Briefings in Bioinformatics
doi.10.1093/bib/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category
OXFORD

Data and text mining

Learning global dependencies and multi-semantics within heterogeneous graph for predicting disease-related lncRNAs

Ping Xuan [1, 2], Shuai Wang [1], Hui Cui [3], Yue Zhao [2], Tiangang Zhang [4,*], Peiliang Wu [1,*]

[1]School of Information Science and Engineering (School of Software), Yanshan University, Qinhuangdao 066004, China, [2]School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China, [3]Department of Computer Science and Information Technology, La Trobe University, Melbourne 3083, Australia, [4]School of Mathematical Science, Heilongjiang University, Harbin 150080, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Long non-coding RNAs (lncRNAs) play an important role in the occurrence and development of diseases. Predicting disease-related lncRNAs can help to understand the pathogenesis of diseases deeply. Existing methods mainly rely on multi-source data related to lncRNAs and diseases when predicting the associations between lncRNAs and diseases. There are interdependencies among node attributes in a heterogeneous graph composed of all lncRNAs, diseases and miRNAs. The meta-paths composed of various connections between them also contain rich semantic information. However, existing methods neglect to integrate attribute information of intermediate nodes in meta-paths.

Results: We propose a novel association prediction model, GSMV, to learn and deeply integrate the global dependencies, semantic information of meta-paths, and node-pair multi-view features related to lncRNAs and diseases. We firstly formulate the global representations of the lncRNA and disease nodes by establishing a self-attention mechanism to capture and learn the global dependencies among node attributes. Second, starting from the lncRNA and disease nodes, respectively, multiple meta-pathways are established to reveal different semantic information. Considering that each meta-path contains specific semantics and has multiple meta-path instances which have different contributions to revealing meta-path semantics, we design a graph neural network (GNN) based module which consists of a meta-path instance encoding strategy and two novel attention mechanisms. The proposed meta-path instance encoding strategy is used to learn the contextual connections between nodes within a meta-path instance. One of the two new attention mechanisms is at the meta-path instance level, which learns rich and informative meta-path instances.The other attention mechanism integrates various semantic information from multiple meta-paths to learn the semantic representation of lncRNA and disease nodes. Finally, a dilated convolution-based learning module with adjustable receptive fields is proposed to learn multi-view features of lncRNA-disease node pairs. The experimental results prove that our method outperforms seven state-of-the-art comparing methods for lncRNA-disease association prediction. Ablation experiments demonstrate the contributions of the proposed global representation learning, semantic information learning, pairwise multi-view feature learning, and the meta-path instance encoding strategy. Case studies on three cancers further demonstrate our method's ability to discover potential disease-related lncRNA candidates.

Contact: zhang@hlju.edu.cn or peiliangwu@ysu.edu.cn

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

Ping Xuan, PhD (Harbin Institute of Technology), is a professor at the School of Computer Science and Technology, Heilongjiang University, Harbin, China. Her current research interests include computational biology, complex network analysis, and medical image analysis.

Shuai Wang, is studying for his master's degree in the School of Information Science and Engineering (School of Software) at Yanshan University, Qinhuangdao, China. His research interests include complex network analysis and deep learning.

Hui Cui, PhD (The University of Sydney), is a lecturer at Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia. Her research interests lie in data-driven and computerized models for biomedical and health informatics.

Yue Zhao, is studying for her master's degree in the School of Computer Science and Technology at Heilongjiang University, Harbin, China. Her research interests include complex network analysis and deep learning.

Tiangang Zhang, PhD (The University of Tokyo), is an associate professor of the Department of Mathematical Science, Heilongjiang University, Harbin, China. His current research interests include complex network analysis and computational fluid dynamics.

Peiliang Wu, PhD (Yanshan University), is a professor at the School of Information Science and Engineering (School of Software) , Yanshan University, Qinhuangdao, China. His current research interests include machine learning, image analysis and intelligent robot.

1 Introduction

Long non-coding RNAs (lncRNAs) of over 200 nucleotides in length are RNAs that are not involved in coding for proteins [1–3]. A growing number of studies have shown that abnormal expression of lncRNAs accompanies the onset and progression of diseases [4–6]. Therefore, discovering lncRNAs that are aberrantly expressed in specific diseased tissues contributes to exploring disease pathogenesis and facilitates the diagnosis and treatment of diseases.

Computational prediction methods for lncRNA-disease association can screen potential disease-associated lncRNA candidates [7, 8], which reduces the cost and time of biological experiments. Existing methods are classified into three main categories. The methods in the first category utilize biological information of lncRNAs, including genomic location, tissue specificity, and expression profile. Clark *et al.* and Li *et al.* predicted lncRNA-disease associations by using known genomic location [9, 10]. However, genomic locations-based methods cannot be used for lncRNAs without neighboring genes. Issue specificity and expression profiles have also been used by Chen *et al.*and Lin *et al.* [11, 12]. However, these two methods are hindered by the limited information of tissue-specific expression and the low-level expression of lncRNAs.

The second category focuses on developing models based on machine learning strategies. Chen *et al.* hypothesized that functionally similar lncRNAs are more likely to be involved in identical disease processes. Afterwards, they designed a prediction model based on Laplace regularized least squares (LRLSDA) to predict lncRNA-associated lncRNA candidates [13]. Chen *et al.* introduced two new lncRNA functional similarity calculation methods to improve the performance of LRLSDA [14]. A couple of methods infer disease-related lncRNA candidates by using random walk [15–17] and the Naive Bayes classifie [18]. Other machine learning methods, including support vector machines(SVM) [19, 20], matrix factorization [21–23] , and random forest [24, 25] have also been used to predict lncRNA-disease associations. In addition, Li *et al.* used a strategy of ensemble learning to integrate different types of strong predictive performance models, including support vector machines, non-negative matrix factorization, and KATZ [26]. However, these methods fail to integrate other data involved in the disease process, including proteins and miRNA.

The third category of prediction models are based on deep learning methods to improve the prediction performance. Convolutional neural network (CNN) based models were built to infer the propensity of lncRNA-disease associations [27–29]. Xuan *et al.* proposed a neighbor topology encoding strategy to learn the neighbor topology of lncRNA and disease nodes in heterogeneous graphs [30]. In addition, several methods predict disease-related lncRNA candidates by using autoencoders [31–33], generative adversarial networks [34, 35], graph convolutional networks [36–38], and graph attention networks [39, 40]. These methods have achieved better predictive performance, as deep learning-based models can learn the deep relationships between lncRNAs and diseases. These previous GNN-based prediction methods learn the representation of a target lncRNA (disease) node mainly by aggregating the attributes of the nodes in its neighborhood, under the hypothesis that the characteristics of the neighbors are more likely to be similar to that of the target node. However, the nodes with similar attributes are not only located around the target node's neighbourhood but also in the regions far away from the target node [41–43]. Besides, the previous methods did not have the capacity to extract the rich semantic information in the heterogeneous graph composed of lncRNAs and diseases.

In this work, we propose a lncRNA-disease association prediction model, GSMV, to learn and integrate the global dependencies between nodes, semantic information from meta-paths, and multi-view features of node pairs. The contributions of our model are summarized below.

- We proposed a novel strategy based on self-attention to learn the global dependencies among attributes of all the lncRNA, disease and miRNA nodes in the lncRNA-disease-miRNA heterogeneous graph. Moreover, we deeply integrated the global features by global dependency learning and the information from the local neighborhoods learned by GNN.
- Multiple meta-paths are established to enable the extraction and formulation of the similarity connections and association connections across lncRNA, miRNA, and disease nodes. Most of the existing meta-path based methods only aggregated the neighbours based on a meta-path for target nodes while neglecting the attribute information of intermediate nodes within a meta-path instance. Therefore, we proposed a new meta-path instance encoding strategy (MIES) to utilize the local dependencies among the attributes of lncRNA, disease and miRNA nodes within the meta-path instance to encode features of the instance.
- For each meta-path, we propose an attention mechanism at meta-path instance level to discriminate the contributions from multiple meta-path instances for specific semantic learning. Since multiple kinds of meta-paths have different contributions to lncRNA-disease association prediction, we also design an attention mechanism at meta-path type level to fuse the diverse semantics from multiple meta-paths.
- We design a module based on dilated convolutions to encode the features of a pair of lncRNA and disease nodes from multiple different views. Our new approach is powerful in learning more abstract and global features from a large receptive field for a pair of lncRNA and disease nodes, and richer local and specific features from a small receptive field. Comprehensive evaluations and comparisons with seven prediction methods, ablation experiments, and case studies prove that our approach achieves superior prediction performance.

