OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–15
https://doi.org/10.1093/bib/bbab604
Problem Solving Protocol

Multi-channel graph attention autoencoders for disease-related lncRNAs prediction

Nan Sheng, Lan Huang, Yan Wang, Jing Zhao, Ping Xuan, Ling Gao and Yangkun Cao

Corresponding authors: Lan Huang, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: huanglan@jlu.edu.cn; Yan Wang, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: wy6868@jlu.edu.cn

Abstract

Motivation: Predicting disease-related long non-coding RNAs (lncRNAs) can be used as the biomarkers for disease diagnosis and treatment. The development of effective computational prediction approaches to predict lncRNA-disease associations (LDAs) can provide insights into the pathogenesis of complex human diseases and reduce experimental costs. However, few of the existing methods use microRNA (miRNA) information and consider the complex relationship between inter-graph and intra-graph in complex-graph for assisting prediction.
Results: In this paper, the relationships between the same types of nodes and different types of nodes in complex-graph are introduced. We propose a multi-channel graph attention autoencoder model to predict LDAs, called MGATE. First, an lncRNA-miRNA-disease complex-graph is established based on the similarity and correlation among lncRNA, miRNA and diseases to integrate the complex association among them. Secondly, in order to fully extract the comprehensive information of the nodes, we use graph autoencoder networks to learn multiple representations from complex-graph, inter-graph and intra-graph. Thirdly, a graph-level attention mechanism integration module is adopted to adaptively merge the three representations, and a combined training strategy is performed to optimize the whole model to ensure the complementary and consistency among the multi-graph embedding representations. Finally, multiple classifiers are explored, and Random Forest is used to predict the association score between lncRNA and disease. Experimental results on the public dataset show that the area under receiver operating characteristic curve and area under precision-recall curve of MGATE are 0.964 and 0.413, respectively. MGATE performance significantly outperformed seven state-of-the-art methods. Furthermore, the case studies of three cancers further demonstrate the ability of MGATE to identify potential disease-correlated candidate lncRNAs. The source code and supplementary data are available at https://github.com/sheng-n/MGATE.

Keywords: disease-related lncRNAs prediction, graph autoencoders, graph-level attention mechanism, random forest

Introduction

With the advancement of RNA sequencing technology, the complexity of the genome has been further revealed in recent years. Studies have shown that 1–2% of protein-coding genes in the human genome are stably transcribed; the rest of the RNA has no protein-coding function [1, 2]. Those genes are named non-coding RNA (ncRNA). Especially, long non-coding RNAs (lncRNAs) are a type of ncRNA with a length of more than 200 nucleotides. They have played an important role in various life activities and w idely participated in a host of physiological and pathological processes [3–5]. Many medical experiments have proved that lncRNA is widely associated with various diseases, including cancers [6], nervous system diseases [7] and cardiovascular diseases [8]. For example, lncRNA CTA-929C8 is overexpressed in brain tissue, about 1000 times higher than other normal tissues [9]. Therefore, this lncRNA can be used as a potential diagnostic marker for Alzheimer's disease.

Many biological experiments have been designed to accurately predict disease-related lncRNAs, but they also have limitations such as being time-consuming,

Nan Sheng is a PhD candidate in the College of Computer Science and Technology in the Jilin University. His research interests include bioinformatics and deep learning.
Lan Huang is a professor at the College of Computer Science and Technology in the Jilin University. Her research primarily focuses on bioinformatics, data mining and machine learning.
Yan Wang is a professor at the College of Computer Science and Technology in the Jilin University. His research primarily focuses on bioinformatics, data mining and machine learning.
Jing Zhao is a clinical assistant professor in the department of biomedical informatics at the Ohio State University College of Medicine. Her research interests lie in statistical genomics and predictive modeling for disease progression.
Ping Xuan is a professor at the School of Computer Science and Technology in the Heilongjiang University. Her current research interests include computational biology, complex network analysis and deep learning.
Ling Gao is studying for his master's degree in the School of Computer Science and Technology at Heilongjiang University. His research interests include complex network analysis and deep learning.
Yangkun Cao is a PhD candidate in the School of Artificial Intelligence in the Jilin University. His research interests include computational biology and deep learning.
Received: October 5, 2021. **Revised:** December 8, 2021. **Accepted:** December 27, 2021

expensive and prone to failure [10]. Therefore, based on existing experimental data, developed computational methods to predict lncRNA-disease associations (LDAs) can improve the limitations. A computational model, named LRLSLDA, was first proposed by Chen et al. to predict the LDAs [11]. The model integrated lncRNA expression profiles and LDA data and used Laplacian regularized least squares approach to predict disease-related lncRNAs. Many computational methods were designed to predict the LDAs, and biologists had further designed experiments to verify them and discover true associations in recent years, most of which can be grouped into four main categories. The first category of approaches is based on information flow propagation. Sun et al. constructed a heterogeneous network using the similarity network of lncRNA and disease, and known LDAs, and a random walk with restart (RWR) is employed to predict LDAs scores [12]. With the accumulation of biological data, there is a need to fuse lncRNA-associated and disease-associated multi-source data information to infer LDAs. Each data source provides a complex mechanism for disease and the interaction of different biomolecules. Therefore, Yu et al. integrated lncRNA–miRNA interaction, miRNA–disease association and lncRNA–gene association and used the bi-random walk model to identify the LDAs probability of between lncRNA and disease [13].

The second category of methods is leveraged non-negative matrix factorization. Fu et al. used 11 relational data sources to decompose the multiple data matrices into low-rank matrices through matrix triple-factorization and assigned different weights to them to select and integrate data sources [14]. Lu et al. regarded the LDAs problem as a recommendation system problem, and then used the induction matrix method to predict LDAs [15]. Xuan et al. built a heterogeneous graph by introducing miRNA–disease association and lncRNA–miRNA interaction, and inferred the potential LDA based on probability matrix factorization [16]. Several matrix factorization approaches have been improved to identify disease-linked lncRNA [17–19].

The third category of approach utilized machine learning. Lan et al. exploited a Web server to calculate two lncRNA similarities and five disease similarities, and support vector machine is employed as a classifier to identify LDAs [20]. Yu et al. first constructed LMD triple-layer heterogeneous network and introduced a project-based collaborative filtering algorithm to update the LMD triple-layer heterogeneous network; naive Bayesian classifier is used to discover disease-related lncRNAs [21]. Zhou et al. used lncRNA, protein, drug, disease, miRNA and other data to construct a heterogeneous graph, and utilized high-order proximity reserve embedding to fuse the node itself feature and the behavior feature. Then, the random forest (RForest) was used as a classifier to predict potential LDAs [22].

Deep learning has great success in bioinformatics in recent years. The fourth category of methods is mainly deep learning. Xuan et al. proposed several classical models, such as CNNDLP [23] and CNNLDA [24], using convolutional autoencoder (CAE) and convolutional neural network (CNN). A framework based on the bidirectional generative adversarial network was proposed by Yang et al., which uses encoders and generators to learn lncRNA and disease features in potential space and uses discriminators as classifiers to identify undiscovered relations between lncRNA and disease [25]. Zeng et al. used singular value decomposition and neural networks to capture linear and nonlinear features of lncRNA and disease, respectively, to predict LDAs [26]. Some methods combine machine learning with deep learning. For example, Lan et al. combined autoencoders and matrix factorization algorithms [27], and Sheng et al. integrated random walks and convolutional and variance autoencoders [28].

The generation of unstructured data has accelerated the emergence of the graph data model, and graph neural network has become one of the most popular topics in scientific research. Currently, graph neural networks have been widely utilized in computational chemistry and computational biology. Xuan et al. combined graph CAEs and CNN to predict disease-related candidate lncRNAs [29]. Shi et al. developed an end-to-end model based on variational graph autoencoder (VGAE) and graph autoencoder (GAE) to extract low-dimensional representation from high-dimensional features for predicting LDAs [30]. Two methods based on graph convolutional networks (GCNs) are proposed: GAMCLDA [31] and GAERF [32]. GAMCLDA constructs a computing model based on matrix completion of GAE, which uses GCNs to obtain local topology structures and latent vector of lncRNA and diseases. Finally, the LDA matrix was reconstructed using the inner product of lncRNA and disease features as the decoder. GAERF thinks that the structure of the previous deep learning model is too complex, and a lot of model parameters have to be tuned. Therefore, a prediction model based on GAEs and the RForest is proposed. Although the above graph neural network approaches have achieved relatively good performance for predicting LDAs, they still have room for improvement. In particular, these two graph models can exploit the specific relationships of the lncRNA and disease nodes in the inter-graph and intra-graph to improve prediction performance. In addition, there are all based on graph neural networks, which encode the constructed heterogeneous networks to extract the topological structure information of lncRNA and diseases. However, these methods do not consider the node information contained in the LMD complex-graph, inter-graph and intra-graph. For complex-graph, the relationship between nodes is regarded as equally important to capture the feature lncRNA and disease nodes. But the inter-graph contains the interaction and correlation relationship between different types of nodes, and the intra-graph contains the similarity relationship between the same type of nodes. Moreover, the complementarity of these relations cannot

be ignored, and their importance is different for the final node feature representation.

