studies.

MATERIALS AND METHODS

To predict disease-related lncRNAs, our AGLDA method incorporates an adaptive hypergraph and pairwise gated convolution, as shown in Figure 1. We constructed a heterogeneous graph by the association, similarity, and interaction between lncRNAs, diseases, and miRNAs. Hypergraphs are established to embed biological perspective information on the association between lncRNA and diseases. The multiorder topology information on heterogeneous graphs, the biological characteristic information on hypergraphs, and the node pair attribute information are combined to infer an association score between each pair of lncRNA−disease nodes.

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

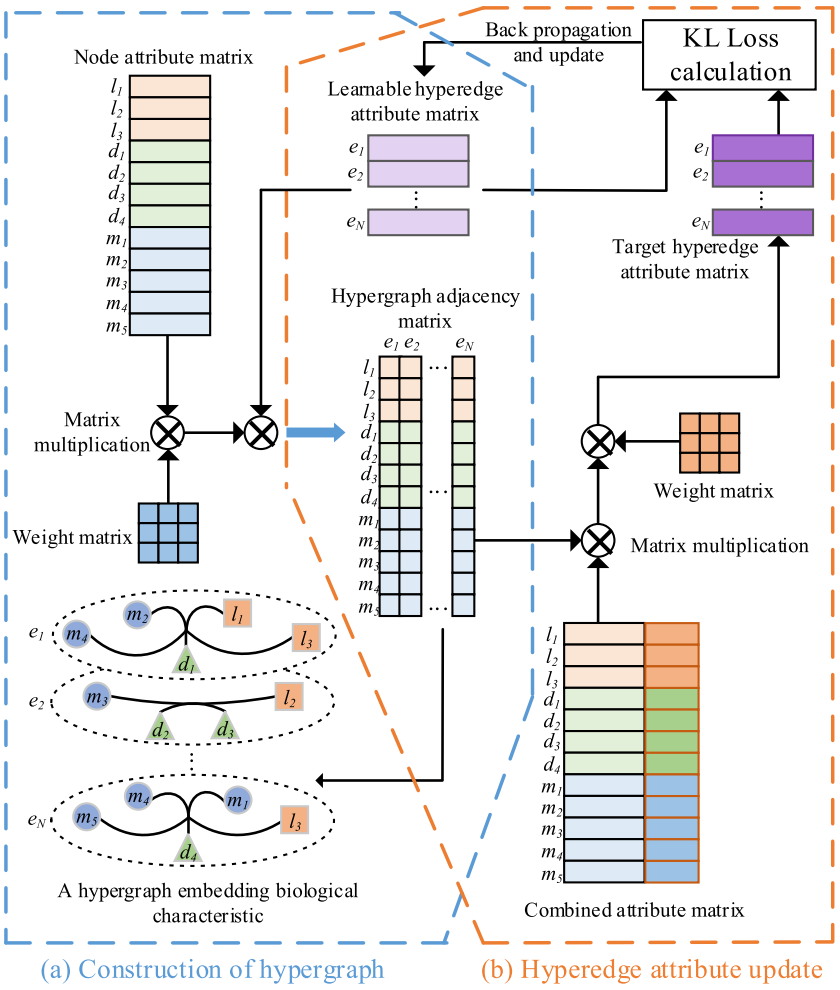


Figure 2. Process of updating hyperedge attributes and hypergraph topologies. (a) Construction of the hypergraph based on hyperedge attributes. (b) Updating the hyperedge attributes.

Data Set. The lncRNA−disease−miRNA similarity, lncRNA−disease and miRNA−disease association, and lncRNA−miRNA interaction are obtained from a previous work[36] for predicting lncRNA−disease association. The data set includes 2687, 13,559, and 1002 pairs of lncRNA−disease associations, disease−miRNA associations, and lncRNA−miRNA interactions, respectively, including 240 lncRNAs, 405 diseases, and 495 miRNAs.

Construction of the lncRNA−Disease−miRNA Heterogeneous Graph. Our heterogeneous graph $\mathcal{G}_o = (\mathcal{V}, \mathcal{E}_o)$ represents the similarity between lncRNA, disease, and miRNA, where the node set $\mathcal{V}$ consists of the node subsets of lncRNA $\mathcal{V}_{lr}$, disease $\mathcal{V}_{di}$, and miRNA $\mathcal{V}_{mr}$. The edge set $\mathcal{E}_o$ is denoted by the association interaction matrix M_{ai} and the similarity matrix M_{si}.

The association interaction matrix M_{ai} is defined as

$$M_{ai} = \begin{cases} M_{ai}^{lr-di} \in \{0, 1\}^{|\mathcal{V}_{lr}|\times|\mathcal{V}_{di}|} \\ M_{ai}^{lr-mr} \in \{0, 1\}^{|\mathcal{V}_{lr}|\times|\mathcal{V}_{mr}|} \\ M_{ai}^{di-mr} \in \{0, 1\}^{|\mathcal{V}_{di}|\times|\mathcal{V}_{mr}|} \end{cases} \quad (1)$$

where M_{ai}^{lr-di} (M_{ai}^{di-mr}) and M_{ai}^{lr-mr} denote the lncRNA−disease (disease−miRNA) association matrix and the lncRNA−miRNA interaction matrix, respectively. $|\mathcal{V}|_{lr}$, $|\mathcal{V}|_{di}$, and $|\mathcal{V}|_{mr}$ represent the number of lncRNA, disease, and miRNA, respectively. If $(M_{ai}^{lr-di})_{ij}((M_{ai}^{di-mr})_{ij})$ is 1, then lncRNA (miRNA) l_i and disease d_j are associated, or 0, otherwise. If lncRNA l_i and miRNA m_j represent a known interaction, then $(M_{ai}^{lr-mr})_{ij} = 1$, and $(M_{ai}^{lr-mr})_{ij} = 0$, otherwise.

The similarity matrix M_{si} is defined as

$$M_{si} = \begin{cases} M_{si}^{lr} \in [0, 1]^{|\mathcal{V}_{lr}|\times|\mathcal{V}_{lr}|} \\ M_{si}^{di} \in [0, 1]^{|\mathcal{V}_{di}|\times|\mathcal{V}_{di}|} \\ M_{si}^{mr} \in \{0, 1\}^{|\mathcal{V}_{mr}|\times|\mathcal{V}_{mr}|} \end{cases} \quad (2)$$

where M_{si}^{lr} (M_{si}^{di}, M_{si}^{mr}) denotes the lncRNA (disease, miRNA) similarity matrix. According to Wang's method,[37] a directed acyclic graph (DAG) was constructed for each disease, and the graph includes all the disease terms about the disease. The similarity of two diseases was calculated based on their DAGs. Following the previous methods,[37,38] the similarity between two lncRNAs (miRNAs) was calculated based on the diseases associated with them. For instance, for two lncRNAs, such as $\mathcal{V}_i^{lr}$ and $\mathcal{V}_j^{lr}$, S_i^{lr} and S_j^{lr} are the disease sets that these two lncRNAs are associated with. The similarity between S_i^{lr} and S_j^{lr} was regarded as the similarity between $\mathcal{V}_i^{lr}$ and $\mathcal{V}_j^{lr}$. A higher

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

value of $(M_{si}^{lr})_{ij}$ ($(M_{si}^{di})_{ij}$, $(M_{si}^{mr})_{ij}$) corresponds to a greater similarity between lncRNA l_i (disease d_i, miRNA m_i) and l_j (d_j, m_j).