2 Materials and methods

We propose an prediction model GSMV (Figure 1) to predict lncRNA candidates related to a given disease. A triple-layer heterogeneous group is constructed to integrate similarities and associations between lncRNAs, diseases, and miRNAs. GSMV consists of three components that are used to learn different information. The global encoding and semantic encoding modules work to learn the global and semantic representations of the lncRNA and disease nodes in the heterogeneous graph, respectively. The purpose of the pairwise multi-view feature encoding module is to capture the feature representations of a pair of lncRNA and disease from multiple different views. These three representations are integrated by the full connection layer to further assess the lncRNA-disease association scores.

2.1 Dataset

The datasets used in this study are from a previous work [21] and include 240 lncRNAs, 405 diseases, and 495 miRNAs, with 2687 pairs of lncRNA and disease associations, 13559 pairs of miRNA and disease associations, and 1002 pairs of miRNA and lncRNA interactions. The raw data were extracted from the LncRNADisease database [44], HMDD database [45], and starBasev2.0 database [46].

2.2 LncRNA-disease-miRNA heterogeneous graph

We constructed a triple-layer heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ using associations and similarities related to lncRNAs, miRNAs, and diseases (Figure 2). The node sets $\mathcal{V} = \{V^{lnc} \cup V^{dis} \cup V^{mir}\}$ consist of the lncRNA set V^{lnc}, disease set V^{dis}, and miRNA set V^{mir}. Edge $e_{ij} \in \mathcal{E}$ connects a pair of nodes $v_i, v_j \in \mathcal{V}$. According to the type of nodes connected by edges, $\mathcal{E}$ is divided into inter-layer and intra-layer edges, which are represented by the inter-layer association matrix O and intra-layer similarity matrix I, respectively.

The inter-layer association matrix O is defined as

$$O = \begin{cases} O^{lnc-dis} \in \mathbb{R}^{N_{lnc} \times N_{dis}} \\ O^{mir-dis} \in \mathbb{R}^{N_{mir} \times N_{dis}} \\ O^{lnc-mir} \in \mathbb{R}^{N_{lnc} \times N_{mir}} \end{cases}, \tag{1}$$

where N_{lnc}, N_{dis} and N_{mir} denote the number of lncRNAs, diseases, and miRNAs in the dataset, respectively. $O^{lnc-dis}$, $O^{mir-dis}$ and $O^{lnc-mir}$ represent the lncRNA-disease association matrix, miRNA-disease associations matrix, and lncRNA-miRNA interactions matrix, respectively.

Given $l_i \in V^{lnc}(m_i \in V^{mir})$, $d_j \in V^{dis}$, if $O_{ij}^{lnc-dis} = 1(O_{ij}^{mir-dis} = 1)$, then $l_i(m_i)$ is known to be associated with d_j. Further, $O_{ij}^{lnc-dis} = 0(O_{ij}^{mir-dis} = 0)$ indicates that $l_i(m_i)$ has not been observed to be associated

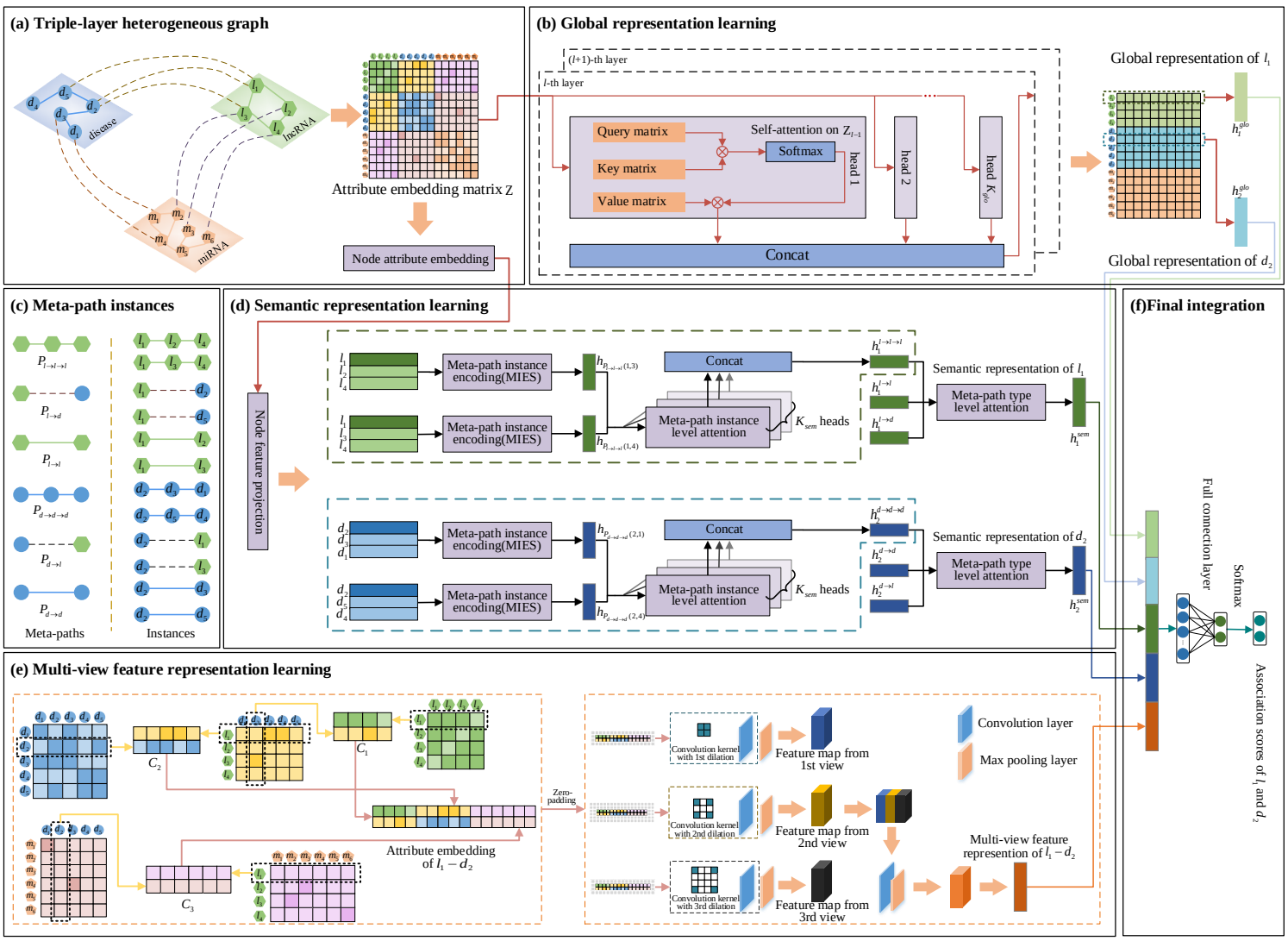

Fig. 1. Framework of the proposed GSMV model. **(a)** the triple-layer heterogeneous graph is constructed to integrate the similarities and associations among lncRNAs, diseases, and miRNAs **(b)** the first component of GSMV learn the global representations of each node by self-attention mechanism **(c)** various meta-paths about the lncRNA and disease nodes and their instances are constructed **(d)** the second component learns the semantic representations of lncRNA and disease nodes by graph neural network **(e)** the third component learns the pairwise multi-view features of a pair of lncRNA and disease nodes by dilated convolutions **(f)** three representations are integrated for estimating the association score of the lncRNA and disease node pair.

with d_j at present. For $m_j \in V^{mir}$, if $O_{ij}^{lnc-mir} = 1$, there is a known interaction between l_i and m_j. Otherwise, $O_{ij}^{lnc-mir} = 0$.

The intra-layer similarity matrix I is defined as

$$I = \begin{cases} I^{lnc} \in \mathbb{R}^{N_{lnc} \times N_{lnc}} \\ I^{dis} \in \mathbb{R}^{N_{dis} \times N_{dis}} \\ I^{mir} \in \mathbb{R}^{N_{mir} \times N_{mir}} \end{cases}, \tag{2}$$

where I^{lnc}, I^{dis} and I^{mir} contain the similarities between lncRNAs, between diseases, and between miRNAs, respectively.