To solve the above drawbacks and limitations, this work proposes a multi-channel graph attention autoencoder learning-based framework, named MGATE, to capture lncRNA and disease information contained in LMD complex-graph, inter-graph and intra-graph. The main contributions of our model are summarized as follows:

- A triple-layer complex-graph was constructed, which consisted of weighted inter-layer edges and intra-layer edges to represented and learned information from multi-data sources, involving the similarity and correlation among lncRNA, disease and miRNA.
- We proposed three representation learning and encoding modules including complex-graph learning module, inter-graph learning module and intra-graph learning module. The GAE is utilized in the complex-graph module to achieve the comprehensive relationship between lncRNA and disease in complex-graph. We introduced the inter-graph learning module and intra-graph learning module, which are modeled by GAE, to learn the subtle connections between lncRNA, miRNA and disease nodes.
- A graph-level attentional integration strategy and combined training method are proposed to adaptively integrate the node representations extracted from complex-graph, inter-graph and intra-graph, and to ensure the consistency of the node representation learned by the three modules. Furthermore, multiple classifiers are used to explore the impact on prediction performance. Finally, we used RForest as a classifier to predict LDAs.
- We evaluated and verified the prediction performance of our model by designed ablation studies, compared with the recently published methods, top-k recall, case studies of three diseases and effects of different parameters. The evaluation results proved that MGATE has superior performance for LDAs prediction.

Materials and methods

In this paper, a novel graph representation learning approach based on a multi-channel graph attention autoencoder is proposed, MGATE, to identify disease-related lncRNAs. The framework of MGATE is depicted in Figure 1. The MGATE consists of the following parts. First, we built a triple-layer complex graph based on the similarity and correlation among lncRNA, miRNA and disease (Figure 1A). Secondly, GAE is utilized to the encoder for complex-graph, intra-graph and inter-graph, respectively, to learn comprehensive and subtle information about lncRNA, miRNA and disease (Figure 1B). Thirdly, an integration strategy based on an attention mechanism is applied, and the whole module is optimized by combined training (Figure 1C). Finally, the learned low-dimensional representation is inputted RForest classification algorithm to calculate the association score of lncRNA and disease pairwise (Figure 1D).

Dataset

The datasets used are mainly obtained from previous work [14]. This dataset documented 240 lncRNAs, 495 miRNAs and 405 diseases, including 2687 experimentally validated associations from the LncRNADisease [33] and Lnc2Cancer databases [34]. The HMDD database recoded 13 559 miRNA-disease associations [35]. The starBase database contains 1002 lncRNA–miRNA interactions between 240 lncRNAs and 405 diseases [36].

Construction and representation of complex-graph, inter-graph and intra-graph

In this section, we will focus on how to construct a triple-layer complex-graph. This complex graph mainly consists of two parts: inter-graph and intra-graph. We combined six types of networks: LDA network, lncRNA–miRNA interaction network, miRNA–disease association network, lncRNA–lncRNA similarity network, disease–disease similarity network and miRNA–miRNA similarity network. Two types of edges are mainly included the complex-graph. One is the inter-layer edge, which is primarily composed of original association or interaction; the other is the intra-layer edge, which mainly consists of the similarities between the same nodes.

Specifically, a weighted LMD complex-graph $G_{\text{complex}} = (V, E)$ was constructed based on lncRNA, miRNA and disease layers. V is a node set composed of the set consisting of lncRNA, miRNA and disease nodes, represented as $V = \{V_{\text{lncRNA}} \cup V_{\text{miRNA}} \cup V_{\text{disease}}\}$. Edge $(v_i, v_j) \in E$ is the set of edges between nodes V, mainly composed of inter-layer edges and intra-layer edges. We define the weight matrix $M_{\text{complex}} = (A, S)$ to represent the adjacency matrix of the complex-graph G_{complex}. A is the inter-association matrix for inter-edges and S is the intra-similarity matrix for intra-edges. Inter-association matrix represents the relationship between different types of nodes, that is, the original association or interaction. Intra-similarity matrix devotes the relationship between nodes of the same type, that is, similarity relationship.

Representation of inter-association matrix

Inter-association matrix A is mainly composed of LDA matrix, lncRNA–miRNA interaction matrix and miRNA–disease association matrix, which is defined as follows:

$$A = \begin{cases} A^{\text{lncRNA-disease}} \in \mathbb{R}^{N_l \times N_d} \\ A^{\text{lncRNA-miRNA}} \in \mathbb{R}^{N_l \times N_m} \\ A^{\text{miRNA-disease}} \in \mathbb{R}^{N_m \times N_d} \end{cases}, \quad (1)$$

where N_l, N_m and N_d represent the number of lncRNA, miRNA and disease, respectively. $A^{\text{lncRNA-disease}}$ matrix recorded the relationship between N_l lncRNAs and N_d diseases. If an lncRNA v_i has an experimentally verified relationship with disease v_j, then $A_{ij}^{\text{lncRNA-disease}} = 1$, otherwise it is 0. Similarly, $A^{\text{miRNA-disease}}$ matrix is

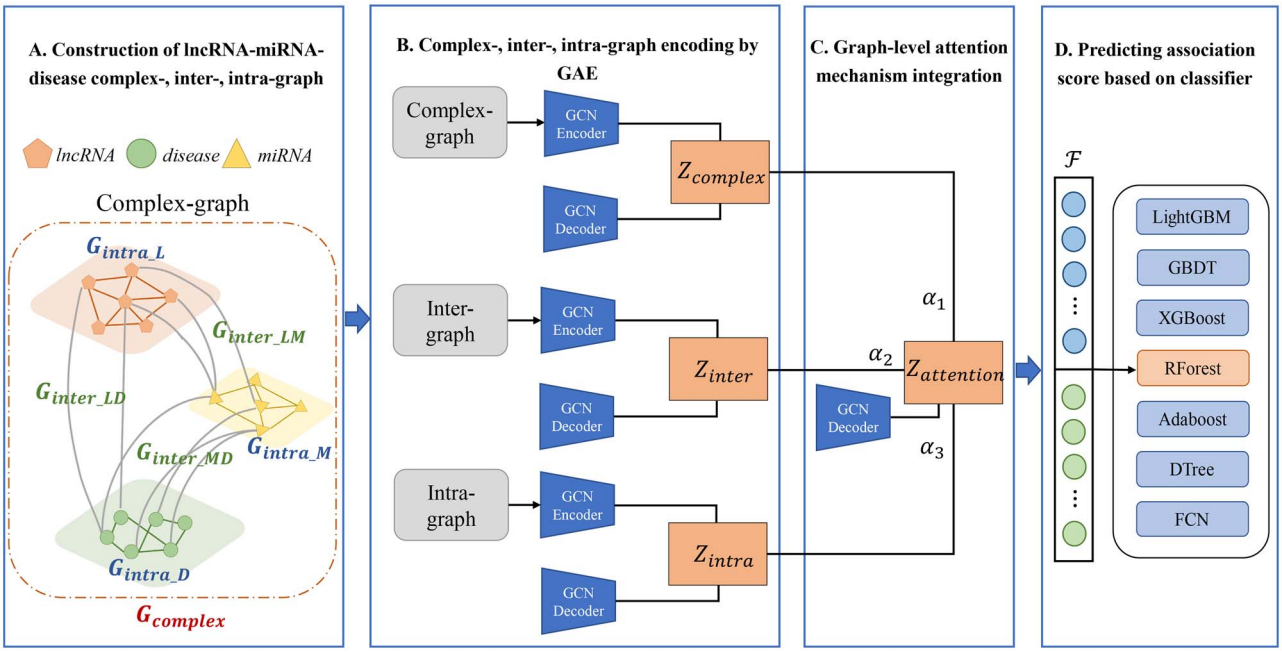

Figure 1. Overview of our proposed MGATE for predicting lncRNA-disease association. **(A)** Construction of complex-graph, inter-graph and intra-graph using the similarity and correlation among lncRNA, disease and miRNA. **(B)** Encoding of complex-graph, inter-graph and intra-graph by GAE to extract comprehensive and subtle representations. **(C)** Adaptive integration of multi-representations based on graph-level attention mechanism. **(D)** Predicting association scores of lncRNAs and disease pairs based on classifiers.

the association between N_m miRNAs and N_d diseases. $A_{ij}^{\text{miRNA-disease}} = 1$, if an association between miRNA and disease has been observed. $A_{ij}^{\text{miRNA-disease}} = 0$ means that there is no association between miRNA and disease has been observed. More and more studies have proved that lncRNA and miRNA interact and participate in the physiological and pathological processes of diseases together. Therefore, we introduced the $A^{\text{lncRNA-miRNA}}$ interaction matrix. Similarly, it also uses 0 or 1 to store the interactions. If there is a known interaction, $A_{ij}^{\text{lncRNA-miRNA}} = 1$, and for an unknown interaction, $A_{ij}^{\text{lncRNA-miRNA}} = 0$.

Representation of intra-similarity matrix

The intra-similarity matrix S is mainly constituted of the lncRNA–lncRNA similarity matrix, miRNA–miRNA similarity matrix and disease–disease similarity matrix, and is expressed as follows:

$$S = \begin{cases} S^{\text{lncRNA-lncRNA}} \in \mathbb{R}^{N_l \times N_l} \\ S^{\text{miRNA-miRNA}} \in \mathbb{R}^{N_m \times N_m} \\ S^{\text{disease-disease}} \in \mathbb{R}^{N_d \times N_d} \end{cases}, \tag{2}$$

where $S^{\text{lncRNA-lncRNA}}$ and $S^{\text{miRNA-miRNA}}$ are, respectively, the functional similarity matrix of lncRNA and miRNA, calculated by related diseases according to the approach achieved by Chen et al. and Wang et al. [37, 38]. The values range of $S_{ij}^{\text{lncRNA-lncRNA}}$ and $S_{ij}^{\text{miRNA-miRNA}}$ is between 0 and 1. $S^{\text{disease-disease}}$ is based on the method of Wang et al. [38], using the directed acyclic graphs to calculate the sematic similarity of diseases. For $S_{ij}^{\text{disease-disease}} \in [0, 1]$, it represents the similarity between disease d_i and d_j.