The adjacency matrix $F \in [0, 1]^{|\mathcal{V}|\times|\mathcal{V}|}$ of the heterogeneous graph is established as

$$F = \begin{bmatrix} M_{si}^{lr} & M_{ai}^{lr-di} & M_{ai}^{lr-mr} \\ M_{ai}^{lr-di^T} & M_{si}^{di} & M_{ai}^{di-mr} \\ M_{ai}^{lr-mr^T} & M_{ai}^{di-mr^T} & M_{si}^{mr} \end{bmatrix} \tag{3}$$

where M^T denotes the transposition of M. $|\mathcal{V}|$ represents the number of lncRNA−disease−miRNA. The i-th line F_i of F denotes the feature vector of the i-th node, for which the i-th node is similar, associated, or interacts with all lncRNAs, diseases, and miRNAs. Therefore, F is regarded as a node attribute matrix.

Constructing a Hypergraph with Biological Perspective Information. An lncRNA (miRNA) is often involved in multiple disease processes, and several lncRNAs (miRNAs) may synergistically participate in the same disease process. Based on this biological characteristic, we construct a hypergraph $\mathcal{G}_h = (\mathcal{V}, \mathcal{E}_h)$ to embed information from this biological perspective. $\mathcal{E}_h$ is a set of hyperedges, which represent the internal connections of some lncRNA−disease and disease−disease. For example, in Figure 1a, the hyperedge e_2 denotes the involvement of lncRNA l_2 in the disease development of d_2 and d_3. The hyperedge e_1 represents the synergistic joining of lncRNAs l_1 and l_3 to disease d_1.

In Figure 2a, we also establish a hypergraph attribute matrix $E \in \mathbb{R}^{N\times d}$, where N is the number of hyperedges and d is the size of the hyperedge embedding. The hypergraph adjacency matrix $H \in \mathbb{R}^{|\mathcal{V}|\times N}$ is formed based on the hyperedge attribute matrix E and the node attribute matrix F.

$$H = \text{LeakyRelu}(D_T^{-1/2} F D_T^{-1/2} W_H E^T) \tag{4}$$

where $D_T^{-1/2} F D_T^{-1/2}$ denotes the Laplace normalization of F, $D_T \in \mathbb{R}^{|v|\times|v|}$ is the diagonal degree matrix of F, $W_H \in \mathbb{R}^{|v|\times d}$ is a learnable weight matrix, and LeakyRelu is nonlinear activation function. The i-th column of H represents the i-th hyperedge in the hypergraph. The purpose of introducing F to update H ensures that each hyperedge of the hypergraph considers the attribute information on all nodes. E is randomly initialized and learnable, so the hypergraph dynamically changes during training. The learnable hyperedge attributes are helpful for forming the dynamic hypergraph topology during the training process. In this way, the hypergraph could reflect the biological characteristics about the associations among multiple lncRNA and disease nodes.

Learning Biological Characteristics Associated with lncRNA (miRNA) and Diseases with Multi-Order Neighbor Topology. We next construct a learning model based on DHNN and MHGCN used to combine the biological information on $\mathcal{G}_h$ with the multiorder topology information on $\mathcal{G}_o$.

Biological Characteristic Information Learning (DHNN). The hypergraph contains the biological characteristics associated with lncRNAs (miRNA) and diseases. To better integrate the topology and node attributes of the hypergraph, we first used a learning strategy based on hypergraph neural networks. As the input hypergraph (H, F), the l-th layer output X_B^l is represented as

$$X_B^l = \text{LeakyReLu}(D_B^{-1/2} H W_{\mathcal{E}} D_{\mathcal{E}}^{-1} H^T D_B^{-1/2} X^{l-1} W_B^l) \tag{5}$$

where $W_{\mathcal{E}}$ is the hyperedge weight matrix initialized as the identity matrix. W_B^l is the weight matrix of the l-th layer DHNN. D_B and $D_{\mathcal{E}}$ are the degree matrices of the nodes and hyperedges, respectively, where $(D_B)_{i,i} = \sum_{e\in\mathcal{E}_h} H_{e,i}$ and $(D_{\mathcal{E}})_{i,i} = \sum_{v\in\mathcal{V}} H_{i,v}$, which together function to normalize $HW_{\mathcal{E}}H^T$. $HW_{\mathcal{E}}D_{\mathcal{E}}^{-1}H^T$ denotes the adjacency matrix based on the hyperedge distribution for all nodes. $X^0 = F$ is the attribute matrix for all lncRNAs, diseases, and miRNAs. X_B^l represents the node attribute matrix embedded with biological characteristics information.

Fusion of Multi-Order Neighbor Topology Information (MHGCN). The hypergraph neural network focuses more on the neighbor node attributes of the source lncRNA (disease, miRNA) in $\mathcal{G}_h$, neglecting the neighbor topology learning of lncRNA−disease−miRNA $\mathcal{G}_o$. X_T^1 combines the first-order neighbor attribute feature of each source node for encoding to obtain the attribute feature of each node. Similarly, X_T^l fuses the attribute matrix learned from the l-order neighbor topology

$$X_T^l = \text{LeakyRelu}(D_T^{-1/2} F D_T^{-1/2} X_T^{l-1} W_T^l) \tag{6}$$

where W_T^1 represents the weight matrix of the l-th layer MHGCN and $X_T^0 = F$ and X_T^l denote the node attribute matrix embedded with the l-order topology information. In the same way, X_B^l considers the embedding of the biological characteristic reflected by the l-order nodes in the hypergraph. Therefore, we supplement the corresponding topological embedding X_T^l with the biological characteristic embedding X_B^l of the l-th layer DHNN to form X^l

$$X^l = X_B^l + X_T^l \tag{7}$$

where we represent the last layer output embedded as X^L. The original node attribute matrix F contains more detailed features, so we fuse F and X^L to form P

$$P = [X^L \| F] \tag{8}$$

where $\|$ is the concatenate operation. The i-th line of P, P_i, represents the feature vector of the i-th node containing biological characteristics, topological features, and original attributes.