The intra-layer similarity matrix takes values between 0 and 1, reflecting the level of similarity between two nodes of the same type, with higher scores representing more similarity between them. The disease similarity is calculated using the directed acyclic graph (DAG) composed of diseases, which was presented by Wang *et al.* [47]. LncRNA and miRNA similarities are calculated using the methods presented by Chen *et al.* and Wang *et al.*, respectively, which are on the basis that the more diseases two lncRNAs (miRNAs) are associated with, the more similar these two lncRNAs (miRNAs) are [47, 48].

Given O and I, the attribute matrix of a heterogeneous graph is defined as

$$Z = \begin{bmatrix} I^{lnc} & O^{lnc-dis} & O^{lnc-mir} \\ O^{lnc-dis^T} & I^{dis} & O^{mir-dis^T} \\ O^{lnc-mir^T} & O^{mir-dis} & I^{mir} \end{bmatrix}, \tag{3}$$

where A^T denotes the transpose of A. The i-th row of Z contains the associations and similarities of node $v_i \in \mathcal{V}$ with all lncRNAs, miRNAs, and diseases and can be considered as the attribute embedding of v_i, renamed as h_i. $\{h_i | 0 < i < N_{lnc}\}$ denotes the set of attribute embeddings for all lncRNAs. $\{h_i | N_{lnc} < i < N_{lnc} + N_{dis}\}$ and $\{h_i | N_{lnc} + N_{dis} < i < N_{lnc} + N_{dis} + N_{mir}\}$ contain the attribute embeddings for all diseases and all miRNAs, respectively.

2.3 Global encoding based on self-attention mechanism

Based on the biological premise that the lncRNAs with similar functions are more likely to be involved in the similar disease processes [13], we learn the representation of a target lncRNA (disease) node by aggregating the attributes of the nodes whih high similarity to it. Meanwhile, the nodes with similar attributes may locate the regions where are far away from the target node, which is ignored by previous GNN-based methods that only aggregate local neighbors. That is, there are also global dependencies among the attributes of all the nodes in the whole heterogeneous network. Inspired by the Transformer proposed in [49], we designed a global encoding module based on the self-attention mechanism to capture the global dependencies from all lncRNAs, miRNAs, and diseases nodes (Figure 1(b)). The module consists of N_{glo} encoding layers, and we explain the learning process for global dependencies using the l-th layer as an example.

We introduce the multi-head attention mechanism, which helps reduce variance during learning. For the m-th attention head, we first obtain the

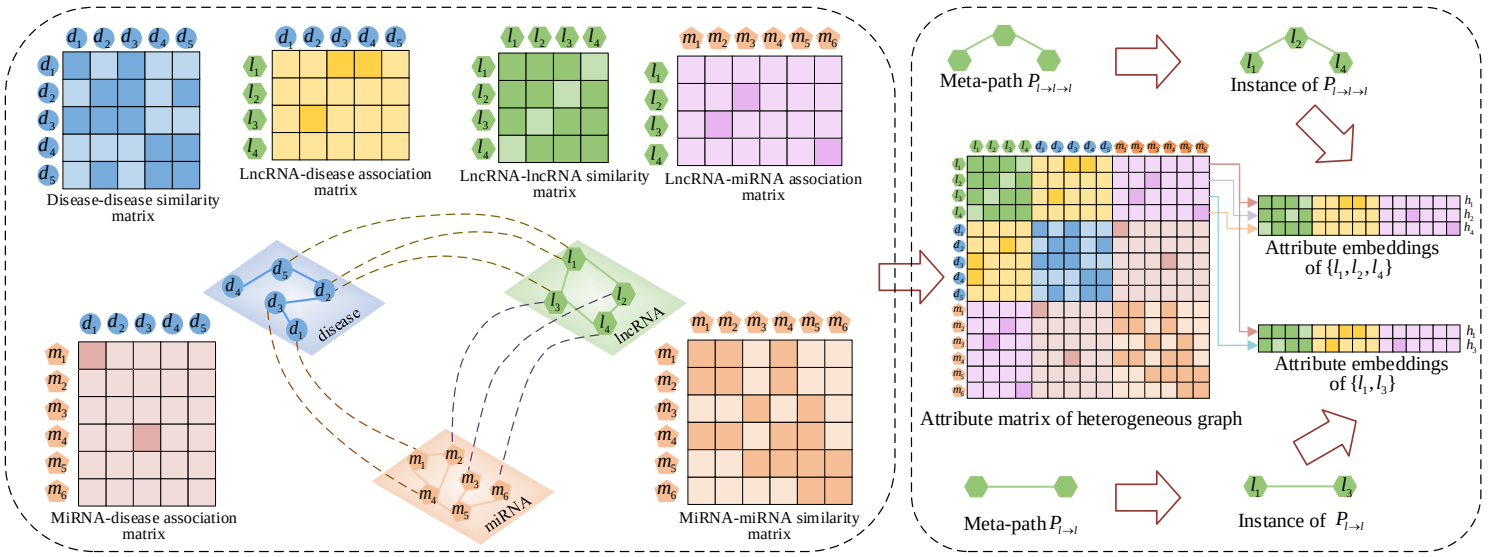

Fig. 2. Construction of the lncRNA-disease-miRNA heterogeneous graph based on multi-source data, and taking l_1 as an example to explain the attribute embedding of the nodes within a meta-path instance.

query matrix $Q_l^m \in \mathbb{R}^{\frac{N^l}{K_{glo}}}$, key matrix $K_l^m \in \mathbb{R}^{\frac{N^l}{K_{glo}}}$, and value matrix $V_l^m \in \mathbb{R}^{\frac{N^l}{K_{glo}}}$ with different linear transformations as follows,

$$\begin{aligned} Q_l^m &= W_l^{m,Q} \cdot Z_{l-1} \\ K_l^m &= W_l^{m,K} \cdot Z_{l-1} \\ V_l^m &= W_l^{m,V} \cdot Z_{l-1} \end{aligned}, \tag{4}$$

where K_{glo} is the number of attention heads, and $\frac{N^l}{K_{glo}}$ is the dimention of per head. $W_l^{m,Q}$, $W_l^{m,K}$ and $W_l^{m,V}$ denote weight matrices. Z_{l-1} represents the global feature matrix obtained at layer $l-1$, and Z_0 is the original attribute matrix Z. Subsequently, we calculate the dot product of Q_l^m and K_l^m to form the attention score matrix $Q_l^m {K_l^m}^T$. The i-th row, $(Q_l^m {K_l^m}^T)_i$, of $Q_l^m {K_l^m}^T$ records the attention scores of all lncRNA, miRNA, and disease nodes on $v_i \in \mathcal{V}$. We normalize $Q_l^m {K_l^m}^T$ and multiply it by V_l^m to form the global feature matrix of the m-th head,

$$Z_l^m = \text{softmax}(\frac{Q_l^m \cdot {K_l^m}^T}{\sqrt{d}}) \cdot V_l^m, \tag{5}$$

where $d = \frac{N^l}{K_{glo}}$, which helps to smooth the gradient during training. Finally, the output of the l-th layer is obtained by concatenating the results of K_{glo} attention heads,

$$Z_l = \Big\|_{m=1}^{K_{glo}} Z_l^m, \tag{6}$$

where $||$ denotes the concatenation operation. After the learning of N_{glo} encoding layers, we can obtain $Z_{N_{glo}}$. Its rows $i (i < N_{lnc})$ and $j (N_{lnc} < j < N_{lnc} + N_{dis})$ are the global representations of $l_i \in V^{lnc}$ and $d_j \in V^{dis}$, renamed as h_i^{glo} and h_j^{glo}, respectively.

2.4 Semantic encoding based on graph neural network enhanced by meta-path

Many nodes in the heterogeneous graph $\mathcal{G}$ can be connected by different paths, which have different semantics and are called meta-paths [50]. The meta-path P of length l is defined as

$$u_1 \xrightarrow{e_1} u_2 \xrightarrow{e_2} \cdots \xrightarrow{e_l} u_{l+1}, \tag{7}$$

where u_i denotes the node type, and e_i denotes the type of edge connecting u_i and u_{i+1}. A meta-path instance is defined as a sequence of nodes in the heterogeneous graph following the structure defined by P. For Instance, let l, d, and m denote lncRNA node type, disease node type, and miRNA node type, respectively. In Figure 1, two lncRNA nodes are connected by multiple meta-paths (e.g., $l \to l \to l$ and $l \to l$). For the meta-path $l \to l \to l$, $l_1 \to l_2 \to l_4$ is its meta-path instance related to the target node l_1. In this study, our model learns the semantic representations of a lncRNA nodes according to $l \to d$, $l \to l$ and $l \to l \to l$. The corresponding representations are denoted as $P_{l\to d}$, $P_{l\to l}$ and $P_{l\to l\to l}$, respectively. When $d \to l$, $d \to d$ and $d \to d \to d$ are established to learn the semantic representations of a disease nodes, we denote the representations as $P_{d\to l}$, $P_{d\to d}$ and $P_{d\to d\to d}$, respectively (Figure 1(c)).