Representation of triple-layer complex-graph matrix

Given inter-association matrix A and intra-similarity S matrix, the adjacency matrix $M_{\text{complex}} \in \mathbb{R}^{N_v \times N_v}$ of complex-graph G_{complex}, $N_v = N_l + N_d + N_m$, which can be defined as follows:

$$M_{\text{complex}} = \begin{bmatrix} S^{\text{lncRNA-lncRNA}} & A^{\text{lncRNA-disease}} & A^{\text{lncRNA-miRNA}} \\ A^{\text{lncRNA-disease}\ T} & S^{\text{disease-disease}} & A^{\text{miRNA-disease}\ T} \\ A^{\text{lncRNA-miRNA}\ T} & A^{\text{miRNA-disease}} & S^{\text{miRNA-miRNA}} \end{bmatrix}, \tag{3}$$

where A^T represents the transposed matrix of A. In addition, we defined the adjacency matrix M_{complex} after row-normalization as the feature matrix X_{complex}.

$$X_{\text{complex}} = \begin{bmatrix} X^{\text{lncRNA}} \\ X^{\text{disease}} \\ X^{\text{miRNA}} \end{bmatrix} \tag{4}$$

X_{complex} is a matrix of $N_v \times N_v$, where each row is a feature vector for a node in V.

Representation of inter-graph and intra-graph matrix

We can construct an inter-graph, which mainly includes three categories: (1) lncRNA–disease inter-graph, which is composed of the relationship between lncRNAs and diseases. (2) lncRNA–miRNA inter-graph, which consists of the correlation between lncRNAs and miRNAs. (3) miRNA–disease inter-graph, which consists of the association between miRNAs and disease. In particular, let

$G_{inter} = \{G_{inter\_LD}, G_{inter\_LM}, G_{inter\_MD}\}$ denote inter-graph, then its adjacency matrix M_{inter} is expressed as follows:

$$M_{inter} = \begin{cases} M_{inter\_LD} = \begin{bmatrix} S^{lncRNA\text{-}lncRNA} & A^{lncRNA\text{-}disease} \\ A^{lncRNA\text{-}disease\,T} & S^{disease\text{-}disease} \end{bmatrix} \in \mathbb{R}^{N_{ld} \times N_{ld}} \\ M_{inter\_LM} = \begin{bmatrix} S^{lncRNA\text{-}lncRNA} & A^{lncRNA\text{-}miRNA} \\ A^{lncRNA\text{-}miRNA\,T} & S^{miRNA\text{-}miRNA} \end{bmatrix} \in \mathbb{R}^{N_{lm} \times N_{lm}} \\ M_{inter\_MD} = \begin{bmatrix} S^{miRNA\text{-}miRNA} & A^{miRNA\text{-}disease} \\ A^{miRNA\text{-}disease\,T} & S^{disease\text{-}disease} \end{bmatrix} \in \mathbb{R}^{N_{md} \times N_{md}} \end{cases} \quad (5)$$

where $N_{ld} = N_l + N_d$, $N_{lm} = N_l + N_m$, $N_{md} = N_m + N_d$. Here, we represent the node feature matrix of inter-graph as follows:

$$X_{inter} = \begin{cases} X_{inter\_LD} = \begin{bmatrix} X^{lncRNA} \\ X^{disease} \end{bmatrix} \in \mathbb{R}^{N_{ld} \times N_v} \\ X_{inter\_LM} = \begin{bmatrix} X^{lncRNA} \\ X^{miRNA} \end{bmatrix} \in \mathbb{R}^{N_{lm} \times N_v} \\ X_{inter\_MD} = \begin{bmatrix} X^{miRNA} \\ X^{disease} \end{bmatrix} \in \mathbb{R}^{N_{md} \times N_v} \end{cases} \quad (6)$$

For intra-graph $G_{intra} = \{G_{intra\_L}, G_{intra\_M}, G_{intra\_D}\}$, it consists mainly of edges of the same type nodes. Its adjacency matrix M_{intra} is defined in equation 7.

$$M_{intra} = \begin{cases} M_{intra\_L} = S^{lncRNA\text{-}lncRNA} \in \mathbb{R}^{N_l \times N_l} \\ M_{intra\_M} = S^{miRNA\text{-}miRNA} \in \mathbb{R}^{N_m \times N_m} \\ M_{intra\_D} = S^{disease\text{-}disease} \in \mathbb{R}^{N_d \times N_d} \end{cases} \quad (7)$$

The feature matrix of intra-graph can be represented as

$$X_{intra} = \begin{cases} X_{intra\_L} = X^{lncRNA} \in \mathbb{R}^{N_l \times N_v} \\ X_{intra\_M} = X^{miRNA} \in \mathbb{R}^{N_m \times N_v} \\ X_{intra\_D} = X^{disease} \in \mathbb{R}^{N_d \times N_v} \end{cases} \quad (8)$$

Multi-channel GAEs encoding

GAE is a multi-layer graph neural network architecture that uses end-to-end training method to learn the low-dimensional representation of nodes from graph-structured data, which can directly extract the structural information and node information of the network. It has been widely used in bioinformatics. To effectively capture comprehensive information of nodes, GAE is applied for encoding and decoding the triple-layer complex-graph. The GAE module is also utilized to extract subtle information of nodes contained in inter-graph and intra-graph to assist prediction. A graph-level attention representation integration and combined optimization are adopted to effectively assist the joint representation learning, which assures the multiple representation complementarity from the complex-, inter- and intra-graph. The overall diagram of our proposed method is shown in Figure 1.

Complex-graph learning and encoding modules by GAE

Given the adjacency matrix $M_{complex}$ and node feature matrix $X_{complex}$ of LMD complex-graph $G_{complex}$. GAE is used to encode the comprehensive information of nodes, where the encoder uses a GCN as shown by Figure 1.

Encoder. Since the diagonal values of $M_{complex}$ are the similarity values of lncRNA–lncRNA, disease–disease and miRNA–miRNA, which have been set to 1, the adjacency matrix does not have to be further processed to add self-loop. In this study, we defined the normalized adjacency matrix $\hat{M}_{complex}$ of $M_{complex}$ as $\hat{M}_{complex} = D^{(-1/2)} M_{complex} D^{(-1/2)}$, where D is the diagonal degree matrix. Specifically, given a complex-graph with adjacency matrix $\hat{M}_{complex}$, the hierarchical propagation rule formula for GCN is formulated as

$$Z^{(l+1)}_{complex} = \sigma\left(\hat{M}_{complex} Z^{(l)}_{complex} W^{(l)}_{complex}\right), \quad (9)$$

where $Z^{(l)}_{complex}$ is the node embedding representation at the l-th layer, and the initial $Z^{(0)}_{complex} = X_{complex}$. $W^{(l)}_{complex}$ is the trainable weight matrix of l-th layer in GCN and $\sigma(\cdot)$ is the nonlinear activation function *Relu*. Different layers of embedding can capture different structural information of complex-graphs. For example, the first layer collects the direct connection information between lncRNA and disease nodes, while multi-layer graph convolution can capture the multi-hop neighbor information between lncRNA and disease nodes through iterative update embedding.

Decoder. The purpose of the decoder is to reconstruct as much input as possible from the latent representation of the encoder. In order to reconstruct the LMD complex-graph, a bilinear decoder was used:

$$\hat{Z}_{complex} = \mathbf{sigmoid}\left(Z_{complex} W_{complex} Z^{T}_{complex}\right), \quad (10)$$

where $\mathbf{sigmoid}(x) = \frac{1}{1+e^{-x}}$ represents the activation function and $Z^{T}_{complex}$ is the transpose of $Z_{complex}$. $W_{complex}$ is the weight matrix of $d \times d$; d is the dimension of the final node embedding.

Objective function. To make the reconstructed matrix $\hat{Z}_{complex}$ consistent with the original input matrix $M_{complex}$, we use the mean-square error loss function to optimize.

$$L_{complex} = \left\| \hat{Z}_{complex} - M_{complex} \right\|^2 \quad (11)$$

In this paper, two-layer GCN is used to encode the information of lncRNA, miRNA and disease. The final

embedding of lncRNA, miRNA and disease nodes is represented as

$$Z_{\text{complex}} = \begin{bmatrix} Z_{\text{complex\_L}} \\ Z_{\text{complex\_D}} \\ Z_{\text{complex\_M}} \end{bmatrix}, \tag{12}$$

where $Z_{\text{complex\_L}}$ denotes the complex-representation of the lncRNA by graph convolution encoding from complex-graph, and $Z_{\text{complex\_D}}$ and $Z_{\text{complex\_M}}$ denote the complex-representation of disease and miRNA from complex-graph, respectively.

Inter-graph and intra-graph learning and encoding modules by GAE

In reality, the information in a complex-graph is not completely irrelevant to the information in inter-graph and intra-graph. The feature representation of lncRNA and disease may be related to information in a complex-graph or an inter-graph and an intra-graph, or three. Therefore, we need to derive the information contained in the complex-graph, which should also be extracted from the corresponding inter-graph and intra-graph. To this end, GAE is also used to encode inter-graph and intra-graph to extract specific and subtle representations of lncRNA, miRNA and disease.