Dynamic Learning of Hyperedge Attributes. The hyperedge attribute matrix E is randomly initialized and can be adaptively adjusted during the learning process. H contains all lncRNAs, diseases, and miRNAs node hyperedge topologies, and P is the feature matrix learned by each node. $H^T\hat{F}$ fully integrates the topology and node attributes to form an estimated hyperedge feature matrix, which then assists E in learning. Therefore, the target hyperedge attribute matrix $\hat{E} \in \mathbb{R}^{N\times b}$ (Figure 2b) is defined as

$$\hat{E} = \text{LeakyRelu}(H^T P W_E) \tag{9}$$

where W_E is a weight matrix. KL divergence can measure the difference between $\hat{E}$ and E, so we establish a KL divergence loss L_E between E and $\hat{E}$

$$L_E = \gamma(E) \log\left(\frac{\gamma(E)}{\gamma(\hat{E})}\right) \tag{10}$$

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

where $(x) = \sum_{p=1}^{|\mathcal{V}|}\sum_{q=1}^{d}\mathrm{SoftMax}(x)$ and SoftMax is a nonlinear activation function. Minimizing L_E can help E learn better biological characteristics.

Pairwise Attribute Learning by Gated CNN (GCNN). If a pair of lncRNA and disease such as l_i and d_j has similarities, associations, and interactions with more common lncRNAs, diseases, and miRNAs, it is more possible that l_i is associated with d_j. On the basis of the biological premise, we stack the attribute feature of l_i and d_j to obtain the pairwise attribute matrix F_y

$$F_y = [F_i; F_j] \tag{11}$$

where $[;]$ is the stack operation. The 1-st to $|\mathcal{V}_{lr}|$-th columns of F_y represent the similarities and associations between l_i−d_j and all lncRNAs. The $(|\mathcal{V}_{lr}| + 1)$-th to $(|\mathcal{V}_{lr}| + |\mathcal{V}_{di}|)$-th columns of F_y denote the similarities and associations between l_i−d_j and all diseases. The $(|\mathcal{V}_{lr}| + |\mathcal{V}_{di}| + 1)$-th to $(|\mathcal{V}_{lr}| + |\mathcal{V}_{di}| + |\mathcal{V}_{mr}|)$-th columns of F_y are the associations and interactions between l_i−d_j and all miRNAs.

We next established a GCNN module to further enhance and screen for features that are beneficial for predicting lncRNA−disease associations in F_y. This module includes two convolution-pooling layers. As shown in Figure 1c, $Conv_1$ and $Pool_1$ represent the first convolutional layer and the first pooling layer, respectively. It fuses the features of l_i and the ones of d_j to form the feature map Y_{fu}

$$Y_{fu} = \mathrm{MaxPooling}(\mathrm{LeakyRelu}(W_{fu} * F_y + b_{fu})) \tag{12}$$

where * is a convolutional operation and MaxPooling is the maximum pooling operation. W_{fu} denote the convolution kernel and b_{fu} is the bias vector.

The second convolutional layer and pooling layer are denoted as $Conv_2$ and $Pool_2$ (Figure 1c). The second layer includes convolution and pooling along with a gating mechanism to enhance the node learning of the attributes (Figure 3). Y_{fu} first uses convolution operations to increase the dimensionality to form a richer pairwise attribute matrix Y_{ri}

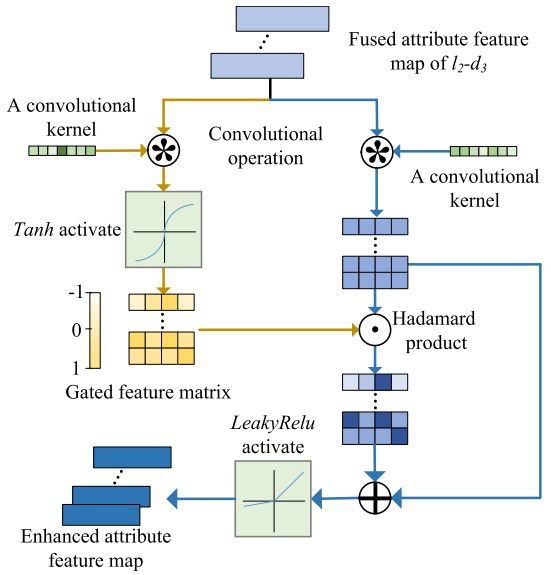


Figure 3. Illustration of the proposed pairwise attribute learning strategy enhanced by gated convolution.

$$Y_{ri} = W_{ri} * Y_{fu} + b_{ri} \tag{13}$$

where Y_{fu} undergoes another convolution operation and passes through the gated feature matrix Y_{ga} with activation function tanh

$$Y_{ga} = \tanh(W_{ga} * Y_{fu} + b_{ga}) \tag{14}$$

where the value of $(Y_{ga})_{ij}$ is between −1 and 1. The closer $(Y_{ga})_{ij}$ is to 1, the more important is the attribute of the node pair. Otherwise, the closer $(Y_{ga})_{ij}$ is to −1, the more trivial are the attributes of the node pair. Therefore, this construction represents a gated mechanism for adjusting the original attribute matrix Y_{ri}. Y_{ga} and Y_{ri} form an enhanced pairwise attribute matrix Y_{en}

$$Y_{en} = \mathrm{MaxPooling}(\mathrm{LeakyRelu}(Y_{ga} \otimes Y_{ri} + Y_{ri})) \tag{15}$$

where $\otimes$ is the Hadamard product operation. Due to the more detailed features in the original attribute matrix Y_{ri}, we added it to the learning process of Y_{en}.

Evaluation and Optimization of Association Scores. For each node pair, lncRNA l_i and disease d_j, we extract and concatenate the feature vectors P_i and P_j. To combine the pairwise feature, the concatenated vector enters the fully connected layer to obtain the l_i−d_j fused feature vector Z_f

$$Z_f = \mathrm{LeakyRelu}(W_f[P_i \| P_j] + b_f) \tag{16}$$

where W_f is a weight matrix and b_f is the bias vector. Y_{en} forms a more representative pairwise vector Z_s through the fully connected layers

$$Z_s = \mathrm{LeakyRelu}(W_s Y_{en} + b_s) \tag{17}$$

To obtain a pair of lncRNA−disease association probabilities y_p, we concatenate and feed Z_f and Z_s into the fully connected layer. Here, y_p is defined as

$$y_p = \mathrm{SoftMax}(W_y[Z_f \| Z_s] + b_y) \tag{18}$$

where Softmax is a nonlinear activation function. Then, $y_p = [y_{p,0}, y_{p,1}]$, where $y_{p,0}$ and $y_{p,1}$ are the probabilities of a pair of lncRNA being associated with the disease, respectively.