As different meta-paths always imply specific semantic information, we design a semantic encoding module that enhances the graph neural network with meta-paths (Figure 1(d)). We describe the module using lncRNA l_i as an example, and the semantic representations of disease nodes are learned in a similar manner.

We design a type-specific projection matrix to project the original attribute embedding of $h_i \in \mathbb{R}^{N_{lnc}+N_{dis}+N_{mir}}$ into the low-dimensional feature space owing to its sparsity. The projection vector of l_i is $\overline{h_i} \in \mathbb{R}^{N_{fea}}$,

$$\overline{h_i} = W_{lnc} \cdot h_i, \tag{8}$$

where W_{lnc} denotes weight matrix and N_{fea} is the dimention of the projection space.

2.4.1 Feature encoding of meta-path instances

Local dependencies exist between lncRNA, miRNA, and disease nodes within meta-path instances. Taking meta-path $P_{l\to l\to l}$ as an example, $P_{l\to l\to l}(i,j)$ denotes the meta-path instance $l_i \to l_k \to l_j$ connecting target node l_i and meta-path based neighbor l_j. We establish a meta-path instance encoding strategy (MIES) for encoding meta-path instances based on the self-attention mechanism [49], which works by encoding $P_{l\to l\to l}(i,j)$ as a feature vector $h_{P_{l\to l\to l}(i,j)}$ using the local dependencies of l_i on l_j and l_k. We obtain the query vector Q_i by performing a linear transformation $\mathbb{R}^{N_{fea}}$ on the target node l_i as follows,

$$Q_i = W_{l-l-l}^Q \cdot \overline{h_i}, \tag{9}$$

where W_{l-l-l}^Q represent weight matrix. We transform l_j and l_k into the key vector K_i and the value vector V_i,

$$\begin{aligned} K_i &= W_{l-l-l}^K \cdot \overline{h}_t \\ V_i &= W_{l-l-l}^V \cdot \overline{h}_t \end{aligned}, \tag{10}$$

where W_{l-l-l}^K and W_{l-l-l}^V are weight matrices, $t = \{j,k\}$. In sequence, we compute the similarity between Q_i and K_i and multiply it with V_i to obtain the feature vector $h_{P_{l\to l\to l}(i,j)}$ of the meta-path instance $l_i \to l_k \to l_j$,

$$h_{P_{l\to l\to l}(i,j)} = \text{softmax}(Q_i \cdot {K_i}^T) \cdot V_i \ . \tag{11}$$

2.4.2 Node semantic-specific representation learning

Each meta-path typically has several instances that contribute differently to the semantic representation of the target node. Therefore, at the meta-path instance level, an attention mechanism is designed for learning more important meta-path instances. Unlike the traditional graph attention network [51], which focuses on the neighbor nodes directly connected to the target node, we can learn the attribute information of all nodes within a meta-path instance by using the proposed MIES. Given the meta-path P related to l_i, the attention score e_{ij}^P of the meta-path instance $P(i,j)$ is defined as

$$e_{ij}^P = \text{LeakyReLU}(a_P^T \cdot [\overline{h_i}||h_{P(i,j)}]), \tag{12}$$

where LeakyReLU indicates a nonlinear activation function. a_P is the attention vector of the meta-path P, and $h_{P(i,j)}$ denotes the feature vector of $P(i,j)$. The normalized attention weight α_{ij}^P of $P(i,j)$ is denoted as

$$\alpha_{ij}^P = \frac{\exp(e_{ij}^P)}{\sum_{s\in N_i^P}\exp(e_{is}^P)}, \tag{13}$$

where N_i^P denotes the set of meta-path based neighbors of l_i. The semantic-specific representation of l_i for meta-path P is h_i^P,

$$h_i^P = \text{elu}(\sum_{s\in N_i^P}\alpha_{is}^P \cdot h_{P(i,s)}), \tag{14}$$

We also establish multi-head attention to stabilize the semantic learning process. We obtain the final h_i^P by concatenating K_{sem} independent attentions,

$$h_i^P = \Big\|_{m=1}^{K_{sem}} \text{elu}(\sum_{s\in N_i^P}\alpha_{is}^{P,m} \cdot h_{P(i,s)}), \tag{15}$$

where elu is a nonlinear activation function.

2.4.3 Fusion of multiple semantic representations of nodes

Given meta-paths $P_{l\to l}, P_{l\to d}, P_{l\to l\to l}$, the semantic-specific representation of l_i is denoted as $h_i^{P_{l\to l}}$,$h_i^{P_{l\to d}}$, $h_i^{P_{l\to l\to l}}$. Each meta-path reflects a specific semantic that contributes distinctly to the prediction of lncRNA-disease associations. Therefore, an meta-path type-level attention mechanism is porposed which helps fuse multiple semantics. The attention score at the level of meta-path type is s_i^P,

$$s_i^P = q^T \cdot \tanh(M^{att} \cdot h_i^P + b^{att}), \tag{16}$$

where tanh represents a nonlinear activation function. $P \in \{P_{l\to l}, P_{l\to d}, P_{l\to l\to l}\}$, M^{att} and b^{att} is the attention vector , and q^T denote the learnable parameters. β_i^P represents the normalized attention weight,

$$\beta_i^P = \frac{\exp(s_i^P)}{\sum_{k\in\{P_{l\to l},P_{l\to d},P_{l\to l\to l}\}}\exp(s_i^k)}. \tag{17}$$

We obtain the semantic representation h_i^{sem} of l_i by weighting $h_i^{P_{l\to l}}$, $h_i^{P_{l\to d}}$, $h_i^{P_{l\to l\to l}}$,

$$h_i^{sem} = \sum_{P\in\{P_{l\to l},P_{l\to d},P_{l\to l\to l}\}}\beta_i^P \cdot h_i^P + h_i^P. \tag{18}$$

2.5 Pairwise multi-view features encoding based on CNN with multi receptive fields

2.5.1 Attribute embedding of lncRNA-disease node pair

In case $l_i \in V^{lnc}$ and $d_j \in V^{dis}$ have similarities, associations, or interactions with more common lncRNAs, diseases, and miRNAs, l_i and d_j are more likely to be related. Based on this biological premise, an embedding strategy is proposed to obtain attribute embeddings of lncRNA-disease node pairs from the multi-source data.

Given the association matrices $O^{lnc-dis}$, $O^{mir-dis}$, and $O^{lnc-mir}$, and the similarity matrices I^{lnc}, I^{dis}, and I^{mir}, the attribute embedding matrix H_{ij} of the node pair $l_i - d_j$ is obtained according to the following steps. We first combine the i-th row of I^{lnc}, $I_{i,*}^{lnc}$, and the j-th column of $O^{lnc-dis}$, $O_{*,j}^{lnc-dis}$, to form C_1,

$$C_1 = [I_{i,*}^{lnc}; O_{*,j}^{lnc-dis}], \tag{19}$$

where ; is a stacking operation. $I_{i,*}^{lnc}$ represents the similarities of l_i to all lncRNAs. $O_{*,j}^{lnc-dis}$ denotes associations of d_j with all the lncRNAs. $O_{i,*}^{lnc-dis}$ is taken from row i of $O^{lnc-dis}$, which covers the associations between l_i and all the diseases. $I_{j,*}^{dis}$ is taken from row j of I^{dis}, which covers the similarities between d_j and all the diseases. They are stacked into C_2,

$$C_2 = [O_{i,*}^{lnc-dis}; I_{j,*}^{dis}]. \tag{20}$$

C_3 is obtained by stacking the i-th row of $O^{lnc-mir}$, $O_{j,*}^{lnc-mir}$, and the j-th column of $O^{mir-dis}$, $O_{*,j}^{mir-dis}$, which contain interactions of l_i with all the miRNAs and associations between d_j and all the miRNAs, respectively,

$$C_3 = [O_{i,*}^{lnc-mir}; O_{*,j}^{mir-dis}]. \tag{21}$$

The attribute embedding matrix $H_{ij} \in \mathbb{R}^{2\times(N_{lnc}+N_{dis}+N_{mir})}$ is defined as

$$\begin{aligned} H_{ij} &= [C_1\ C_2\ C_3] \\ &= \begin{bmatrix} I_{i,*}^{lnc} & O_{i,*}^{lnc-dis} & O_{i,*}^{lnc-mir} \\ O_{*,j}^{lnc-dis} & I_{j,*}^{dis} & O_{*,j}^{mir-dis} \end{bmatrix}. \end{aligned} \tag{22}$$

2.5.2 Multi-view features extraction

Owing to their fixed receptive field, traditional CNNs can only capture features from a single view. With dilated convolution, the receptive field of the convolution kernel can be expanded to be a variable size depending on the dilation rate. Different sizes of receptive fields can extract features from different views, whereas larger receptive fields result in more abstract and global features, and smaller receptive fields mean more local and specific features. Therefore, we designed a pairwise multi-view feature encoding module to extract and fuse the multi-view features of lncRNA and disease node pair ($l_i - d_j$) from H_{ij} (Figure 1(e)).