Encoder. The adjacency matrix M_{inter} of inter-graph and the feature matrix X_{inter} can be received through the above section. Given an input inter-graph $G_{\text{inter}}(M_{\text{inter}}, X_{\text{inter}})$, GCN is utilized as an encoder to learn node embedding of the inter-graph. The l-th layer output $Z_{\text{inter}}^{(l+1)}$ can be defined as follows:

$$Z_{\text{inter}}^{(l+1)} = \sigma\left(D_{\text{inter}}^{-\frac{1}{2}} M_{\text{inter}} D_{\text{inter}}^{-\frac{1}{2}} Z_{\text{inter}}^{(l)} W_{\text{inter}}^{(l)}\right), \tag{13}$$

where $W_{\text{inter}}^{(l)}$ is the weight matrix of l-th layer in GCN. The node feature matrix X_{inter} is initialized to $Z_{\text{inter}}^{(0)}$, $\sigma(\cdot)$ is the activation function *Relu*. In particular, D_{inter} is a diagonal degree matrix of M_{inter}. We define the output embedding representation of the last layer as $Z_{\text{inter}} = \{Z_{\text{inter\_LD}}, Z_{\text{inter\_LM}}, Z_{\text{inter\_MD}}\}$. In this way, we can learn the node embedding of lncRNA, disease and miRNA in $Z_{\text{inter\_LD}}$, $Z_{\text{inter\_LM}}$ and $Z_{\text{inter\_MD}}$, which extracts the subtle representation in inter-graph, where $Z_{\text{inter\_LD}} = \begin{bmatrix} Z_{\text{inter\_L1}} \\ Z_{\text{inter\_D1}} \end{bmatrix}$, $Z_{\text{inter\_LM}} = \begin{bmatrix} Z_{\text{inter\_L2}} \\ Z_{\text{inter\_M1}} \end{bmatrix}$ and $Z_{\text{inter\_MD}} = \begin{bmatrix} Z_{\text{inter\_M2}} \\ Z_{\text{inter\_D2}} \end{bmatrix}$.

Decoder. To reconstruct the adjacency matrix of the inter-graph, we decode the encoder representation of the inter-graph.

$$\hat{Z}_{\text{inter}} = \text{sigmoid}\left(Z_{\text{inter}} W_{\text{inter}} Z_{\text{inter}}^{\text{T}}\right), \tag{14}$$

where $W_{\text{inter}} \in \mathbb{R}^{d \times d}$ is the weight matrix and $Z_{\text{inter}}^{\text{T}}$ denotes the transpose of Z_{inter}. Similarly, the loss function L_{inter} is designed to optimize the error between the reconstructed inter-graph and the original inter-graph.

$$L_{\text{inter}} = \left\| \hat{Z}_{\text{inter}} - M_{\text{inter}} \right\|^2, \tag{15}$$

where $L_{\text{inter}} = \{L_{\text{inter\_LD}}, L_{\text{inter\_LM}}, L_{\text{inter\_MD}}\}$, $L_{\text{inter\_LD}}$ is the loss of the optimized reconstructed lncRNA–disease inter-graph and the original lncRNA–disease inter-graph, $L_{\text{inter\_LM}}$ and $L_{\text{inter\_MD}}$ are the loss of the optimized lncRNA–miRNA inter-graph and miRNA–disease inter-graph, respectively. The two-output embedding $Z_{\text{inter\_L1}}$ and $Z_{\text{inter\_L2}}$ can be fused into one node embedding of lncRNA.

$$Z_{\text{inter\_L}} = \left(Z_{\text{inter\_L1}} + Z_{\text{inter\_L2}}\right)/2 \tag{16}$$

Similarly, the node embedding of disease $Z_{\text{inter\_D}}$ and the node embedding of miRNA $Z_{\text{inter\_M}}$ can be obtained in the same way. Therefore, Z_{inter} is redefined as $Z_{\text{inter}} = \{Z_{\text{inter\_L}}, Z_{\text{inter\_D}}, Z_{\text{inter\_M}}\}$.

Similarly, GAE is also applied to encode the subtle information of miRNA and disease from intra-graph. Given the intra-graph G_{intra} with adjacency matrix M_{intra} and feature matrix X_{intra}, encoder parameters and decoder parameters are trained by loss function L_{intra}. When the training is finished, the intra-graph node embedding indicates that $Z_{\text{intra}} = \{Z_{\text{intra\_L}}, Z_{\text{intra\_D}}, Z_{\text{intra\_M}}\}$ can be obtained.

Attention graph-level representation integration strategy

We not only need to extract representation in complex-graph, inter-graph and intra-graph, but also these three kinds of representation should be effectively fused. To achieve the best representation, an attentional graph-level representation integration strategy is designed to combine complex-representation in complex-graph and inter-representation in inter-graph, and intra-representation in intra-graph. The strategy adaptively integrates multi-graph knowledge by adaptive weights, and the weights are reflected by the attentional score vector.

Attention strategy. Given the embedding representation $Z = \{Z_{\text{complex}}, Z_{\text{inter}}, Z_{\text{intra}}\}$ learned from complex-graph, inter-graph and intra-graph, each graph representation is assigned different attention weight α_i to obtain the final embedding representation of lncRNA, miRNA and disease. The attention score α_i is calculated as

$$s_i = h\left(W_{att} Z_i + b_{att}\right) \tag{17}$$

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j \in 3} \exp(s_j)}, \tag{18}$$

where W_{att} and b_{att} are weight matrix and bias vector, respectively. h is a parameter vector and s_i represents the information score of the i-th embedding representation.

α is the attention score and the value range is (0,1). The important information can be screened out for fusion through the graph-level attention mechanism. The final embedding after the integration of attention enhancement is expressed as

$$Z_{\text{attention}} = \sum_{i} \alpha_i Z_i, \tag{19}$$

to more effectively integrate the embedding representations of complex-graph, inter-graph and intra-graph. We perform independent decoding operations on the node embedding representation $Z_{\text{attention}}$ to adequately integrate the node embedding representation and maintain the consistency of the embedding representation at different graph levels. Similarly, a decoder is used to reconstruct the original complex-graph adjacency matrix, which is defined as follows:

$$\hat{Z}_{\text{attention}} = \textbf{sigmoid}\left(Z_{\text{attention}} W_{\text{att\_d}} Z^{\text{T}}_{\text{attention}}\right), \tag{20}$$

where $\textbf{sigmoid}(\cdot)$ is the activation function and $Z^{\text{T}}_{\text{attention}}$ is the transpose of $Z_{\text{attention}}$. $W_{\text{att\_d}}$ is a trainable weight matrix with dimension $d \times d$. To make the reconstructed matrix $\hat{Z}_{\text{attention}}$ consistent with the original input matrix M_{complex}, the mean square error loss function is applied to optimize.

$$L_{\text{attention}} = \left\| \hat{Z}_{\text{attention}} - M_{\text{complex}} \right\|^2 \tag{21}$$

Overall objective function and combine optimization

In this work, we design the combine optimization method to optimize the whole module. This approach allows complex-representation, inter-representation and intra-representation complement each other, and also ensure the consistency in the final learning of the embedding representation. The entire representation learning framework is optimized by the following overall objective function

$$L = L_{\text{complex}} + L_{\text{inter}} + L_{\text{intra}} + L_{\text{attention}}, \tag{22}$$

where Adam algorithm is adopted for optimization. Our multi-channel graph attention autoencoders are trained by the back propagation algorithm. When the training process is completed, the learned procreative node embedding representation is denoted by equation 23.

$$Z_{\text{attention}} = \begin{bmatrix} Z_{\text{attention\_L}} \\ Z_{\text{attention\_D}} \\ Z_{\text{attention\_M}} \end{bmatrix}, \tag{23}$$

where $Z_{\text{attention\_L}}$ represents the feature matrix of lncRNA, and $Z_{\text{attention\_D}}$ and $Z_{\text{attention\_M}}$ are the feature matrix of disease and miRNA, respectively.

Prediction association scores based RForest

Multi-channel graph attention module was utilized to encode LDA complex-graph, inter-graph and intra-graph, the embedding representation $Z_{\text{attention}}$ of lncRNAs and diseases are obtained. We investigate the effect of different classifiers on the prediction performance, and the detailed study is presented in the following section. Finally, the RForest will then be applied as a supervised learning model to predict the association score of lncRNA and disease pairwise. RForest was proposed by Breiman et al., which is an ensemble learning algorithm based on Bagging [39]. It is composed of multiple decision trees, and the prediction is given by averaging the output of multiple decision trees. RForest has a high classification accuracy rate, especially in the face of noise and sparse data, and has good robustness. RForest is widely devoted to various data mining competitions and industries. Based on the above work, we can achieve lncRNA feature representation $Z_{\text{attention\_L}}$ and disease feature representation $Z_{\text{attention\_D}}$ after the fusion of complex-information, inter-information and intra-information. As shown in Figure 2, for a node pair of $l_1 - d_2$, the first row of $Z_{\text{attention\_L}}$ is the feature vector of l_1, $Z_{\text{attention\_L},1}$, and the second row of $Z_{\text{attention\_D}}$ represents the feature vector of d_2, $Z_{\text{attention\_D},2}$. $Z_{\text{attention\_L},1}$ and $Z_{\text{attention\_D},2}$ can be concatenated, and we can get the feature vector $\mathcal{F}$ of pairwise $l_1 - d_2$. $\mathcal{F}$ is used as the input of RForest to predict the final association probability of lncRNA-disease pairwise. A higher association score between the lncRNA l_1 and the disease d_2 demonstrates that l_1 is more possible to be related to d_2.

Experimental results and discussions

In this section, we first briefly introduce our experimental setup and evaluation metrics. Then we compare with seven existing methods to show the effectiveness of MGATE. In addition, we verify the importance of different module by designing ablation experiments, and analyze the performance effect of node embedding dimension and classifier selection. Finally, three common diseases are selected for case studies to further confirm the power of MGATE in predicting underlying LDAs.