We used the cross-entropy loss function to optimize the prediction model. This loss of the lncRNA and disease between the predicted score y_p and the real label y_l is L_C

$$L_C = \sum_{b=1}^{B}[-y_l \log y_{p,1} - (1 - y_l)\log(1 - y_{p,1})] \tag{19}$$

where B denotes the number of samples trained. If the association between a pair of lncRNA and disease is known, then $y_l = 1$. Otherwise, $y_l = 0$. The total loss L_T of the model is defined as

$$L_T = L_C + \alpha L_E \tag{20}$$

where the loss function L_T is optimized through the Adam algorithm. The value of α is used to adjust the contribution of the KL divergence loss to the total loss calculation.

EXPERIMENTAL EVALUATION AND ANALYSIS

Evaluation Strategy. A 5-fold cross-validation method is used to evaluate the performance of AGLDA and that of the compared models. The positive sample consists of a data set with known lncRNA−disease associations randomly divided

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

into five equal parts. Negative samples represent all of the unknown associations. The training set includes four-fifths of the positive samples and an equal number of randomly sampled negative samples. The remaining samples are considered as the test set.

The area under the receiver operating characteristic (ROC) Curve (AUC),[39] the area under the precision−recall curve (AUPR),[40] and the top k recall rate are used as evaluation indicators for model prediction performance. To compare all models, we averaged the AUC and AUPR averages across each fold. Biologists tend to choose predicted candidates with higher rankings to guide further experimentation, so the recall rate of the top k (30, 60, ..., 240) is used to measure the model capability to correctly identify associations.

Experimental Details. The AGLDA model is trained on an NVIDIA GeForce GTX 1050Ti graphic card with 8G memory. The PyTorch framework is used for model training configured with learning rate, weight decay rate, batch, and epochs of 0.0003, 0.001, 32, and 150, respectively. The hyperedge matrix E contains 200 hyperedges with a feature dimension of 300. We use two layers of DHNN and MHGCN, where the output feature dimensions for each layer are 600 nm and 300, respectively. The first and second convolutions have 32 sizes of 2×7 and 64 sizes of 1×7 filters, respectively. The experiment results about the hyperparameter selection are listed in the Supporting Information SF1.

Ablation Experiments. We evaluate their contribution by removing the main components of AGLDA (Table 1), which include biological characteristic topological learning (BCTL), pairwise node attribute learning (GCNN), and dynamic learning of hyperedge attributes (DLHA). The combined models of BCTL, DLHA, and GCNN acquired the best AUC of 0.988 and AUPR of 0.684. Removing only BCTL, the AUC (AUPR) drops by 0.5% (6.6%), suggesting that the biological and topological features help improve prediction performance. Without GCNN, the AUC (AUPR) reduces by 3.2% (12.6%) less than the complete AGLDA. GCNN has the most contribution to the increased prediction performance. The possible reason is the embedding of the lncRNA−disease node pair, which includes the direct similarity, association, and interaction relationships between the pair and all the lncRNAs, diseases, and miRNAs. These relationships provide the effective information for the lncRNA−disease association prediction. The biological characteristic topological learning includes biological characteristic learning based on hypergraph neural networks (DHNN) and topology fusion based on graph convolutional networks (MHGCN). The AUC (AUPR) of AGLDA decreases by 2% (3.3%) in the absence of DHNN. Without MHGCN, the AUC (AUPR) reduces by 1% (4.4%). These results indicate that DHNN and MHGCN contribute significantly to the model performance. Finally, the complete model achieves 0.1 and 2.7% improvement in AUC and AUPR, respectively, compared to those of the model without DLHA. The possible reason is that updating the hyperedge attributes continuously is helpful for forming the dynamic hypergraph topology. Furthermore, the topology is beneficial for reflecting the biological characteristics about the associations among multiple lncRNAs, miRNAs, and diseases.

Table 1. Ablation Experimental Results of Our Method

	BCTL	GCNN	DHNN	MHGCN	DLHA	AGLDA
average AUC	0.983	0.956	0.986	0.987	0.987	**0.988**
average AUPR	0.618	0.558	0.651	0.640	0.657	**0.684**

There is another fusion strategy in which the topological information from the last layer of MHGCN and the biological characteristic information from the last layer of DHNN are fused (fusion at the last layer). We compared the strategy and our original strategy (fusion at each layer) and demonstrated the experimental result in Table 2. The AUC and AUPR of our strategy are 0.988 and 0.684, respectively, and they are 0.1 and 2.2% higher than the values of the fusion strategy at the last layer. The possible reason is that fusion of the information from the MHGCN encoding layer and the one from the DHNN encoding layer is beneficial for fully integrating the biological characteristics about the associations and the heterogeneous graph topology.

Table 2. Prediction Results for the Different Strategies and Data

strategy	AUC	AUPR
fusion at each layer	0.988	0.684
fusion at the last layer	0.987	0.662
dynamic hyperedge	0.988	0.684
static hyperedge	0.987	0.652
complete heterogeneous graph	0.988	0.684
G_p	0.969	0.643

There is a static hyper-edge construction strategy. A hyperedge is formed by the top k similar miRNAs, the associated diseases, and the interacted miRNAs with a lncRNA l_i. We compared the static hyperedge construction strategy and the dynamic one. As shown in Table 2, the AUC and AUPR of the dynamic strategy are 0.1 and 3.2%, respectively, higher than those of the static strategy. The possible reason is that the dynamic strategy could update the hyperedge attributes and the hypergraph topology, which reveals the biological characteristics about the lncRNA−disease associations.

A heterogeneous graph G_p was established, and it was only composed of the associations and interactions among the lncRNAs, diseases, and miRNAs. Our model (AGLDA) was conducted on G_p and obtained an AUC of 0.909 and an AUPR of 0.334. The AUC and AUPR for the complete heterogeneous graph with similarity connections were 1.9 and 4.1%, respectively, higher than those for G_p. It showed the lncRNA similarities, the disease ones, and the miRNA ones are important for the improved prediction performance.