To preserve the marginal information, we perform a zero-padding operation on H_{ij} to obtain $\tilde{H}_{ij} \in \mathbb{R}^{L\times H}$. The convolution layer contains N_{conv} filters whose length and width are l_c and w_c, respectively. Given the dilatation rate d, the region affected when the q-th filter is moved to the m-th row and the n-th column of $\tilde{H}_{ij}$ is $Z_{q,m,n}^d$,

$$Z_{q,m,n}^d = \tilde{H}_{ij}(m : m + (w_c - 1)\times d, n : n + (l_c - 1)\times d), \tag{23}$$

where $q \in [1, N_{conv}]$, $m \in [1, 2+L+d-d\times w_c]$, $n \in [1, 2+H+d-d\times l_c]$. By applying the q-th filter to $Z_{q,m,n}^d$, we can obtain the element value $Z_q^d(m,n)$ in row m and column n on the q-th feature map,

$$Z_q^d(m,n) = \text{ReLU}(W_{conv}(i,j,q) * Z_{q,m,n}^d + b_{conv}(q)), \tag{24}$$

where * represents the convolution operation, and ReLU indicates a nonlinear activation function. W_{conv} and b_{conv} represent weight matrix and bias vector, respectively.

The max pooling layer is used to select the more important features on the feature graph, and we set the length and width of the filter to l_p and w_p, respectively. The element value $P_{q,m,n}^d$ in row m and column n on the q-th feature graph via the max pooling layer is obtained as,

$$P_{q,m,n}^d = \max(Z_q^d(m : m + w_p, n : n + l_p)). \tag{25}$$

The feature Y^d of $\tilde{H}_{ij}$ from view d is obtained by passing $\tilde{H}_{ij}$ through the convolution-pooling layer described above. Given $d = d_1, d_2, d_3$, we can separately obtain Y^1, Y^2 and Y^3, which reflect pairwise features from different views. To fuse the multi-view features, we concatenate Y^1, Y^2, Y^3 and then feed it to another convolution-pooling layer to obtain $\hat{Y}$. h_{ij}^{pai} is obtained by feeding $\hat{Y}$ to a fully connected layer, which denote the multi-view feature representation at the level of node pairs,

$$h_{ij}^{pai} = \text{ReLU}(W_{flat} \cdot \hat{Y} + b_{flat}), \tag{26}$$

where W_{fla} denote weight matrix and b_{fla} is bias vector.

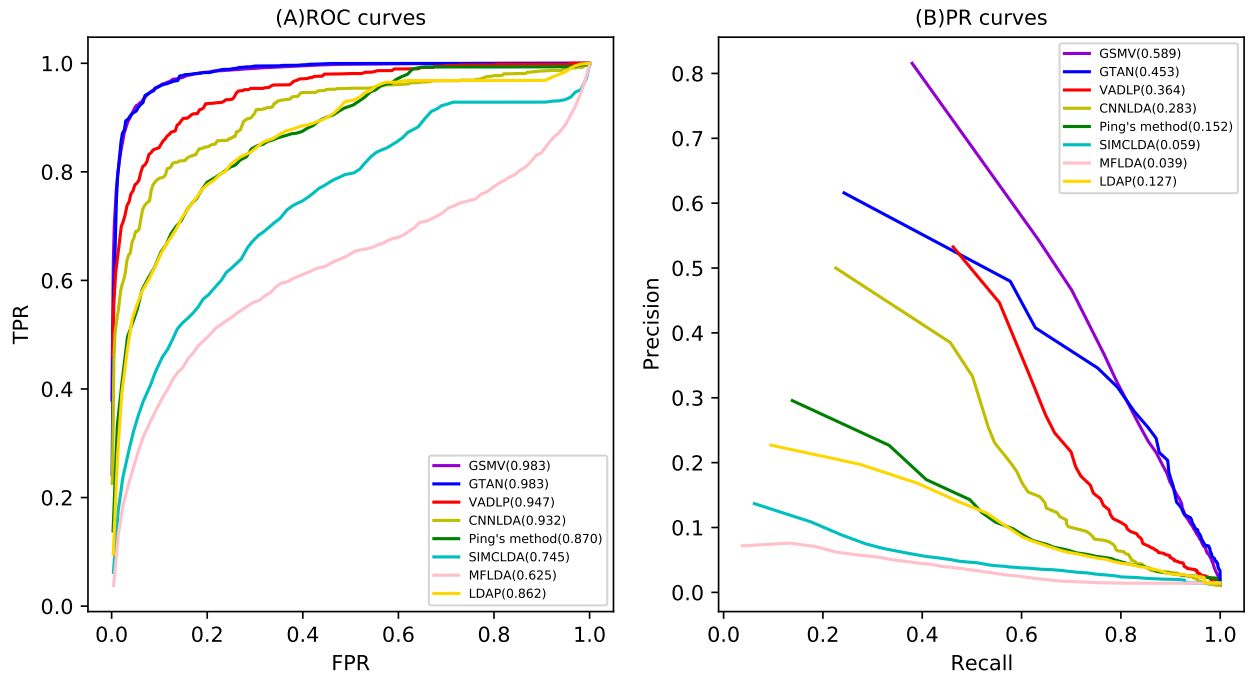

Fig. 3. Average ROC and PR curves for five-fold cross-validation of our method and the compared methods over 405 diseases.

2.6 Final integration and training

The global representation of $l_i \in V^{lnc}$ is denoted as h_i^{sem}, and its semantic representation is h_i^{glo}. h_j^{glo} and h_j^{sem} represent the global and semantic representations of $d_j \in V^{dis}$, respectively. The pairwise multi-view feature representation of $l_i - d_j$ is h_{ij}^{pai}. As described in Figure 1(f), they are concatenated to form the final representation h_{ij} of l_i and d_j,

$$h_{ij} = [h_i^{glo}||h_j^{glo}||h_i^{sem}||h_j^{sem}||h_{ij}^{pai}]. \tag{27}$$

y_{pre} denotes the probability distribution of whether l_i and d_j are related,

$$y_{pre} = \mathrm{softmax}(W_{agg} \cdot h_{ij} + b_{agg}), \tag{28}$$

where W_{agg} is weight matrix and b_{agg} represents bias vector.

During training, we use back propagation and gradient descent algorithms to optimize our model. The cross-entropy loss of our model is defined as

$$loss = -\sum_{i=1}^{N}\sum_{j=1}^{c} z[j] \cdot \log(y_{pre}[j]), \tag{29}$$

where N denotes the batch size of training samples, and z represents the real label. $c = 2$, $y_{pre}[1]$ and $y_{pre}[2]$ are the probabilities that l_i and d_j are associated or not, respectively.

3 Experimental Evaluations and Discussions

3.1 Parameter settings

The global encoding module contained two encoding layers with output feature dimensions of 800 and 480 ($N^1 = 800$, $N^2 = 480$), and the number of attention heads in each layer is 8 ($K_{glo} = 8$). For the semantic encoding module, the number of attention heads K_{sem} was set to 8, and the dimension of the projection space $N_{fea} = 800$ was set to 800. In the pairwise multi-view feature encoding module, the convolution-pooling layer used to extract multi-view features had 16 convolution filters of size 2×2 and an dilation rate d_1, d_2, d_3 of 1, 2, 3, respectively. The convolution-pooling layer used for feature fusion contained 32 convolution filters of size 2×4. The size of the pooling filter in both convolution-pooling layers was 2×5. GSMV was implemented based on the machine learning framework PyTorch, using an NVIDIA GeForce GTX TITAN X GPU card to accelerate the training. The training epoch was set to 100, and the weight decay and learning rate were 0.001 and 0.0003, respectively.