Experimental setup and evaluation metrics

MGATE is implemented in Python based on the PyTorch deep learning framework. In addition, all experiments are conducted on the Nvidia GeForce GTX 2060Ti graphics card with 32GB memory. In MGATE, there are some hyperparameters, such as node embedding dimension d, graph convolution layer number l and learning rate lr. Here, we set the learning rate lr to 0.0001. Two-layer graph convolution is used to encode the latent feature vector of the node in the LMD complex-graph, inter-graph and intra-graph. The Xavier normal distribution is utilized for weight initialization in the GCN layer. The node embedding dimension is set as 256. Additionally, in the following section, the model performance effects of

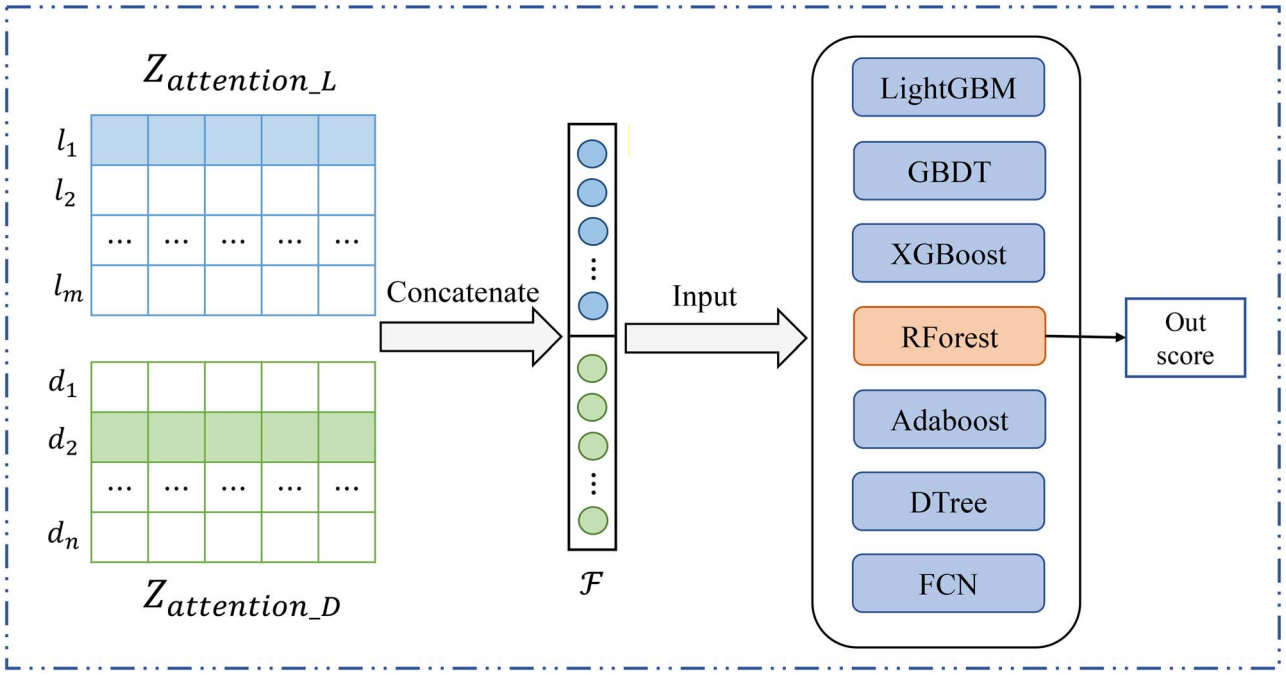

Figure 2. Constructing lncRNA-disease pairs and predicting association scores based on random forest.

different embedding dimensions d are investigated. For RForest, the grid search algorithm is adopted to find the optimal parameters. When $max\_depth = 10$ and $n\_estimators = 500$, the best prediction performance is obtained.

In our experiment, 5-fold cross-validation is adopted to evaluate the prediction performance of MGATE. The data include 2687 known LDAs and 94 513 unobserved associations. They are randomly partitioned into five equal parts, four parts are adopted for training and the remaining one part for testing. In each fold cross-validation, we randomly selected samples with the unknown LDAs whose number is equal to observed associations for training, and used the remaining unknown LDAs for testing. In each cross-validation, the lncRNA–lncRNA similarity is recalculated using the training dataset, where the known associations used for testing are removed. LDA prediction can be treated as a classification task. Several classification evaluation metrics are utilized to evaluate the prediction performance, including accuracy (Acc), area under receiver operating characteristic (ROC) curve (AUC), area under precision-recall (PR) curve (AUPR), F1-score (F1), Matthews correlation coefficient (Mcc) and recall rates under different top k values. These evaluation metrics are defined as follows:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}, \tag{1}$$

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP}, \tag{2}$$

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}, \tag{5}$$

where TP (TN) denotes true positive (negative) samples and FP (FN) denotes false positive (negative) samples. We plot ROC curves based on TPR and FPR, and PR curves based on Precision and Recall to show the performance of our proposed model. The AUC and AUPR are adopted to all-round evaluate the results of the MGATE. The recall rate of the top k is also calculated to evaluate MGATE performance.

Ablation experiments

For the MGATE, the complex-graph, inter-graph and intra-graph contain the comprehensive and subtle information of lncRNA and disease nodes. We designed a series of ablation experiments to effectively verify the importance of the complex-graph, inter-graph and intra-graph. The ablation experimental results are provided in Table 1. Without a complex-graph module, we observe that the average AUC and the average AUPR decrease by 1.3% and 12.1%, respectively. Without an inter-graph

Table 1. Results of ablation experiments on our method

Complex-graph	Inter-graph	Intra-graph	Graph-level attention	Average AUC	Average AUPR
X	✓	✓	✓	0.951	0.292
✓	X	✓	✓	0.957	0.356
✓	✓	X	✓	0.960	0.381
✓	✓	✓	X	0.961	0.385
✓	✓	✓	✓	**0.964**	**0.413**

module, the average AUC and average AUPR are 0.7% and 5.7% lower than that of our model. Without an intra-graph module, the prediction performance has decreased by 0.4% and 3.2% in AUC and AUPR. The ablation study proves the important contribution of these three modules. In addition, the ablation experiments are performed to testify the contribution of attentional mechanism fusion based on different graph levels. Without a graph-level attentional mechanism, we perform an average fusion for the three embedding representations. As shown in Table 1, compared with our model without graph-level attention, the AUC and AUPR of this model improved by 0.3% and 2.8%, respectively, which proves the contribution of the graph-level attention fusion mechanism.

The experimental results showed that integrating the comprehensive information of LMD complex-graph and the subtle information of the corresponding inter-graph and intra-graph could effectively improve the prediction performance of LMDs. In addition, the comprehensive representation obtained by the complex-graph module is more significant than the subtle representation obtained by the inter-graph and intra-graph. One possible reason is that the complex-graph contains richer node information, such as similarity relationship, association relationship and interaction relationship. The inter-graph results made the second most contribution. Compared with complex-graph, the inter-graph can directly capture the correlation and interaction between lncRNAs and diseases. Node subtle information can be extracted from the corresponding intra-graph, which is an indispensable part of our model, and which is able to help improve the prediction performance.

Comparison with seven other methods

To evaluate the superior performance of our proposed model, we compared MGATE with seven state-of-the-art approaches that were designed for LDAs prediction, which are VGAELDA [30], GAERF [32], CNNDLP [23], GCNLDA [29], Ping's method [40], SIMCLDA [15] and MFLDA [14]. These methods include information flow propagation-based methods, matrix factorization-based methods, deep learning-based methods and graph neural network-based methods. For a fair comparison, all methods were evaluated using 5-fold cross-validation, and their default parameters were used. In particular, a brief description of the comparison methods are listed in supplementary file SF1.

The ROC and PR curves of MGATE and all the comparison approaches over all the 405 diseases are shown in Figure 3. Our model of MGATE achieved the highest prediction performance with an average AUC of 0.964 and an average AUPR of 0.413, which were 0.9% and 3.9% higher than the second-best method VGAELDA based on dual GAEs, respectively. GAERF was a graph-neural-network-based approach that achieved the third-best prediction performance, with an average AUC and average AUPR lower than MGATE by 1.1% and 8.1%. As shown by Figure 3 (A), the average AUC was 1.4% higher than CNNDLP, 1.6% better than GCNLDA, 9.4% better than Ping's method, 21.9% better than SIMCLDA and 33.8% better than the poorly performing MFLDA. Figure 3 (B) shows the average AUPR over 405 diseases; MGATE reached the best AUPR of 0.413. Its average AUPR is also 14.6%, 18.8%, 26.1%, 35.4% and 37.4%, higher than CNNDLP, GCNLDA, Ping's method, SIMCLDA and MFLDA, respectively. Additionally, Table 2 also reports that MGATE outperforms all baselines in terms of accuracy, F1-score and MCC. The superior performance of MGATE mainly benefits from the fusion of comprehensive information and subtle information from complex-graph, inter-graph and intra-graph.

The results show that MGATE, VGAELDA and GAERF have the best performance and the decent performance, respectively, indicating that the method considering graph neural network can effectively predict LDAs. Nevertheless VGAELDA and GAERF failed to consider the specific subtle information in inter-graph and intra-graph. CNNDLP is based on CNN and CAE, and GCNLDA is based on CNN and GCN; their performance is similar in AUC and AUPR. One of the possible explanations is that both CNNDLP and GCNLDA used deep learning methods. Ping's method also achieved the sixth prediction performance because of the calculation and utilization of multiple lncRNA and disease similarities. However, based on the shallow model and without considering the lncRNA similarity and disease similarity, the prediction performance of SIMCLDA and MFLDA is much worse than other methods.