Performance Comparison. We compare AGLDA with seven advanced methods for predicting disease-associated lncRNAs, including GAIRD,[34] GSMV,[33] MGLDA,[32] GTAN,[31] iLncDA-LTR,[41] VADLP,[30] and CNNLDA.[29]

- GAIRD: The method integrated the homogeneous and heterogeneous information from the lncRNA−-disease−miRNA network and inferred the disease-related lncRNAs by the graph convolution and group convolution.
- GSMV: The prediction model was constructed based on meta-path and transformer to encode the semantic information from multiple perspectives and the dependencies among the lncRNA, disease, and miRNA nodes.
- MGLDA: The method sampled subgraphs based on multiple meta-paths, and then, the subgraph topologies

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

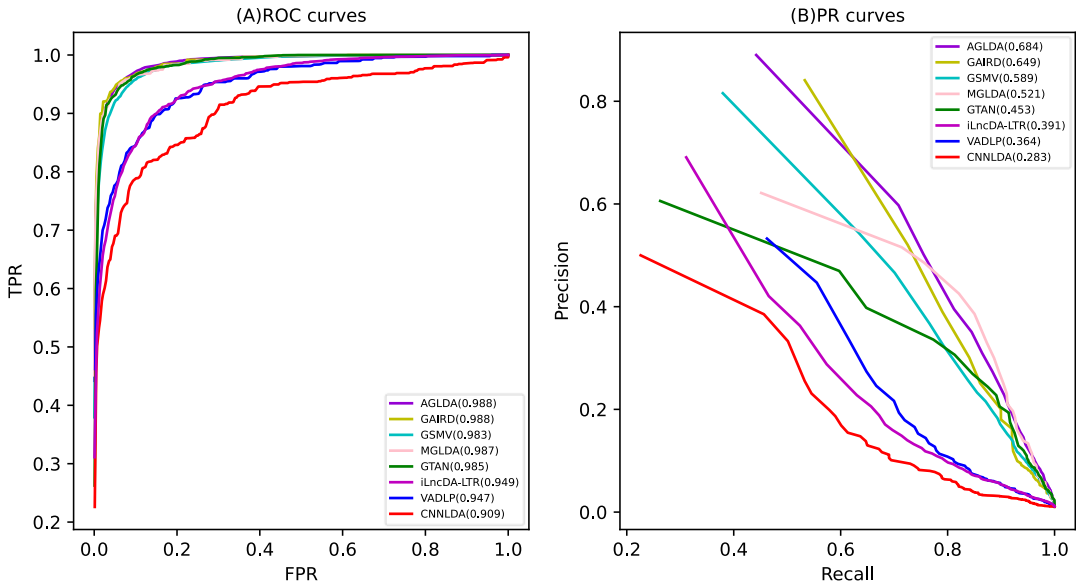


Figure 4. ROC and PR curves of AGLDA and the compared methods.

were encoded and integrated by the graph convolutional autoencoder and the topology-level attention.

- GTAN: GTAN encoded the neighbor topology of each lncRNA (disease) node and the pairwise attributes by the graph attention mechanism and convolutional neural network, respectively.
- iLncDA-LTR: The lncRNA similarities and the disease similarities were calculated by the DOSE package and the Needleman Wunsch alignment method. The disease-related candidate lncRNAs were predicted by Adaboost, Xgboost, and k-NN.
- VADLP: The strategy based on the convolutional and variation autoencoders was designed to learn the attributes and the attribute distribution for each pair of lncRNA and disease nodes.
- CNNLDA: It learned the original feature representation and the attention one for a lncRNA-disease node pair by a proposed dual convolutional neural networks.

All of the methods use the same training and testing sets in the cross-validation operation. The average ROC and PR curves of AGLDA and the comparison models are shown in Figure 4. AGLDA and GAIRD produced the best AUC of 0.988, which are 0.1% higher than MGLDA, 0.5% more than GSMV, and GTAN, 3.2% higher than iLncDA-LTR, 4.1% more than VADLP, and 5.6% more than CNNLDA. Meanwhile, AGLDA acquires the best AUPR of 0.684, which is 3.5, 9.5, 16.3, 23.1, 29.3, 32, and 40.1% higher than GSMV, MGLDA, GTAN, iLncDA-LTR, VADLP, and CNNLDA, respectively. Therefore, our method achieves a significant improvement in ARPR compared with that of other methods.

For this evaluation, each prediction method obtained 405 AUC and AUPR results for 405 diseases. When AGLDA is compared to another method, a Paired Wilcoxon test is calculated through the 405 pairs of AUC and AUPR values. The results of statistical testing (p-value <0.05) indicate that the performance of AGLDA significantly exceeds that of other models, as seen in Table 3.

CNNLDA is a model based on convolutional neural networks that demonstrate poor performance. Compared with this model, the performances of VADLP and GTAN

Table 3. Paired Wilcoxon Test Results for AGLDA and Each Compared Method

	p-value of AUC	p-value of AUPR
GAIRD	9.3465×10^{-7}	1.7015×10^{-4}
GSMV	6.9191×10^{-7}	2.1133×10^{-10}
MGLDA	9.6223×10^{-7}	4.3750×10^{-10}
GTAN	1.2942×10^{-6}	1.4860×10^{-8}
iLncDA-LTR	1.3863×10^{-6}	1.3855×10^{-4}
VADLP	1.7271×10^{-6}	2.1963×10^{-4}
CNNLDA	2.1799×10^{-6}	1.1057×10^{-4}

are improved because they leverage a variational autoencoder and graph neural network, respectively. CNNLDA and VADLP have AUCs and AUPRs lower than those of the other methods. It is probably because these two methods learned the attributes only from a pair of lncRNA and disease nodes, while they ignored capture of the topology from the heterogeneous graph composed of the lncRNA, disease, and miRNA nodes. GSMV and MGLDA establish meta paths with semantics and predictive models for graph inference, but they are limited by the number of meta paths. GAIRD constructs a graph convolution and random walk model to refine the attributes and topology of high-order node neighbors, achieving the second-highest performances. The improvement of our method over GAIRD is likely due to the extraction and dynamic changes of the lncRNA−disease biological characteristics from a hypergraph perspective.

The top k recall rates of the candidate lncRNA disease associations for AGLDA and other methods are presented in Figure 5. The AGLDA method achieves the best recall rates of 94.17% in the top 30, 96.62% in the top 60, and 97.42% in the top 90. This result indicates that our method correctly identifies more disease-related lncRNA candidates in the top rank list compared to other methods. The GAIRD achieves the second highest recall rate of 94.02% in the top 30, 96.51% in the top 60, and 97.34% in the top 90. GSMV has lower recall of rates 92.37, 94.16, and 96.74% compared to that of GAIRD, when k is 30, 60, and 90, respectively. The recall rates of MGLDA are higher than that of GTAN in the top 30, top 60, and top 90. The former receives 91.19, 93.34, and 96.64%,