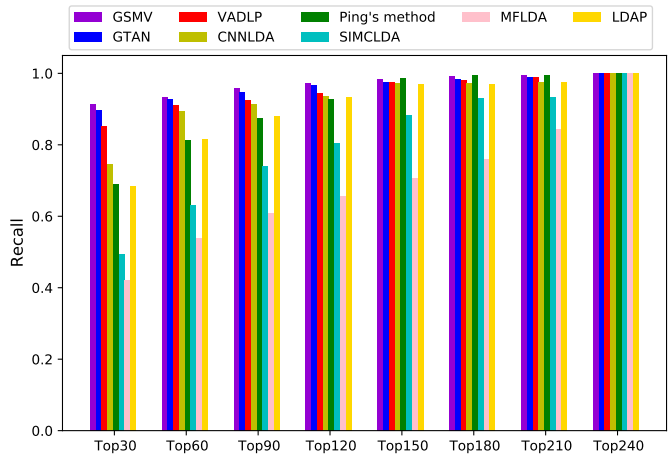

Fig. 4. Recall rates of GSMV and all the comparison methods at different top k values.

Table 1. Results of ablation experiments in our method.

GPL	SPL	PMVEL	Average AUC	Average AUPR
✗	✓	✓	0.976	0.551
✓	✗	✓	0.972	0.544
✓	✓	✗	0.948	0.433
✓	✓	✓	**0.983**	**0.589**

3.2 Evaluation metrics

Five-fold cross-validation helps to assess the generalization ability and prevent overfitting for evaluating the prediction performance of GSMV. We took all known lncRNA-disease associations as positive samples and divided them randomly into five equal parts. An unknown lncRNA-disease association means that there is no evidence from the biological experiment to confirm a pair of lncRNA and disease nodes are associated with each other. All the unknown lncRNA-disease associations are regarded as negative samples. In each fold, we adopted four parts of the positive samples and an equal number of randomly selected negative samples as the training set, whereas the test set consisted of the remaining part of the positive samples and all the remaining negative samples.

The area under the receiver operating characteristic (ROC) curve (AUC) [52] was used as an evaluation metric. The dataset contains 2687 known lncRNA-disease associations, while the number of the unobserved lncRNA-disease pairs

Table 2. Results of the experiments for comparing the proposed meta-path instance encoding strategy with the previous strategies.

	$GSMV_{HAN}$	$GSMV_{avg}$	$GSMV_{linear}$	$GSMV_{rot}$	$GSMV$
Average AUC	0.978	0.979	0.979	0.980	**0.983**
Average AUPR	0.569	0.572	0.575	0.582	**0.589**

Table 3. Results of the paired Wilcoxon test for comparing GSMV and all the other methods.

	GTAN	VADLP	CNNLDA	Ping's method	SIMCLDA	MFLDA	LDAP
p-value of AUC	3.7419e-02	2.8498e-03	2.7062e-04	8.1854e-07	2.4303e-04	5.9560e-06	9.0107e-07
p-value of AUCPR	8.3938e-04	6.0215e-04	6.6206e-04	4.5843e-08	9.0348e-05	2.0746e-05	4.9248e-08

is 94513. Therefore, we also selected the area under the precision-recall (PR) curve (AUPR) as the evaluation metric due to AUPR is more informative than AUC under such kind of cases when the known lncRNA-disease associations and the unknown associations are very imbalanced [53]. Given that the top-ranked candidates are more likely to be selected by biologists for further validation, we calculated the recall of the top $k \in [30, 60, \cdots, 240]$ candidates.

3.3 Ablation experiments

To verify the effectiveness of global representation learning (GPL), semantic representation learning (SPL), and pairwise multi-view feature learning (PMVFL), we conducted ablation experiments (Table 1).The final model including GPL, SPL, and PMVFL achieved the highest AUC (AUC = 0.983) and AUPR (AUPR = 0.589). Compared with the final model, the AUC and AUPR showed a decrease of 0.7% and 3.8%, respectively, after removing GPL. Ignoring SPL resulted in a 1.1% drop in AUC and 4.5% drop in AUPR. The result suggests that both GPL and SPL can help improve the performance of lncRNA-disease association prediction. Compared with the model without PMVFL, our method improved AUC and AUPR by 3.5% and 15.6%, respectively, which demonstrates the contribution of PMVFL for predicting the disease-related lncRNA candidates.

To estimate the effectiveness of the proposed MIES for lncRNA-disease association prediction, we compared it with existing methods (Table 2). Unlike our method, a previous method [54] neglected the intermediate nodes in the meta-path instance and aggregated the attributes of the meta-path based neighbors directly. This strategy was used to replace the MIES in GSMV to form $GSMV_{HAN}$. Wang *et al.* proposed a series of meta-path instance encoders, including mean, linear, and relational rotation encoders [55]. We built $GSMV_{avg}$, $GSMV_{linear}$ and $GSMV_{rot}$ by replacing the MIES in GSMV with these three encoders, respectively.

As observed in Table 2, the performance of $GSMV_{HAN}$ in terms of both AUC and AUPR was lower than that of the other four methods, indicating that aggregate meta-path instances are necessary in lncRNA-disease association prediction. This is mainly because the attribute information of intermediate nodes within the meta-path instances contribute to the SPL of nodes. $GSMV_{avg}$ and $GSMV_{linear}$ neglect the context of nodes composing the meta-path instances when encoding the feature vectors for the meta-path instances; as a result, they performed worse than $GSMV_{rot}$ and GSMV. Both $GSMV_{rot}$ and GSMV learned the relationships between nodes within the meta-path instance, but GSMV achieved a higher AUC and AUPR. This is mainly because GSMV learns the dependencies of the target node on all other nodes within the meta-path instance by using self-attention mechanism.

3.4 Comparison with other methods

Seven of the best performing lncRNA-disease association prediction methods were used for comparison with GSMV, including GTAN [30], VADLP [33], CNNLDA [27], Ping's method [56], SIMCLDA [22], MFLDA [21], and LDAP [19] (see details in Supplementary File SF1). They were implemented according to the best reported free parameters, and the same training and test sets as GSMV were used in the cross-validation.

Figure 3 shows that GSMV obtained the highest average AUC of 0.983, which is higher than VADLP by 3.6%, higher than CNNLDA by 5.1%, higher than Ping's method by 11.3%, higher than SIMCLDA by 23.8%, higher than MFLDA by 35.8%, and higher than LDAP by 12.1%. The average AUPR of GSMV was 0.589. It was higher than that of GTAN, VADLP, CNNLDA, Ping's method, SIMCLDA, and MFLDA, LDAP by 13.6%, 22.5%, 30.6%, 43.7%, 53%, 55%, and 46.2%, respectively. We obtained 405 AUCs (AUPRs) for each prediction method for 405 diseases. We conducted a Paired Wilcoxon test using 405 AUC pairs (AUPR pairs) for any two methods. The statistical results in Table 3 indicated that GSMV is significantly superior to other comparing methods in terms of both AUC and AUPR, where p-values are less than 0.05.

SIMCLDA and MFLDA use methods based on non-negative matrix factorization and show poor performance. Compared with these two methods, Ping's method and LDAP have been improved, as they are based on information flow propagation and SVM, respectively, but they do not learn relevant information at the level of lncRNA-disease node pairs. GSMV and CNNLDA both use CNNs, and GSMV achieves the fusion of extracted features from multiple views, thus outperforming CNNLDA. The improvement of GSMV over GTAN and VADLP is mainly due to the rich semantic information extracted from multiple meta-paths and the dependencies between nodes.