Figure 4 shows the recall rate under different k cutoffs. MGATE is persistently superior to the comparison methods under different k value, because of the learning of comprehensive representation and subtle representation in complex-graph, inter-graph and intra-graph. The results suggest that MGATE can predict more real LADs among the top-k candidates. When k was 30,

Table 2. The performance of MGATE against competitive approaches

Method	Acc	AUC	AUPR	F1	Mcc
MDLDA	0.724	0.626	0.039	0.192	0.152
SIMCLDA	0.781	0.745	0.059	0.275	0.239
Ping's method	0.892	0.870	0.152	0.421	0.407
GCNLDA	0.957	0.948	0.225	0.575	0.541
CNNDLP	0.969	0.950	0.267	0.628	0.609
GAERF	0.971	0.953	0.332	0.694	0.681
VGAELDA	0.975	0.955	0.374	0.731	0.713
MGATE	**0.982**	**0.964**	**0.413**	**0.782**	**0.748**

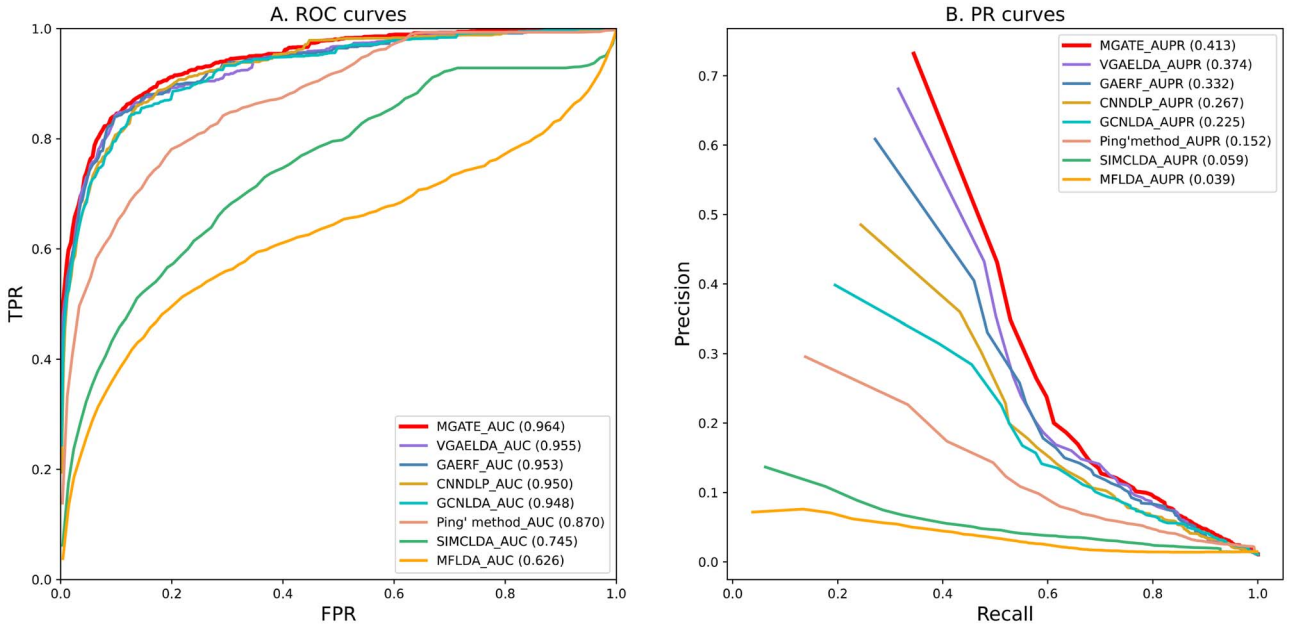

Figure 3. ROC curve and PR curve of MGATE with all comparison methods for all the 405 diseases.

MGATE achieved the highest recall rate of 89.5%. The performance of VGAELDA and GAERF was relatively close, where VGAELDA obtained the second-best recall rate of 88.7% and GAERF achieved the third-best recall rate of 88.4%. CNNDLP and GCNLDA obtained similar recall rates of 87.9% and 87.3%, respectively. When k increases from 60 to 120, the recall rate of our model was also the best performance, and the recall rate was 95.1%, 96.6% and 97.4%, respectively. VGAELDA had the second-best recall rates with 94.8%, 96.2% and 97.0%. GAERF ranked 94.3%, 96.1% and 97.0% in the top 60, 90 and 120. The recall rates of CNNDLP and GCNLDA remained close. The recall rates of CNNDLP were 93.6%, 95.8% and 96.1%, while the recall rate of GCNLDA were 93.2%, 95.1% and 96.0%. The recall rate of SIMCLDA and MFLDA was been consistently lower than that of Ping's method. When k ranges from 30 to 120, the recall rates of Ping's method are 68.9%, 81.3%, 87.5% and 92.7%.

The performance effects of different dimensions of node embedding

Node embedding dimension is the main hyperparameter of MGATE. We designed experiments to compare the effects of different embedding dimensions on MGATE performance under two layers of encoders. We choose the value of hyperparameter d, from {32, 64, 128, 256, 512}, to change the node embedding dimension. Table 3 demonstrated the performance changes of MGATE with different embedding dimensions in 5-fold cross validation. MGATE achieves the best performance when the embedding dimension is 256. But as the dimensionality increases, the performance gradually decreases. This indicates that when the dimension is high, the feature of lncRNA and disease are sparse, which leads to the introduction of redundant features and the generation of noise. As the node embedding dimension decreases, the performance also gradually decreases. The result demonstrates that when the dimension is too low, the lncRNA and disease information cannot be effectively represented, which leads to performance degradation. According to the results of many experiments, we finally selected 256 as the default node embedding dimension.

The performance effects of different classifiers

To evaluate the prediction performance of MGATE more comprehensively, we compared the classifier RForest

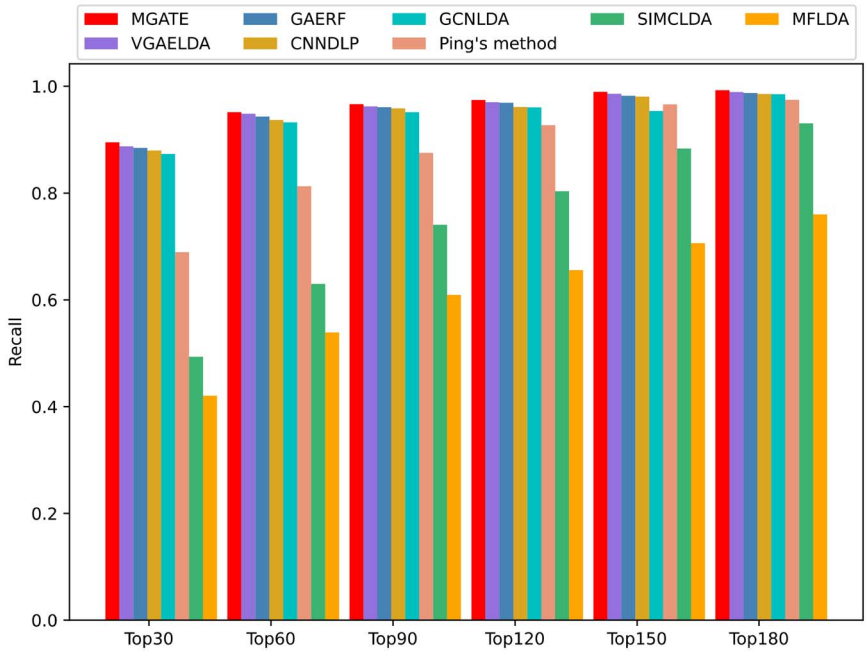

Figure 4. The recall values at different top *k*.

Table 3. The prediction performance on different embeding dimensions

	32	64	128	256	512
Average AUC	0.914	0.928	0.953	**0.964**	0.955
Average AUPR	0.309	0.320	0.363	**0.413**	0.379

Table 4. The prediction performance on different classifier

	RForest	LightGBM	GBDT	XGBoost	Adaboost	DTree	FCN
Average AUC	**0.964**	0.949	0.947	0.946	0.915	0.877	0.862
Average AUPR	**0.413**	0.327	0.310	0.271	0.288	0.162	0.296

with other classic machine learning classifiers and deep learning classifiers, such as Light Gradient Boosting Machine (LightGBM) [41], Gradient Boosting Decision Tree (GBDT) [42], Extreme Gradient Boosting (XGBoost) [43], Adaptive Boosting (Adaboost) [44], Decision Tree (DTree) [45] and fully connected neural network (FCN). According to Figure 5 and Table 4, when the RForest is used, the best performance (AUC = 0.964 and AUPR = 0.413) is achieved. One possible reason is that RFoest is an ensemble learning method based on Bagging, which can better choose and combine lncRNA and disease features for classification. Better performance is also achieved when using LightGBM, GBDT and XGBoost, which are all ensemble learning approaches based on boosting. We also find that when FCN is adopted for classification, poor performance is achieved (AUC = 0.862, AUPR = 0.296). One possible explanation is that deep learning methods cannot do a good job in classification and prediction, when there are fewer and unobvious feature vectors.

Case studies: colorectal cancer, breast cancer and prostate cancer

To further demonstrate the prediction performance of our proposed model to detect potential LDAs, we establish case studies for three cancers, containing colorectal cancer, breast cancer and prostate cancer. For each cancer, we ranked the lncRNA candidates in descending order based on their predicted LDA score.