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

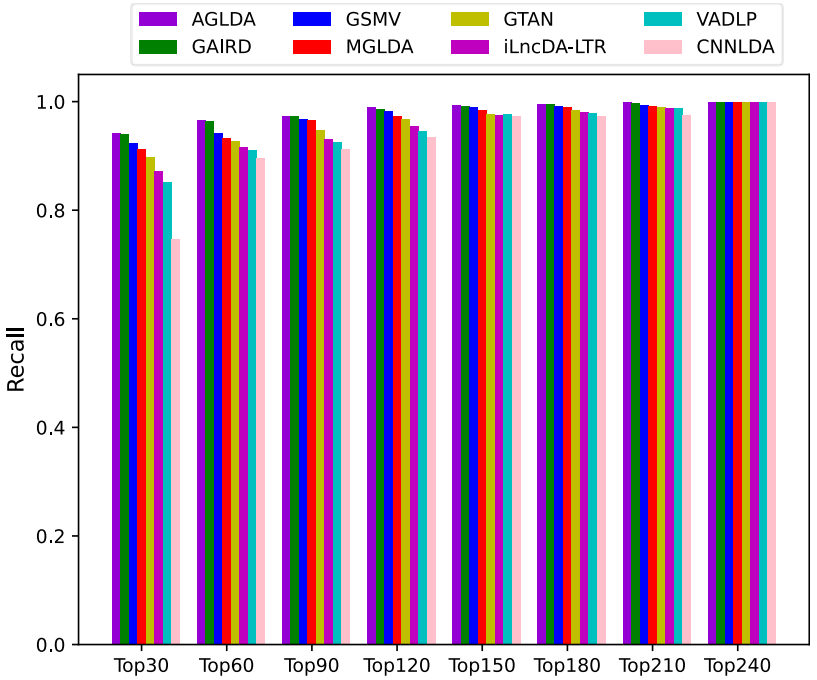


Figure 5. Recall rates of all the methods at different k values.

Table 4. Top 15 Prostate Cancer-Related lncRNA Candidates[a]

rank	lncRNA name	description	rank	lncRNA name	description
1	PCA3	L^a, L^b	9	SCHLAP1	L^a, L^b
2	GAS5	L^a, L^b	10	C1QTNF9B-AS1	L^a
3	IGF2-AS	L^a	11	TTTY15	L^a
4	CDKN2B-AS1	L^a	12	PCAT1	L^a, L^b
5	H19	L^a	13	NEAT1	L^a
6	PVT1	L^a, L^b	14	HOTTIP	L^a, L^b
7	CTBP1-AS	L^a, L^b	15	PCGEM1	L^a, L^b
8	MALAT1	L^a, L^b			

[a]L^a: lnc2Cancer; L^b: lncRNADisease.

while the latter receives rates of 89.72, 92.81, and 94.83%, respectively. The recall rates of iLncDA-LTR in the top 30, 60, and 90 are 87.26, 91.60, and 93.06%, respectively. The corresponding recall rates of VADLP are 85.16, 91.08, and 92.47%, respectively. Finally, the recall rates of CNNLDA are not as good as those of the other methods, with recall rates of 74.64, 89.52, and 91.34%, respectively.

Case Studies. Considering case studies featuring prostate cancer, breast cancer, and esophageal cancer, we demonstrate the ability of AGLDA to identify diseases related to lncRNA candidates. All candidate lncRNAs related to each disease are sorted in descending order according to their association scores. We selected the top 15 candidates for each of three diseases as potential candidates. The top 15 lncRNA candidates for these three disease rankings are listed in Tables 4–6.

Bao et al.[42] extracted the experimentally supported lncRNA–disease associations from PubMed by searching for relevant keywords. The database Lnc2Cancer[43] provides information on the regulatory mechanisms, biological functions, and clinical applications of lncRNA in cancer. The lncRNA and disease associations included in the database were obtained from biological experiments, and any abnormal expression of lncRNA was artificially determined. Table 4 shows the top 15 lncRNAs associated with prostate cancer, 9 lncRNAs of which are contained in lncRNADisease,[42] confirming the disease association. Of this, all 15 candidates appear in Lnc2Cancer, suggesting their up- or down-regulation in the occurrence of prostate cancer. The top 15 candidates for esophageal cancer and breast cancer (Tables 5, and 6), 25 (24) candidates are reported in lncRNADiease (Lnc2Cancer). This analysis suggests that the candidates identified by AGLDA are directly associated with the corresponding disease. The label "Literarure" indicates the discovery of the lncRNA candidate

Table 5. Top 15 Breast Cancer-Related lncRNA Candidates[a]

rank	lncRNA candidate	evidence	rank	lncRNA candidate	evidence
1	AFAP1-AS1	L^a	9	CDKN2B-AS1	L^a, L^b
2	LINC00675	L^a	10	UCA1	L^a, L^b
3	HOTTIP	L^a, L^b	11	GAS5	L^a, L^b
4	H19	L^a, L^b	12	BCYRN1	L^a, L^b
5	MEG3	L^a, L^b	13	SPRY4-IT1	L^a, L^b
6	HOAIR	L^a, L^b	14	KCNQ1OT1	L^a, L^b
7	MALAT1	L^a, L^b	15	XIST	L^a, L^b
8	PVT1	L^a, L^b			

[a]L^a: lnc2Cancer; L^b: lncRNADisease.

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 6. Top 15 Esophageal Cancer-Related lncRNA Candidates[a]

rank	lncRNA candidate	evidence	rank	lncRNA candidate	evidence
1	HOTAIR	L[a], L[b]	9	MALAT1	L[a], L[b]
2	H19	L[a], L[b]	10	CCAT1	L[a]
3	SPRY4-IT1	L[b]	11	UCA1	L[a], L[b]
4	CDKN2B-AS1	L[b]	12	GAS5	L[a], L[b]
5	SOX2-OT	L[b]	13	NEAT1	L[b]
6	HNF1A-AS1	L[c]	14	CCAT2	L[a]
7	BCYRN1	L[a], L[b]	15	PVT1	L[a], L[b]
8	MEG3	L[a], L[b]			

[a]L[a]: lnc2Cancer; L[b]: lncRNADisease; L[c]: Literature.

HNF1A-AS1, which is highly up-regulated for esophageal cancer.[44] These case studies demonstrate that AGLDA is a powerful and effective way to identify potential lncRNA candidates with related diseases.

Predicting Novel lncRNA−Disease Associations. After comparing these models with advanced models and conducting case studies to evaluate the predictive ability of AGLDA, we next trained the model using all known lncRNA−disease associations to predict novel associations. We regarded all of the known lncRNA−disease associations as the positive training samples. The same number of unobserved lncRNA−disease associations as the positive training samples were selected, and they were used as the negative training samples. We obtained the top 30 high-quality disease-related lncRNAs as recorded in Supporting Information Table ST1.