For the recall rate of the top k, owning to the integration of multi-view features at the level of node pair, global representations, and semantic representations of nodes, our method was superior to others (Figure 4). When $k = 30$, GSMV obtained a recall rate of 91.3%, which was higher than that of the other methods (GTAN: 89.7%, VADLP: 85.1%, CNNLDA: 74.6%, Ping's method: 68.9%, LDAP: 68.5%, SIMCLDA: 49.3%, and MFLDA: 42%). When k was increased from 60 to 120, GSMV maintained the highest recall with 93.4%, 95.9%, and 97.2%, respectively. GTAN ranks second, achieving recall rates of 92.8%, 94.8%, and 96.8%, respectively.VADLP consistently outperformed CNNLDA with recall rates of 85.0%, 91%, and 92.4%, compared to 89.5%, 91.3%, and 93.5% for CNNLDA, respectively. LDAP (81.7%, 88.0%, 93.3%) and Ping's method (81.3%, 87.5%, 92.7%) performed slightly worse, but better than SIMCLDA, which had recall rates of 63%, 74%, and 80%, respectively. MFLDA achieved the lowest recall rate when , and the corresponding recall rates were 53.9%, 61%, and 66%, respectively.

Since in all the compared methods, GTAN, VADLP and CNNLDA utilized the same dataset about the lncRNAs, diseases, and miRNAs with our method, we conducted the control experiments on these methods and our method (GSMV). For each method, we randomly generated some connection edges as the positive samples, and then performed 5-fold cross-validation to evaluate the prediction performance (see details in Supplementary Table 2).

3.5 Case studies

We performed case studies on ovarian cancer, pancreatic cancer, and oesophageal cancer to further prove the capability of GSMV to discover potential disease-related lncRNA candidates. The lncRNA candidates for each cancer were ranked in descending order of association score. The top 15 lncRNA candidates of these three types of cancer are listed in Tables 4, 5, and 6, respectively.

To validate the predicted lncRNA-disease associations,we searched the LncRNADisease database [51], the Lnc2Cancer database [57], and published literature. LncRNA Disease documents the associations between lncRNAs and diseases that are extracted from the published literature. Lnc2Cancer records lncRNAs that are expressed at different levels in normal and diseased tissues, and they are all supported by strong experimental evidence.

Among the top 15 ovarian cancer-related lncRNA candidates (Table 4), 14 were found in LncRNADisease. The result indicates that these lncRNAs are indeed associated with ovarian cancer. Moreover, 14 lncRNA candidates were included in Lnc2Cancer, suggesting that they are expressed at abnormal levels in ovarian cancer tissues. For pancreatic cancer, 14 of the top 15 lncRNA candidates

Table 4. Top 15 ovarian cancer-related lncRNA candidates.

Rank	LncRNA name	Evidence	Rank	LncRNA name	Evidence
1	PVT1	Lnc2Cancer, LncRNADisease	9	GAS5	Lnc2Cancer, LncRNADisease
2	NEAT1	Lnc2Cancer, LncRNADisease	10	MEG3	Lnc2Cancer, LncRNADisease
3	MALAT1	Lnc2Cancer, LncRNADisease	11	HOXA11-AS	Lnc2Cancer, LncRNADisease
4	HOTAIR	Lnc2Cancer, LncRNADisease	12	BCYRN1	LncRNADisease
5	XIST	Lnc2Cancer, LncRNADisease	13	CCAT1	Lnc2Cancer, LncRNADisease
6	CDKN2B-AS1	Lnc2Cancer, LncRNADisease	14	CCAT2	Lnc2Cancer, LncRNADisease
7	UCA1	Lnc2Cancer, LncRNADisease	15	LSINCT5	Lnc2Cancer
8	H19	Lnc2Cancer, LncRNADisease			

Table 5. Top 15 pancreatic cancer-related lncRNA candidates.

Rank	LncRNA name	Evidence	Rank	LncRNA name	Evidence
1	H19	Lnc2Cancer, LncRNADisease	9	GAS5	Lnc2Cancer, LncRNADisease
2	MALAT1	Lnc2Cancer, LncRNADisease	10	NEAT1	Lnc2Cancer, LncRNADisease
3	AFAP1-AS1	Lnc2Cancer, LncRNADisease	11	HULC	Lnc2Cancer, LncRNADisease
4	LINC00675	Unconfirmed	12	CDKN2B-AS1	LncRNADisease
5	HOTAIR	Lnc2Cancer, LncRNADisease	13	LINC-ROR	Lnc2Cancer, LncRNADisease
6	PVT1	Lnc2Cancer, LncRNADisease	14	MIR17HG	LncRNADisease
7	HOTTIP	Lnc2Cancer, LncRNADisease	15	LINC00261	Lnc2Cancer, LncRNADisease
8	MEG3	Lnc2Cancer, LncRNADisease			

Table 6. Top 15 esophageal cancer-related lncRNA candidates.

Rank	LncRNA name	Description	Rank	LncRNA name	Description
1	HOTAIR	Lnc2Cancer, LncRNADisease	9	CCAT1	Lnc2cancer
2	CDKN2B-AS1	LncRNADisease	10	MEG3	Lnc2Cancer, LncRNADisease
3	H19	Lnc2Cancer, LncRNADisease	11	NEAT1	LncRNADisease
4	SPRY4-IT1	LncRNADisease	12	UCA1	Lnc2Cancer, LncRNADisease
5	SOX2-OT	LncRNADisease	13	XIST	Lnc2Cancer, LncRNADisease
6	HNF1A-AS1	Literature	14	CCAT2	Lnc2Cancer
7	MALAT1	Lnc2Cancer, LncRNADisease	15	PVT1	Lnc2Cancer, LncRNADisease
8	BCYRN1	LncRNADisease			

listed in Table 5 were confirmed by LncRNADisease and 12 by Lnc2Cancer. One lncRNA candidate in Table 5 is labeled "Unconfirmed", which means that we did not get the evidence for the lncRNA-disease candidate. Table 6 lists the top 15 lncRNA candidates associated with esophageal cancer, 12 and 9 of which were verified by LncRNADisease and Lnc2Cancer, respectively. Moreover, the HNF1A-AS1 candidate was proved by a recent study [58], which showed that HNF1A-AS1 is highly upregulated in oesophageal cancer tissues. The capability of GSMV to forcast possible lncRNA-disease associations was further demonstrated by case studies of these three types of cancer.

3.6 Prediction of novel lncRNA-disease associations

Finally, we trained GSMV using all known associations between lncRNAs and diseases, and then applied it to predict 405 disease-associated lncRNA candidates. The top 30 lncRNA candidates for all the diseases predicted by GSMV are listed in Supplementary Table ST1.

4 Conclusion

We presented a novel method to encode and integrate the attributes of the lncRNA and disease nodes, and the semantic information from multiple meta-paths for predicting the lncRNAs associated with diseases. The constructed multiple meta-paths contributed to encoding the diverse semantic information of the lncRNA and disease nodes. A framework composed of self-attention mechanism, graph neural networks, and multi-view learning was constructed to learn the global representation, the semantic one, and the multi-view feature one of the lncRNA and disease nodes. We designed an encoding strategy to integrate the attribute information of the intermediate nodes in a meta-path instance and that of the boundary nodes of the meta-path instance. Two attention mechanisms were presented to assign the higher weights to the more important meta-path instances and meta-paths. The cross-validation results showed that both AUC and AUPR of GSMV were higher than those of other comparative methods. GSMV also performed better in retrieving the actual associations between lncRNAs and diseases, as it achieved higher recall rates for the top ranked candidates. GSMV's ability in identifying potential lncRNA-disease associations was further confirmed by the case studies on the diseases including the ovarian, pancreatic, and oesophageal cancers.

Key Points

- A framework to learn three kinds of representations for the lncRNA and disease nodes, where global representation captures the global dependencies among the attributes of all the lncRNA, disease, and miRNA nodes, semantic representation implies the semantic information from multiple meta-paths, and multi-view feature representation reveals features of the lncRNA-disease node pair from multiple views.
- A novel meta-path instance encoding strategy (MIES) to enable the extraction of attribute information from both the intermediate and boundary nodes within a meta-path instance.
- Two new attentional mechanisms at meta-path instance level and meta-path type level to distinguish the different contributions of meta-path instances and meta-paths for semantic learning of the lncRNA and disease nodes.
- Improved prediction performance demonstrated by comparison with seven state-of-the-art methods for lncRNA-disease association prediction, comparison of recall rates at different top k values, and case studies on three cancers. Ablation and comparison experiments also demonstrate the effectiveness of the main components and the proposed MIES, respectively.