Tables 5–7 show the top-rank 20 lncRNA candidates for each cancer. LncRNADisease database, Lnc2Cancer database and published literature are used to confirm and verify the predicted LDAs by MGATE. LncRNADisease documented information on the impact of lncRNA on human disease, which is collected from biological experiments, published literature and model predictions. But we only use the LDAs supported by biological experiments and published literature here.

First of all, as a common intestinal malignant tumor, colorectal cancer (CRC) is one of the digestive tract tumors [46]. The incidence and mortality of colorectal

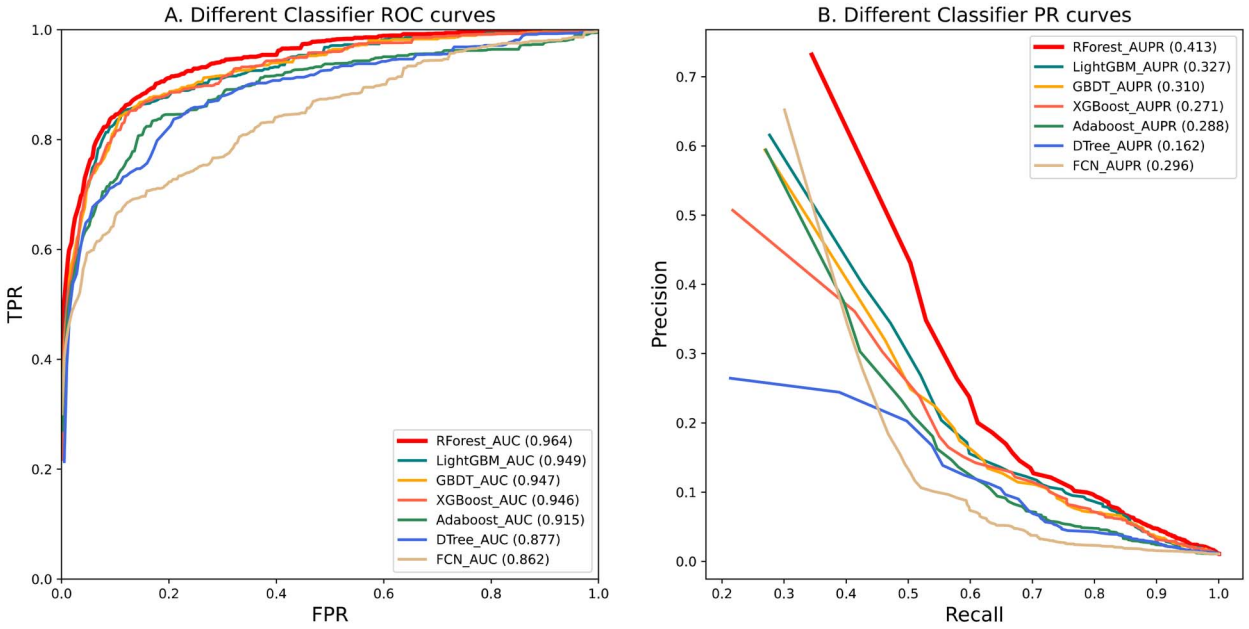

Figure 5. ROC curve and PR curve of the prediction performance on different classifiers.

Table 5. The top 20 lncRNA candidates of colorectal cancer-correlated

Rank	LncRNA name	Source of verification	Rank	LncRNA name	Source of verification
1	AFAP1-AS1	Lnc2Cancer, LncRNADisease	11	TUG1	Lnc2Cancer, LncRNADisease
2	NEAT1	Lnc2Cancer, LncRNADisease	12	CASC2	Lnc2Cancer, LncRNADisease
3	MEG3	Lnc2Cancer, LncRNADisease	13	GAS5	Lnc2Cancer, LncRNADisease
4	HOTAIR	Lnc2Cancer, LncRNADisease	14	SNHG4	literature
5	LSINCT5	Lnc2Cancer	15	CCAT2	Lnc2Cancer, LncRNADisease
6	GHET1	Lnc2Cancer, LncRNADisease	16	HOTAIRM1	Lnc2Cancer, LncRNADisease
7	BANCR	Lnc2Cancer, LncRNADisease	17	CASC16	Unconfirmed
8	KCNQ1OT1	Lnc2Cancer, LncRNADisease	18	HULC	Lnc2Cancer, LncRNADisease
9	CRNDE	Lnc2Cancer, LncRNADisease	19	PRNCR1	Lnc2Cancer, LncRNADisease
10	MALAT1	Lnc2Cancer, LncRNADisease	20	UCA1	Lnc2Cancer, LncRNADisease

Table 6. The top 20 lncRNA candidates of breast cancer-correlated

Rank	LncRNA name	Source of verification	Rank	LncRNA name	Source of verification
1	XIST	Lnc2Cancer, LncRNADisease	11	CDKN2B-AS1	LncRNADisease
2	SOX2-OT	Lnc2Cancer, LncRNADisease	12	LINC-ROR	Lnc2Cancer, LncRNADisease
3	BCYRN1	Lnc2Cancer, LncRNADisease	13	LSINCT5	Lnc2Cancer
4	SPRY4-IT1	Lnc2Cancer, LncRNADisease	14	MIR124-2HG	Literature
5	LINC00472	Lnc2Cancer, LncRNADisease	15	CCAT1	Lnc2Cancer, LncRNADisease
6	ZFAS1	Lnc2Cancer, LncRNADisease	16	AFAP1-AS1	Lnc2Cancer
7	HOTAIR	Lnc2Cancer, LncRNADisease	17	CASC16	Literature
8	LINC-PINT	Literature	18	NBAT1	Lnc2Cancer, LncRNADisease
9	MALAT1	Lnc2Cancer, LncRNADisease	19	PVT1	Lnc2Cancer, LncRNADisease
10	GAS5	Lnc2Cancer, LncRNADisease	20	EGOT	Lnc2Cancer, LncRNADisease

cancer maintain an upward trend in China. Therefore, our first case study prioritizes CRC-linked lncRNAs. As indicated in Table 5, among the 20 top-ranked CRC correlated lncRNA candidates, 18 are confirmed by the Lnc2Cancer and 17 are verified by the LncRNADisease. The results suggest that those lncRNA candidates are associated with CRC. LncRNA SNHG4 has been proven to be associated with CRC in the literature. LncRNA SNHG4 is significantly increased in colorectal cancer tissue samples and cell lines, and it regulates the colorectal cancer cell cycle and cell proliferation by modulating the miR-590-3p/CDK1 axis [47]. The lncRNA CASC16 is not associated with colorectal cancer by the literatures and databases. The potential relationship between CASC16 and human cancer colorectal cancer is still unstudied. But it has been demonstrated to be involved in the

Table 7. The top 20 lncRNA candidates of prostate cancer-correlated

Rank	LncRNA name	Source of verification	Rank	LncRNA name	Source of verification
1	KCNQ1OT1	LncRNADisease	11	EGOT	Lnc2Cancer, LncRNADisease
2	MEG3	Lnc2Cancer, LncRNADisease	12	PVT1	Lnc2Cancer, LncRNADisease
3	H19	Lnc2Cancer, LncRNADisease	13	NEAT1	Lnc2Cancer, LncRNADisease
4	PCGEM1	Lnc2Cancer, LncRNADisease	14	HOTTIP	Lnc2Cancer, LncRNADisease
5	HOTAIR	Lnc2Cancer, LncRNADisease	15	PRINS	Lnc2Cancer, LncRNADisease
6	PCAT5	Lnc2Cancer, LncRNADisease	16	PCA3	Lnc2Cancer, LncRNADisease
7	CDKN2B-AS1	Lnc2Cancer, LncRNADisease	17	LINC00963	Lnc2Cancer, LncRNADisease
8	GAS5	Lnc2Cancer, LncRNADisease	18	UCA1	Lnc2Cancer, LncRNADisease
9	IGF2-AS	Lnc2Cancer, LncRNADisease	19	DANCR	Lnc2Cancer, LncRNADisease
10	TUG1	Lnc2Cancer, LncRNADisease	20	HULC	Lnc2Cancer, LncRNADisease

development of some cancers, such as cervical cancer and stomach cancer.

Secondly, breast cancer (BC) develops from breast tissue, and it is one of the most common malignancies in women, accounting for 30 percent of all cancers in women [48]. It is also the most common cancer in the world and has the highest incidence of all cancers. We chose BC as our second case study and listed the top 20 ranked lncRNA candidates associated with BC in Table 6. Among them, Lnc2Cancer confirmed 15 lncRNA candidates and LncRNADisease confirmed 15 lncRNA candidates, which means that they are abnormally expressed in BC tissue. The lncRNA LINC-PINT, MIR124-2HG and CASC16 are supported by the literature. lncRNA LINC-PINT is downregulated in breast cancer tissues and high expression of LINC-PINT is associated with favorable disease-free survival in breast cancer patients in the TCGA cohort [49]. The lncRNA CASC16 is supported by the latest literature, which revealed that the CASC16 gene mutation is significantly related to breast cancer susceptibility [50]. Another literature verified a candidate, lncRNA MIR124-2HG, whose reduced expression can enhance the proliferation of breast cancer cells targeting BECN1 [51].

Finally, as another high-incidence disease, prostate cancer (PC) is an epithelial malignancy of the prostate gland, and is one of the most common types of cancer in men [52]. Therefore, we consider PC for the third case study. The 20 top-ranked prostate cancer-related lncRNA candidates are recorded in Table 7. All the lncRNA candidates are confirmed by the database, 20 of them are confirmed by the LncRNADisease and 19 of them are successfully funded in the Lnc2Cancer.

In summary, the case study demonstrated that we can infer that MGATE has good capability in discovering potential disease-correlated lncRNAs. Therefore, MGATE can contribute to screen reliable lncRNA candidates.

Prediction of novel disease-related lncRNAs

Finally, MGATE is leveraged to predict disease-related lncRNA candidates. The top-ranked 30 lncRNA candidates predicted by MGATE are supplied in supplementary table ST1 for researchers to download and help biologists discover true novel disease-related lncRNAs in further experiments.

Conclusions

In this paper, we propose a prediction model based on a multi-channel graph attention autoencoder, named MGATE, for predicting new LDA. First, we constructed a triple-layer complex-graph to effectively integrate rich biological information, including the similarities and correlations between lncRNA, miRNA and disease. Then, we implemented a representation learning framework based on a multi-channel GAE, which is used to capture comprehensive information from LMD complex-graph and extract subtle information from the corresponding inter-graph and intra-graph. The graph-level attention mechanism integration and combined optimization strategies are applied to effectively fuse complex-representation, inter-representation and intra-representation. Finally, RForset is used to predict disease-related lncRNA candidates. Comparison with seven state-of-the-art methods in predicting LDA methods and ablation study showed the improved performance in AUC and AUPR for MGATE. It is worth noting that MGATE can more effectively identify true LDAs and rank them as the top candidates, which is proven by the recall rates under different top k values. Case studies on three types of cancer further demonstrate the ability of MGATE. MGATE can serve as a prioritizing tool to screen underlying lncRNA candidates and help discover true LDAs.

We have utilized some biological data to construct complex graphs, but through further integration of multiple source data, our prediction model still has space for improvement. MGATE also has some limitations, such as the efficiency being slower when applied on large-scale lncRNA and disease related graph data. How to effectively fuse multiple data sources is an interesting future work and may further improve the performance of MGATE. In addition, our model can be used as a universal framework for association prediction and has implications for other association prediction tasks, including miRNA–disease association prediction, drug–disease association prediction and drug–target interaction

prediction. In the future, we will consider more lncRNA and disease data, such as lncRNA sequence, gene information and protein information. Although GCN is a powerful method for processing graph data, it has the problem of over-smoothing. Designing a more effective method to alleviate the over-smoothing problem in deep GCN is also the focus of our future research.

Key Points

- A multi-layer complex-graph was constructed, which includes inter-graph and intra-graph, to better extract sophisticated relationships of lncRNA, disease and miRNA for association prediction.
- Three representation and encoding modules were modeled to fully learn the comprehensive and subtle embedding representation of the nodes from complex-graph, inter-graph and intra-graph.
- A graph-level attention mechanism integration module and combine optimization strategy was adopted to fuse the multi-graph node embedding representation and maintain the consistency and complementarity of the embedding representation. Multiple classifiers were investigated and finally Random Forest was used to predict the association score between lncRNA and disease.
- Experimental results in several commomly evolutated metrics indicate that MGATE outperforms seven state-of-the-art lncRNA–disease prediction models. Case studies also further demonstrate the ability of to identify disease-associated lncRNA candidates.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This research was funded by the National Natural Science Foundation of China (62072212, 62172143), Chinese Postdoctoral Science Foundation (2021M691211), the Development Project of Jilin Province of China (20200401083GX, 2020C003), Guangdong Key Project for Applied Fundamental Research (2018KZDXM076), and Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC).

References

1. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012;**482**(7385):339–46.
2. Wang Kevin C, Chang HY. Molecular Mechanisms of Long Noncoding RNAs. *Mol Cell* 2011;**43**(6):904–14.
3. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011;**12**(12):861–74.
4. Chen X, Sun Y-Z, Guan N-N, *et al.* Computational models for lncRNA function prediction and functional similarity calculation. *Brief Funct Genomics* 2018;**18**(1):58–82.
5. Tsai M-C, Manor O, Wan Y, *et al.* Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* 2010;**329**(5992):689.
6. Romano G, Veneziano D, Acunzo M, *et al.* Small non-coding RNA and cancer. *Carcinogenesis* 2017;**38**(5):485–91.
7. Briggs James A, Wolvetang Ernst J, Mattick John S, *et al.* Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron* 2015;**88**(5):861–77.
8. Lorenzen JM, Thum T. Long noncoding RNAs in kidney and cardiovascular diseases. *Nat Rev Nephrol* 2016;**12**(6):360–73.
9. Shi Y, Liu H, Yang C, *et al.* Transcriptomic Analyses for Identification and Prioritization of Genes Associated With Alzheimer's Disease in Humans. *Front Bioeng Biotechnol* 2020;**8**:31.
10. Chen X, Yan CC, Zhang X, *et al.* Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017;**18**(4):558–76.
11. Chen X, Yan G-Y. Novel human lncRNA-Cdisease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;**29**(20):2617–24.
12. Sun J, Shi H, Wang Z, *et al.* Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol Biosyst* 2014;**10**(8):2074–81.
13. Yu G, Fu G, Lu C, *et al.* BRWLDA: bi-random walks for predicting lncRNA-disease associations. *Oncotarget* 2017;**8**(36):60429–46.
14. Fu G, Wang J, Domeniconi C, *et al.* Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 2017;**34**(9):1529–37.
15. Lu C, Yang M, Luo F, *et al.* Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics* 2018;**34**(19):3357–64.
16. Xuan Z, Li J, Yu J, *et al.* A Probabilistic Matrix Factorization Method for Identifying lncRNA–Disease Associations. *Genes* 2019;**10**(2):126.
17. Yu G, Wang Y, Wang J, *et al.* Weighted matrix factorization based data fusion for predicting lncRNA-disease associations. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2018**:572–7.
18. Liu J-X, Gao M-M, Cui Z, *et al.* DSCMF: prediction of LncRNA-disease associations based on dual sparse collaborative matrix factorization. *BMC Bioinformatics* 2021;**22**(3):241.
19. Wang M-N, You Z-H, Wang L, *et al.* LDGRNMF: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* 2021;**424**:236–45.
20. Lan W, Li M, Zhao K, *et al.* LDAP: a web server for lncRNA–disease association prediction. *Bioinformatics* 2016;**33**(3):458–60.
21. Yu J, Xuan Z, Feng X, *et al.* A novel collaborative filtering model for LncRNA–disease association prediction based on the Naive Bayesian classifier. *BMC Bioinformatics* 2019;**20**(1):396.
22. Zhou J-R, You Z-H, Cheng L, *et al.* Prediction of lncRNA-disease associations via an embedding learning HOPE in heterogeneous information networks. *Molecular Therapy-Nucleic Acids* 2021;**23**:277–85.
23. 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-CDisease Associations. *Int J Mol Sci* 2019;**20**(17):4260.
24. 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.

25. Yang Q, Li X. BiGAN: LncRNA-disease association prediction based on bidirectional generative adversarial network. *BMC Bioinformatics* 2021;**22**(1):357.
26. Zeng M, Lu C, Zhang F, *et al.* SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods* 2020;**179**:73–80.
27. Lan W, Lai D, Chen Q, *et al.* LDICDL: LncRNA-disease association identification based on Collaborative Deep Learning. *IEEE/ACM Trans Comput Biol Bioinform* 2020;1–12.
28. 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* 2020;**22**(3).
29. 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.
30. 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):136.
31. Wu X, Lan W, Chen Q, *et al.* Inferring LncRNA-disease associations based on graph autoencoder matrix completion. *Comput Biol Chem* 2020;**87**:107282.
32. 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).
33. 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.
34. 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* 2020;**49**(D1):D1251–8.
35. Huang Z, Shi J, Gao Y, *et al.* HMDD v3.0: a database for experimentally supported human microRNÄCdisease associations. *Nucleic Acids Res* 2018;**47**(D1):D1013–7.
36. Li J-H, Liu S, Zhou H, *et al.* starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-CRNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2013;**42**(D1):D92–7.
37. Chen X, Clarence Yan C, Luo C, *et al.* Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci Rep* 2015;**5**(1):11338.
38. Wang D, Wang J, Lu M, *et al.* Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**(13):1644–50.
39. Breiman L. Random Forests. *Machine Learning* 2001;**45**(1):5–32.
40. 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.
41. Ke G, Meng Q, Finley T, *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 2017;**30**:3146–54.
42. Friedman JHJAos. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 2001;3146–54.
43. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 2016;785–94.
44. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 1997;**55**(1):119–39.
45. Murthy SK. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery* 1998;**2**(4):345–89.
46. Marmol I, Sanchez-de-Diego C, Pradilla Dieste A, *et al.* Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int J Mol Sci* 2017;**18**(1):197.
47. Zhou Z, Tan F, Pei Q, *et al.* lncRNA SNHG4 modulates colorectal cancer cell cycle and cell proliferation through regulating miR-590-3p/CDK1 axis. *Aging* 2021;**13**(7):9838–58.
48. Sharma GN, Dave R, Sanadya J, *et al.* Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology and research* 2010;**1**(2):109–26.
49. Pang B, Wang Q, Ning S, *et al.* Landscape of tumor suppressor long noncoding RNAs in breast cancer. *Journal of Experimental and Clinical Cancer Researchs* 2019;**38**(1):79.
50. Zuo X, Wang H, Mi Y, *et al.* The association of CASC16 variants with breast Cancer risk in a northwest Chinese female population. *Mol Med* 2020;**26**(1):11.
51. Huang R, Zhang Y, Han B, *et al.* Circular RNA HIPK2 regulates astrocyte activation via cooperation of autophagy and ER stress by targeting MIR124-2HG. *Autophagy* 2017;**13**(10):1722–41.
52. Rawla P. Epidemiology of Prostate Cancer. *World journal of oncology* 2019;**10**(2):63–89.