CONCLUSIONS

We proposed a prediction method (AGLDA) to encode the biological characteristics of the lncRNA−disease associations and integrate the dynamic topologies of a hypergraph. The constructed multiple hyperedges are helpful for reflecting the various complex association relationships among multiple lncRNAs and diseases from multiple views. The designed hypergraph neural network-based strategy was able to dynamically update the attributes of hyperedges and generate the dynamic topologies of the hypergraph. The module based on graph convolutional networks can fuse the attributes of multiple types of nodes and their topology from the entire heterogeneous network view. A strategy based on gated convolutional neural networks was presented to assign higher weights to the more informative features of the lncRNA−disease node pair. The cross-validation experimental results demonstrated that AGLDA achieved higher AUC and AUPR than those of the seven prediction methods compared, and the ablation experiments also showed the effectiveness of its major innovations. The recall rates for the top-ranked candidates and the case studies on three diseases showed AGLDA's ability to provide reliable disease-related candidate lncRNAs.

ASSOCIATED CONTENT

Data Availability Statement

The source code and data sets are freely available at https://github.com/pingxuan-hlju/AGLDA.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00245.

Experiment results of the hyperparameter selection (PDF)

Top 30 high-quality disease-related lncRNAs (XLSX)

AUTHOR INFORMATION

Corresponding Author

Tiangang Zhang − *School of Computer Science and Technology and School of Mathematical Science, Heilongjiang University, Harbin 150080, China*; Email: zhang@hlju.edu.cn

Authors

Ping Xuan − *School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China; Department of Computer Science, Shantou University, Shantou 515063, China*; orcid.org/0000-0001-5328-691X

Siyuan Lu − *School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China*

Hui Cui − *Department of Computer Science and Information Technology, La Trobe University, Melbourne 3083, Australia*

Shuai Wang − *School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*; orcid.org/0009-0003-1562-3198

Toshiya Nakaguchi − *Center for Frontier Medical Engineering, Chiba University, Chiba 2638522, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00245

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (62172143, 62372282), STU Scientific Research Initiation Grant (NTF22032), and the Natural Science Foundation of Heilongjiang Province (LH2023F044).

REFERENCES

(1) Bridges, M. C.; Daulagala, A. C.; Kourtidis, A. LNCcation: lncRNA Localization and Function. *J. Cell Biol.* **2021**, *220*, No. e202009045.

(2) Xing, C.; Sun, S.-g.; Yue, Z.-Q.; Bai, F. Role of lncRNA LUCAT1 in Cancer. *Biomed. Pharmacother.* **2021**, *134*, 111158.

(3) Robinson, E. K.; Covarrubias, S.; Carpenter, S. The How and Why of lncRNA Function: an Innate Immune Perspective. *Biochim. Biophys. Acta, Gene Regul. Mech.* **2020**, *1863*, 194419.

(4) Liu, C.; Wu, S.; Lai, L.; Liu, J.; Guo, Z.; Ye, Z.; Chen, X. Comprehensive Analysis of Cuproptosis-Related lncRNAs in Immune Infiltration and Prognosis in Hepatocellular Carcinoma. *BMC Bioinf.* **2023**, *24*, 4.

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

(6) Chen, X.; Yan, C. C.; Zhang, X.; You, Z.-H. Long Non-Coding RNAs and Complex Diseases: From Experimental Results to Computational Models. *Briefings Bioinf.* **2017**, *18*, 558−576.

(7) Clark, M. B.; Johnston, R. L.; Inostroza-Ponta, M.; Fox, A. H.; Fortini, E.; Moscato, P.; Dinger, M. E.; Mattick, J. S. Genome-Wide Analysis of Long Noncoding RNA Stability. *Genome Res.* **2012**, *22*, 885−898.

(8) Li, J.; Gao, C.; Wang, Y.; Ma, W.; Tu, J.; Wang, J.; Chen, Z.; Kong, W.; Cui, Q. A Bioinformatics Method for Predicting Long Noncoding RNAs associated with vascular disease. *Sci. China: Life Sci.* **2014**, *57*, 852−857.

(9) Lin, X.-C.; Zhu, Y.; Chen, W.-B.; Lin, L.-W.; Chen, D.-H.; Huang, J.-R.; Pan, K.; Lin, Y.; Wu, B.-T.; Dai, Y.; et al. Integrated analysis of long non-coding RNAs and mRNA expression profiles