Funding

This work was supported by the Natural Science Foundation of China (62172143, 61972135, U20A20167); Natural Science Foundation of Heilongjiang Province (LH2019F049); Natural Science Foundation of Hebei Province (F202103079); China Postdoctoral Science Foundation

(2020M670939, 2019M650069); Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHLQ18104).

References

[1]Rinn J L, Chang H Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 2012;81:145-166.
[2]Kung J T Y, Colognori D, Lee J T. Long noncoding RNAs: past, present, and future. *Genetics* 2013;193(3):651-669.
[3]Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012;482(7385):339-346.
[4]Wapinski O, Chang H Y. Long noncoding RNAs and human disease. *Trends in cell biology* 2011;21(6): 354-361.
[5]Schmitz S U, Grote P, Herrmann B G. Mechanisms of long noncoding RNA function in development and disease. *Cellular and molecular life sciences* 2016;73(13):2491-2509.
[6]Batista P J, Chang H Y. Long noncoding RNAs: cellular address codes in development and disease. *Cell* 2013;152(6):1298-1307.
[7]Chen X, Yan C C, Zhang X, *et al.* Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings in bioinformatics* 2017;18(4):558-576.
[8]Liu M X, Chen X, Chen G, *et al.* A computational framework to infer human disease-associated long noncoding RNAs. *PloS one* 2014;9(1):e84408..
[9]Clark M B, Johnston R L, Inostroza-Ponta M, *et al.* Genome-wide analysis of long noncoding RNA355 stability. *Genome research* 2012;22(5):885-898.
[10]Li J W, Gao C, Wang Y C, *et al.* A bioinformatics method for predicting long noncoding RNAs associated with vascular disease. *Science China Life Sciences* 2014;57(8):852-857.
[11]Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Scientific reports* 2015;5(1):1-11.
[12]Lin XC, Zhu Y, Chen WB, *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(2):619-628.
[13]Chen X, Yan G Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;29(20):2617-2624.
[14]Chen X, Yan C C, Luo C, *et al.* Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Scientific reports*. 2015;5(1):1-12.
[15]Gu C, Liao B, Li X, *et al.* Global network random walk for predicting potential human lncRNA-disease associations. *Scientific reports* 2017;7(1):1-11.
[16]Li J, Zhao H, Xuan Z, *et al.* A novel approach for potential human LncRNA-disease association prediction based on local random walk. *IEEE/ACM transactions on computational biology and bioinformatics*. 2021;18(3):1049-1059.
[17]Wang L, Shang M, Dai Q, *et al.* Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks. *BMC bioinformatics* 2022;23(1):1-20.
[18]Zhao T, Xu J, Liu L, *et al.* Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Molecular BioSystems* 2015;11(1):126-136.
[19]Lan W, Li M, Zhao K, *et al.* LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 2017;33(3):458-460.
[20]Zhao J, Cheng W, He X, *et al.* Construction of a specific SVM classifier and identification of molecular markers for lung adenocarcinoma based on lncRNA-miRNA-mRNA network. *OncoTargets and therapy* 2018;11:3129-3140
[21]Fu G, Wang J, Domeniconi C, *et al.* Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. *Bioinformatics* 2018;34(9):1529-1537.
[22]Lu C, Yang M, Luo F, *et al.* Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* 2018;34(19):3357-3364.
[23]Xuan Z, Li J, Yu J, *et al.* A probabilistic matrix factorization method for identifying lncRNA-disease associations. *Genes* 2019;10(2):126-145.
[24]Yao D, Zhan X, Zhan X, *et al.* A random forest based computational GSMV for predicting novel lncRNA-disease associations. *BMC bioinformatics* 2020;21(1):1-18.
[25]Wu Q W, Xia J F, Ni J C, *et al.* GAERF: predicting lncRNA-disease associations by graph auto-encoder and random forest. *Briefings in bioinformatics* 2021;22(5):1-18.
[26]Li J, Zeng X, Dou Y, *et al.* LADstackING: Stacking Ensemble Learning-based Computational model for Predicting Potential LncRNA-disease Associations. In:*2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) , Houston, United States: IEEE.* 2021;177-182.
[27]Xuan P, Cao Y, Zhang T, *et al.* Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. *Frontiers in genetics* 2019;10:416-426.
[28]Xuan P, Jia L, Zhang T, *et al.* LDAPred: a method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs. *International journal of molecular sciences* 2019;20(18):4458-4472.
[29]Wei H, Liao Q, Liu B. iLncRNAdis-FB: identify lncRNA-disease associations by fusing biological feature blocks through deep neural network. *IEEE/ACM transactions on computational biology and bioinfor-400 matics* 2021;18(5):1946-1957.
[30]Xuan P, Zhan L, Cui H, *et al.* Graph triple-attention network for disease-related lncRNA prediction. *IEEE Journal of Biomedical and Health Informatics* 2022;1-11
[31]Yang J, Ma S, Jiang X. Predicting LncRNA-disease association by autoencoder and rotation forest. In:*2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, United States: IEEE.* 2019;159-164.
[32]Xuan P, Gong Z, Cui H, *et al.* Fully connected autoencoder and convolutional neural network with attention-based method for inferring disease-related lncRNAs. *Briefings in Bioinformatics* 2022;bbac089.
[33]Sheng N, Cui H, Zhang T, *et al.* Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA-disease association prediction. *Briefings in Bioinformatics* 2021;22(3):bbaa067.
[34]Du B, Tang L, Liu L, *et al.* Predicting LncRNA-disease association based on generative adversarial network. *Current Gene Therapy* 2021;21:1-12.
[35]Yang Q, Li X. BiGAN: LncRNA-disease association prediction based on bidirectional generative adversarial network. *BMC bioinformatics* 2021;22(1):1-17.
[36]Xuan P, Pan S, Zhang T, *et al.* Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells* 2019;8(9):1012-1027.
[37]Wu Q W, Cao R F, Xia J, *et al.* Extra Trees Method for Predicting LncRNA-Disease Association Based on Multi-layer Graph Embedding Aggregation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2021
[38]Fan Y, Chen M, Pan X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Briefings in Bioinformatics* 2022;23(1): bbab361.
[39]Wang L, Zhong C. gGATLDA: lncRNA-disease association prediction based on graph-level graph attention network. *BMC bioinformatics* 2022;23(1): 1-24.
[40]Lan W, Wu X, Chen Q, *et al.* GANLDA: graph attention network for lncRNA-disease associations prediction. *Neurocomputing* 2022;469: 384-393.
[41]Bo D, Wang X, Shi C,*et al.* Beyond low-frequency information in graph convolutional networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* 2021; 35(5):3950-3957.
[42]Chien E, Peng J, Li P, *et al.* Adaptive universal generalized pagerank graph neural network. In:*International Conference on Learning Representations, Virtual Event, Austria: ICLR.* 2021.
[43]Zhu J, Yan Y, Zhao L, *et al.* Beyond homophily in graph neural networks. Current limitations and effective designs. *Advances in Neural Information Processing Systems.* 2020; 33: 7793-7804.
[44]Bao Z, Yang Z, Huang Z, *et al.* LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic acids research* 2019;47(D1): D1034-D1037.
[45]Li J H, Liu S, Zhou H, *et al.* starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research* 2014;2(D1):D92-D97.
[46]Huang Z, Shi J, Gao Y, *et al.* HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic acids research* 2019;47(D1):D1013-D1017.
[47]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):1644-1650.
[48]Chen X, Y an CC, Luo C, *et al.* Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Scientific reports* 2015;5:11338
[49]Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Advances in neural information processing systems* 2017;30.
[50]Sun Y, Han J, Yan X, *et al.* Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* 2011;4(11): 992-1003.
[51]Velickovic P, Cucurull G, Casanova A, *et al.* Graph attention networks. *Stat* 2017;1050: 20.
[52]Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* 2013;4(2): 627.
[53]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(3): e0118432.
[54]Wang X, Ji H, Shi C, *et al.* Heterogeneous graph attention network. In: *The world wide web conference, San Francisco, United States: IW3C2.* 2019, 2022-2032.
[55]Fu X, Zhang J, Meng Z, *et al.* Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In:*Proceedings of The Web Conference 2020, Taiwan, China: IW3C2.* 2020,2331-2341.
[56]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 transactions on computational biology and bioinformatics* 2018;16(2):688-693.
[57]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 research* 2021;49(D1): D1251-D1258.
[58]Hou X, Wen J, Ren Z, *et al.* Non-coding RNAs: new biomarkers and therapeutic targets for esophageal cancer. *Oncotarget* 2017; 8(26): 43571.