reveals the potential role of lncRNAs in gastric cancer pathogenesis. *Int. J. Oncol.* **2014**, *45*, 619−628.
(10) Biswas, A. K.; Zhang, B.; Wu, X.; Gao, J. X. A Multi-Label Classification Framework to Predict Disease Associations of Long Non-Coding RNAs (lncRNAs). *Proceedings of the Third International Conference on Communications, Signal Processing, and Systems*; Springer, 2015; Vol. *322*, pp 821−830..
(11) Chen, X.; Yan, G.-Y. Novel Human lncRNA−Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* **2013**, *29*, 2617−2624.
(12) Zhao, T.; Xu, J.; Liu, L.; Bai, J.; Xu, C.; Xiao, Y.; Li, X.; Zhang, L. Identification of Cancer-Related lncRNAs Through Integrating Genome, Regulome and Transcriptome Features. *Mol. BioSyst.* **2015**, *11*, 126−136.
(13) Xie, G.; Jiang, J.; Sun, Y. LDA-LNSUBRW: lncRNA-Disease Association Prediction Based on Linear Neighborhood Similarity and Unbalanced Bi-Random Walk. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2020**, *19*, 989−997.
(14) Wang, L.; Shang, M.; Dai, Q.; He, P.-a. Prediction of lncRNA-Disease Association Based on a Laplace Normalized Random Walk with Restart Algorithm on Heterogeneous Networks. *BMC Bioinf.* **2022**, *23*, 5.
(15) Lan, W.; Li, M.; Zhao, K.; Liu, J.; Wu, F.-X.; Pan, Y.; Wang, J. LDAP: a Web Server for lncRNA-Disease Association Prediction. *Bioinformatics* **2017**, *33*, 458−460.
(16) Zhao, J.; Cheng, W.; He, X.; Liu, Y.; Li, J.; Sun, J.; Li, J.; Wang, F.; Gao, Y. Construction of a Specific SVM Classifier and Identification of Molecular Markers for Lung Adenocarcinoma Based on lncRNA-miRNA-mRNA Network. *OncoTargets Ther.* **2018**, *11*, 3129−3140.
(17) Wang, M.-N.; You, Z.-H.; Wang, L.; Li, L.-P.; Zheng, K. LDGRNMF: LncRNA-Disease Associations Prediction Based on Graph Regularized Non-Negative Matrix Factorization. *Neurocomputing* **2021**, *424*, 236−245.
(18) Liu, J.-X.; Gao, M.-M.; Cui, Z.; Gao, Y.-L.; Li, F. DSCMF: Prediction of LncRNA-Disease Associations Based on Dual Sparse Collaborative Matrix Factorization. *BMC Bioinf.* **2021**, *22*, 241.
(19) Lu, C.; Yang, M.; Li, M.; Li, Y.; Wu, F.-X.; Wang, J. Predicting Human lncRNA-Disease Associations Based on Geometric Matrix Completion. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2420−2429.
(20) Wu, X.; Lan, W.; Chen, Q.; Dong, Y.; Liu, J.; Peng, W. Inferring LncRNA-Disease Associations Based on Graph Autoencoder Matrix Completion. *Comput. Biol. Chem.* **2020**, *87*, 107282.
(21) Yao, D.; Zhan, X.; Zhan, X.; Kwoh, C. K.; Li, P.; Wang, J. A Random Forest Based Computational Model for Predicting Novel lncRNA-Disease Associations. *BMC Bioinf.* **2020**, *21*, 126.
(22) Wu, Q.-W.; Xia, J.-F.; Ni, J.-C.; Zheng, C.-H. GAERF: Predicting lncRNA-Disease Associations by Graph Auto-Encoder and Random Forest. *Briefings Bioinf.* **2021**, *22*, bbaa391.
(23) Chen, Y.; Wang, J.; Wang, C.; Liu, M.; Zou, Q. Deep Learning Models for Disease-associated circRNA Prediction: a Review. *Briefings Bioinf.* **2022**, *23*, bbac364.
(24) Niu, M.; Wang, C.; Zhang, Z.; Zou, Q. A Computational Model of circRNA-Associated Diseases Based on a Graph Neural Network: Prediction and Case Studies for Follow-Up Experimental Validation. *BMC Biol.* **2024**, *22*, 24.
(25) Feng, H.; Jin, D.; Li, J.; Li, Y.; Zou, Q.; Liu, T. Matrix Reconstruction with Reliable Neighbors for Predicting Potential MiRNA−Disease Associations. *Briefings Bioinf.* **2023**, *24*, bbac571.
(26) Wang, Y.; Zhang, X.; Ju, Y.; Liu, Q.; Zou, Q.; Zhang, Y.; Ding, Y.; Zhang, Y. Identification of Human microRNA-Disease Association Via Low-Rank Approximation-Based Link Propagation and Multiple Kernel Learning. *Front. Comput. Sci.* **2024**, *18*, 182903.
(27) Wang, Z.; Gu, Y.; Zheng, S.; Yang, L.; Li, J. MGREL: A Multi-Graph Representation Learning-Based Ensemble Learning Method for Gene-Disease Association Prediction. *Comput. Biol. Med.* **2023**, *155*, 106642.
(28) Mastropietro, A.; De Carlo, G.; Anagnostopoulos, A. XGDAG: Explainable Gene−Disease Associations Via Graph Neural Networks. *Bioinformatics* **2023**, *39*, btad482.
(29) Xuan, P.; Cao, Y.; Zhang, T.; Kong, R.; Zhang, Z. Dual Convolutional Neural Networks with Attention Mechanisms Based Method for Predicting Disease-related lncRNA Genes. *Front. Genet.* **2019**, *10*, 416.
(30) Sheng, N.; Cui, H.; Zhang, T.; Xuan, P. Attentional Multi-Level Representation Encoding Based on Convolutional and Variance Autoencoders for lncRNA−Disease Association Prediction. *Briefings Bioinf.* **2021**, *22*, bbaa067.
(31) Xuan, P.; Zhan, L.; Cui, H.; Zhang, T.; Nakaguchi, T.; Zhang, W. Graph Triple-Attention Network for Disease-Related lncRNA Prediction. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2839−2849.
(32) Xuan, P.; Zhao, Y.; Cui, H.; Zhan, L.; Jin, Q.; Zhang, T.; Nakaguchi, T. Semantic Meta-Path Enhanced Global and Local Topology Learning for lncRNA-Disease Association Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2023**, *20*, 1480−1491.
(33) Xuan, P.; Wang, S.; Cui, H.; Zhao, Y.; Zhang, T.; Wu, P. Learning Global Dependencies and Multi-Semantics within Heterogeneous Graph for Predicting Disease-Related lncRNAs. *Briefings Bioinf.* **2022**, *23*, bbac361.
(34) Wang, S.; Hui, C.; Zhang, T.; Wu, P.; Nakaguchi, T.; Xuan, P. Graph Reasoning Method Based on Affinity Identification and Representation Decoupling for Predicting lncRNA-Disease Associations. *J. Chem. Inf. Model.* **2023**, *63*, 6947−6958.
(35) Xuan, P.; Bai, H.; Cui, H.; Zhang, X.; Nakaguchi, T.; Zhang, T. Specific Topology and Topological Connection Sensitivity Enhanced Graph Learning for lncRNA−Disease Association Prediction. *Comput. Biol. Med.* **2023**, *164*, 107265.
(36) Fu, G.; Wang, J.; Domeniconi, C.; Yu, G. Matrix Factorization-Based Data Fusion for the Prediction of lncRNA−Disease Associations. *Bioinformatics* **2018**, *34*, 1529−1537.
(37) 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*, 1644−1650.
(38) Chen, X.; Clarence Yan, C.; Luo, C.; Ji, W.; Zhang, Y.; Dai, Q. Constructing lncRNA Functional Similarity Network Based on lncRNA-Disease Associations and Disease Semantic Similarity. *Sci. Rep.* **2015**, *5*, 11338.
(39) Huang, J.; Ling, C. X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299−310.
(40) Saito, T.; Rehmsmeier, M. The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One* **2015**, *10*, No. e0118432.
(41) Wu, H.; Liang, Q.; Zhang, W.; Zou, Q.; El-Latif Hesham, A.; Liu, B. iLncDA-LTR: Identification of lncRNA-Disease Associations by Learning to Rank. *Comput. Biol. Med.* **2022**, *146*, 105605.
(42) Bao, Z.; Yang, Z.; Huang, Z.; Zhou, Y.; Cui, Q.; Dong, D. LncRNADisease 2.0: an Updated Database of Long Non-Coding RNA-Associated Diseases. *Nucleic Acids Res.* **2019**, *47*, D1034−D1037.
(43) Gao, Y.; Shang, S.; Guo, S.; Li, X.; Zhou, H.; Liu, H.; Sun, Y.; Wang, J.; Wang, P.; Zhi, H.; Li, X.; 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*, D1251−D1258.
(44) Hou, X.; Wen, J.; Ren, Z.; Zhang, G. Non-Coding RNAs: New Biomarkers and Therapeutic Targets for Esophageal Cancer. *Oncotarget* **2017**, *8*, 43571−43578.

https://doi.org/10.1021/acs.jcim.4c00245
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX