BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

# BiGAN: LncRNA-disease association prediction based on bidirectional generative adversarial network

Qiang Yang[1] and Xiaokun Li[1,2*]

*Correspondence:
li.xiaokun@163.com
[1] School of Electronic Engineering, Heilongjiang University, Harbin 150080, China
Full list of author information is available at the end of the article

## Abstract

**Background:** An increasing number of studies have shown that lncRNAs are crucial for the control of hormones and the regulation of various physiological processes in the human body, and deletion mutations in RNA are related to many human diseases. LncRNA- disease association prediction is very useful for understanding pathogenesis, diagnosis, and prevention of diseases, and is helpful for labelling relevant biological information.

**Results:** In this manuscript, we propose a computational model named bidirectional generative adversarial network (BiGAN), which consists of an encoder, a generator, and a discriminator to predict new lncRNA-disease associations. We construct features between lncRNA and disease pairs by utilizing the disease semantic similarity, lncRNA sequence similarity, and Gaussian interaction profile kernel similarities of lncRNAs and diseases. The BiGAN maps the latent features of similarity features to predict unverified association between lncRNAs and diseases. The computational results have proved that the BiGAN performs significantly better than other state-of-the-art approaches in cross-validation. We employed the proposed model to predict candidate lncRNAs for renal cancer and colon cancer. The results are promising. Case studies show that almost 70% of lncRNAs in the top 10 prediction lists are verified by recent biological research.

**Conclusion:** The experimental results indicated that our proposed model had an accurate predictive ability for the association of lncRNA-disease pairs.

**Keywords:** LncRNA-disease association, LncRNA sequence similarity, Disease semantic similarity, Bidirectional generative adversarial network

## Background

Conventional molecular biology assumes genetic information is stored primarily in the sequences of genes that code for proteins [1]. However, an increasing number of studies have revealed that protein-coding genes account for only a tiny fraction of human genome (approximately 1.5%), while the other human genes are not involved in the protein-coding sequence [2–5]. In addition, in recent years, an increasing amount of experimental evidence has demonstrated that in most biological processes non-coding

RNAs (ncRNAs) are involved extensively [6, 7]. In particular, long non-coding RNAs (lncRNAs) with a nucleotide sequence length greater than 200 are large and essential non-coding RNAs [8, 9]. Recently, with the improvement of computational power and experimental techniques, thousands of lncRNAs have been discovered, from lower eukaryotes such as paramecia to humans [10]. Although lncRNAs cannot encode proteins, they play important roles in biochemical reactions in the human body, such as protein translation, expression, gene regulation, immune regulation, oncogenesis and tissue development [11]. Currently, there are many accumulated peices of evidence that the association between lncRNAs and diseases is particularly important. Many diseases caused by lncRNAs are complex and difficult to control, such as prostate cancer, colon cancer, Alzheimer's disease, cardiovascular disease, and lung cancer [12–16]. For instance, the oncogenic effect of lncRNA-H19 can be inhibited by the under-regulation of renal carcinoma cells [17]. Therefore, it is essential to predict lncRNA-disease associations. It can help us to understand the biological processes and the molecular mechanisms of human diseases from the perspective of ncRNAs.

In recent years, a large number of computational methods have been proposed to predict lncRNA-disease associations for application in biological experimental verification. These approaches are mainly divided into three categories. The first predicts the correlation between unknown lncRNAs and diseases by sorting out disease similarities, lncRNA similarities, and the association between lncRNAs and diseases based on a random walk. However, if there is no known interaction of relevant lncRNA information on the new disease or no known interaction of relevant disease information on the new lncRNA, it is difficult for these methods to be applied to the relevant association prediction. For example, Sun et al. [18]., restarted the random walk and applyied it to the functional similarity network of lncRNAs. They proposed a computational model called RWRLNCD to detect the associations between diseases and lncRNAs in humans. In addition, Yao et al. [19]., Zhou et al. [20]., also raised a similar calculation approach based on a random walk. However, they focused more on the construction of a heterogeneous network to achieve the purpose of disease association prediction forzz lncRNAs.

The second method utilizes semi-supervised learning methods and machine learning models to extract the feature space between the known lncRNA-disease association and predict the unverified association of the two. In 2013, Chen et al. [22]., created a semi-supervised learning model based on the Laplacian regularized least-squares method (LRLS). In 2016, Lan et al. [21]., blended different data sources and employed a classifier SVM to predict potential interactions between diseases and lncRNAs. Their model solved the problem that the method of LRLS for predicting lncRNA-disease associations was degraded by using two combined classifiers. However, the method proposed by Lan et al. still had great deficiencies in the effective fusion of different lncRNA cores. In 2019, Li et al. [23] proposed a disease gene prioritization based on graph convolutional neural network (PGCN). Their method empolyed end-to-end manner to embedding the hetergeneous network of diseases and genes.

The third category constructs a correlation matrix of lncRNA-disease pairs based on known experimental data. The sequence similarity between lncRNAs and semantic similarity between diseases are integrated to find their associations with genes, to obtain the potential association between lncRNAs and diseases. Such an approach, however, relies

heavily on extensive genetic records. As a result, these models are greatly limited in their predictive tasks. For example, in 2012, Chen et al. [24]., predicted lncRNA-disease associations based on the close relationship between genes through the association between diseases and lncRNA genes. However, the identification of lncRNA position characteristics is still a tough task. In 2015, a computational model called KATZ was proposed by Chen et al. [25]. The main idea of KATZ is to integrate disease semantic similarities and the expression profile of lncRNAs. However, the low expression level of lncRNA inhibited the function of this model.

In the past decade, deep learning has become one of the most popular subjects in scientific research. Many deep learning models have been created by scholars and applied in various fields. At the same time, great success has been achieved in the field of biology. In particular, computational models based on neural networks have made outstanding contributions to the task of prediction [26]. As a neural network, the auto-encoder can learn input data through unsupervised learning, strongly represent potential features, effectively reduce sample noise, and randomly generate data similar to the training data [27].

Therefore, this paper proposed a bidirectional generative adversarial network that uses an encoder and generator to learn high-level features in latent space, and a discriminator to predict lncRNA-disease associations.

## Results

### Parameter settings

In our study, the input length of the BiGAN encoder is 6137 and the output length is 100. The lengths of the generator input and output are opposite to those of the encoder input and output. We applied a fully connected layer and ReLU activation function on each network and employed a cross-entropy function as the loss function. Adam was also used to optimize our model. The number of epochs was set as 5, and the batch size was set as 64 in our predicted model.

### Evaluation metrics

To evaluate the predictive ability of the BiGAN on the association between lncRNA and disease pairs, we validated our proposed model using five-fold and 10-fold cross-validation. Almost all samples were taken as candidates in each cross-validation. Therefore, the distribution closest to the original samples makes the evaluation results highly reliable. We utilized the experimentally verified lncRNA-disease associations as samples, while all the unverified associations of the lncRNA-disease pair were used as candidates. Hence, we could rank each candidate sample based on the predicted score. In the rank list, a threshold was given. Samples with lncRNA- disease association prediction scores above the threshold were considered true positive (TP). For each given threshold, we can find the corresponding TP to determine the true positive ratio (TPR), which is also known as sensitivity. Similarly, we can obtain the false negative (FN) samples among the candidate samples by setting a threshold, and the corresponding false positive ratio (FPR) is also called the 1-specificity. The TPR and the FPR can be calculated as follows:

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

where TP is the number of positive samples, and FN is the number of negative samples whose prediction scores are higher than the threshold but considered as a negative sample. TN is the number of negative samples, and FP is the number of negative samples whose prediction scores are lower than the threshold but considered positive samples.

TPR and FPR denote the proportion of the number of lncRNA-disease association prediction scores over or under a given threshold in the test samples respectively. Therefore, according to the different thresholds, we can plot the receiver operating characteristic (ROC) curve, which is shown in Fig. 1. At the same time, we calculated the area under the ROC curve(AUC) to evaluate the lncRNA-disease association ability of our proposed model [28]. The higher the AUC value is, the better the performance of the BiGAN. When the AUC value of reaches 1, it is considered that the BiGAN can perfectly predict lncRNA- disease associations. When the value is close to 0.5, it is considered to predict the association randomly. To balance the samples of known lncRNA-disease associations and unknown lncRNA-disease associations, we also utilized the precision-recall (PR) curve to estimate our BiGAN [29]. Precision and Recall are defined as follows:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{2}$$

In addition, we utilized statistical parameters to measure the performance of our predicted model, such as the F1-score, accuracy, and Matthews Correlation Coefficient (MCC). The experimental results in the three datasets are shown in Table 1.

### Comparison with other methods

We compared the BiGAN with other four other advanced methods to prove that our model can predict the associations of lncRNA-disease pairs effectively. The three datasets mentioned above were employed as gold standard training sets to evaluate the other

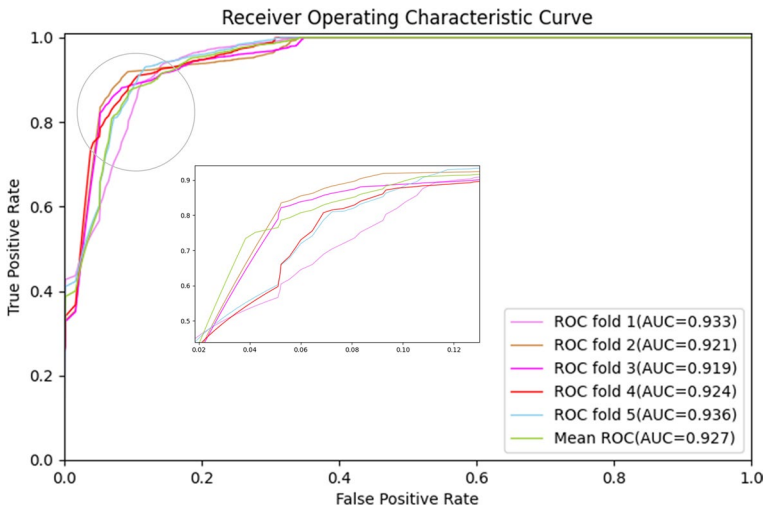


**Fig. 1** ROC curves of the BiGAN in five-fold cross-validation in LncRNADisease dataset

**Table 1** Ten-fold cross-validation results performed by the BiGAN on three datasets

| Dataset | AUC | AUPR | Accuracy | F1-score | MCC |
|---|---|---|---|---|---|
| MNDR | 0.929 | 0.901 | 0.967 | 0.874 | 0.867 |
| LncRNADisease | 0.931 | 0.911 | 0.961 | 0.871 | 0.864 |
| LncRNACancer | 0.934 | 0.905 | 0.979 | 0.873 | 0.864 |

methods. The four methods were the probabilistic prediction model of the association between lncRNAs and disease based on Naive Bayes Classifier [30], the Convolutional Neural Network based on an attention mechanism for the lncRNA genes relationship with diseases (CNNLDA) [31], identifying known lncRNA-disease association by using Topological Information (TILDA) [32], and the web service for discovering lncRNA-disease interaction through mixing multiple biological data resources (LDAP).

As shown in Fig. 2, the AUC (0.931) of the BiGAN is the highest compared to those of the other methods in 10-fold cross-validation on the LncRNADisease dataset. The AUC values of CNNLDA, NBCLDA, TILDA, and LDAP are 0.907, 0.871, 0.817, and 0.806, respectively. The AUC and AUPR values for all methods in five-fold cross-validation are shown in Table 2.

To verify that our prediction model performs well not only in a single dataset, but also in two other datasets. As shown in Fig. 3a, the area enclosed by the ROC curve and the coordinate axis of BiGAN and different models in the dataset Lnc2Cancer in 10-fold cross-validation. In the Lnc2Cancer dataset, except for the fact that the AUC value of

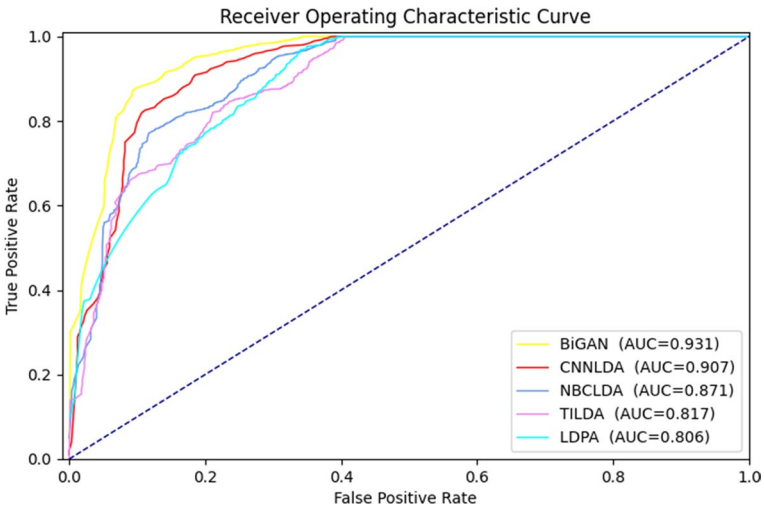


**Fig. 2** ROC curves in 10-fold cross-validation by different methods

**Table 2** The AUC and AUPR values for all methods in five-fold cross-validation

| Five-fold cross-validation | BiGAN | CNNLDA | NBCLDA | TILDA | LDAP |
|---|---|---|---|---|---|
| AUC | 0.927 | 0.914 | 0.821 | 0.815 | 0.776 |
| AUPR | 0.917 | 0.897 | 0.807 | 0.796 | 0.753 |

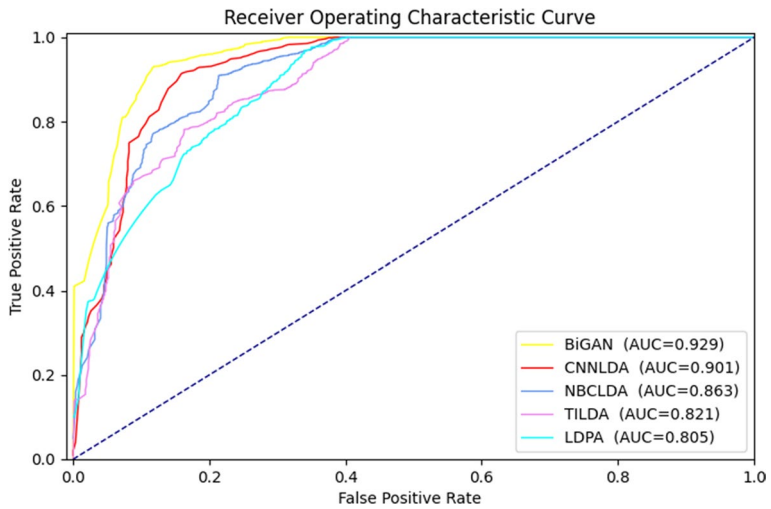


**Fig. 3** Ten-fold cross-validation ROC curves obtained by different methods on the Lnc2Cancer dataset

TILDA was slightly higher than that for LncRNADisease, the AUC values of the models was lower. Most likely, this is because the underlying cancer was more difficult to predict than a common disease. So the performance of most models on Lnc2Cancer dataset was poorer. In the MNDR, the AUC of the BiGAN reached the highest value, 0.934 (Fig. 4). The results in the three datasets demonstrated that the proposed model was not solely capable of efficient prediction of lncRNA-disease association in a specific dataset. As the results show, the BiGAN has strong robustness and generalization ability, which is better than other state-of-the-art models in most datasets.

### Case studies on colon cancer and renal cancer

To demonstrate the ability of the BiGAN to predict the latent association between lncRNAs and diseases, we measured our method based on case studies on the Lnc2Cancer dataset and MNDR dataset.

Colon cancer is one of the most dangerous cancers and one of the main causes of death among humans. The relationship among the codes in the sequence of lncRNAs associated with colon cancer is that these sequences may cause cancer. With the development of cancer research, lncRNAs had become an essential target for colon cancer prevention, diagnosis, and treatment. In the Lnc2Cancer and MNDR datasets, we applied the BiGAN to predict the associations between colon cancer and lncRNAs, and 7 experimentally verified lncRNAs were on the top ten prediction list. ANRIL can suppress the expression of other RNAs in the late phase of the DNA damage response to repair DNA to normal levels [33]. Experimental results show that the control network composed of UCA1 and other RNAs is a potential factor in the treatment of colon cancer [34]. Additionally, the migration ability of colon cancer cells is significantly inhibited and blocked when TUG1 is expressed, and the overexpression of TUG1 may accelerate the cell migration process of colon cancer cells [35].More details are shown in Table 3.

**Table 3** Top ten predicted results between colon cancer and renal cancer by the BiGAN with experimental validation in the literature on Lnc2Cancer dataset

| Colon cancer | | | Renal cancer | | |
|---|---|---|---|---|---|
| **Name of lncRNAs** | **Rank** | **Pubmed ID** | **Name of lncRNAs** | **Rank** | **Pubmed ID** |
| ANRIL | 3 | 23416462 | UCA1 | 5 | 31996265 |
| CCAT1 | 6 | 31039730 | MALAT1 | 1 | 31250518 |
| H19 | 4 | 31602323 | ACTN4 | 6 | Unknown |
| ENST | 8 | Unknown | PVT1 | 3 | 30105850 |
| XIAP-AS1 | 9 | 30892955 | FAL1 | 9 | Unknown |
| P14AS | 7 | Unknown | HOTAIR | 2 | 30105850 |
| GAS5 | 2 | 28722800 | H19 | 7 | 29214011 |
| UCA1 | 5 | 30652355 | RAB31 | 10 | Unknown |
| TUG1 | 1 | 27634385 | NBAT1 | 4 | 31298469 |
| DANCR | 10 | Unknown | MEG3 | 8 | 31071531 |

More than 250 thousand new cases of renal cancer are diagnosed each year, and renal cancer is recognized as one of the top ten common cancers. It is important to find an association between renal cancer progression and the dysregulation of certain lncRNAs. Among all the lncRNA candidates predicted by the BiGAN as being associated with renal cancer, 7 lncRNAs were among the top 10 in the predicted list, (MALAT1 1st, HOTAIR 2nd, PVT1 3rd, NBAT1 4th, UCA1 5th, H19 6th, and MEG3 7th). MALAT1 reduces the expression of miR-203 to promote the expression of BIRC5 and accelerate the occurrence and development of renal cell carcinoma [36]. The long non-coding RNA HOTAIR accelerates $\alpha$-2, 8-salivary transferases in renal cell carcinoma malignancies by wetting pre-miniaturized microRNA-124 [37]. By down-regulating miR-16-5p, lncRNA PVT1 promoted the invasion, proliferation, and epithelial-mesenchymal transformation of renal cell carcinoma cells [38].

According to the above description, the BiGAN can achieve good performance in predicting unknown association between lncRNA-disease pairs. Therefore, our approach can be widely used for predicting unverified lncRNA-disease associations recorded in the databases. All candidate associations are prioritized and the predicted results can be used for future research and experimental validation.

## Discussion

In the experiment, we integrated the comprehensive similarity vectors of known lncRNA-disease correlations to present their relationship as the first step. Then, the BiGAN model was built to predict the unverified associations between lncRNAs and diseases by learning high-level features in latent space from the similarity vectors. Although the BiGAN seems to outperform than other advanced methods in the above evaluation, it still has some room for improvement.

Based on an auto-encoder, the BiGAN can automatically recognize the comprehensive similarity characteristics of lncRNAs and diseases, eliminate noise, and reduce dimensions. It always learns the annotated biological patterns perfectly. However, we found that our proposed model did not achieve the best performance in predicting the association between lncRNAs and diseases. In our research, two main factors may affect the

results. On the one hand, the performance of the BiGAN strongly depends on the similarity eigenvectors which are computed through handcrafted measurements. However, it is not easy to extract the similarity features from high-dimensional data by using these methods. On the other hand, the structure of the BiGAN is based on an auto-encoder whose main idea is to compress the features into low dimensions and learn the latent representation. Thus, we assume that the BiGAN cannot share and propagate information perfectly in each network layer.

In this study, we did not further consider whether the performance would be impacted by the values of the parameters that were set as default. In fact, the parameter settings are significantly important to a certain model because suitable parameters can help the model learn privileged information from the eigenvectors, particularly for complex associated features. In recent years, heteroscedastic dropout has been one of the best regularization techniques for controlling deep neural networks to absorb privileged information. Thus, we will take what has been discussed above as our future work to improve the prediction ability of lncRNA-disease associations.

## Conclusions

In this manuscript, we introduce an unsupervised learning lncRNA- disease association prediction framework called BiGAN. The model includes three main parts, a feature extractor based on similarity algorithm, a bidirectional generator based on autoencoder, and a discriminator that jointly discriminates data and latent space features. We integrated lncRNA sequence similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity to mine the high-level representation of the potential space between lncRNAs and diseases. Ultimately, the BiGAN can effectively predict the associations between lncRNAs based on the latent relationship of the integrated similarity vectors. In 10-fold cross-validation and five-fold cross-validation, our AUC values were 0.931 and 0.927, respectively, indicating the effectiveness of our prediction model. We also compared our model with other state-of-the-art methods, and the results revealed that the BiGAN was superior to other advanced methods. Additionally, we conducted case studies on colon cancer and renal cancer. The case results showed that our proposed model had an accurate predictive ability for the association of lncRNA-disease pairs.

## Methods

### Datasets

To better train our BiGAN model, we collected three experimentally validated datasets from MNDR v3.0, Lnc2Cancer, and LncRNADisease. Below is a brief description of the datasets used.

The first dataset is from the mammalian ncRNA disease repository (MNDR) with coverage and annotation proposed by Lin et al. 24 August 2020 [39]. We extracted association information about human lncRNA-disease pairs in MNDR, consisting of two datasets. One of the databases is experimentally verified association information, covering 742 human diseases, 25,494 human lncRNAs, and 39,783 lncRNA-disease associations, which can be used as a training set. The other dataset is the association

information of predicted lncRNA-disease pairs, covering 231 human diseases, 17,713 human lncRNAs, and 52,144 pieces of association information, which can be used as a validation set.

The second dataset was released on 8 January 2019, and contains experimentally validated lncRNA-disease correlations downloaded from LncRNADisease V2.0 [40]. We also collected another special ncRNA dataset named circRNA whose sequences were sufficiently long (>200) in this dataset. After removing the lncRNA disease pairs that were not labelled with IDs and that lacked features, we deleted duplicate samples describing the lncRNA-disease relationship according to known experimental evidence. From this, we obtained 205,959 interaction associations for 529 human diseases and 19,166 lncRNAs. In addition, 823 circRNAs and 529 human diseases, and 1004 interaction associations were included. This dataset contains more comprehensive information than the other two datasets.

The third dataset was published on 30 June 2020, and contains experimentally proven lncRNA-disease correlations which were based on the Lnc2Cancer V3.0 dataset [41]. After removing the lncRNA disease pairs that were not labelled with IDs and that lacked characteristics, we deleted duplicate samples describing the lncRNA-disease relationships according to known experimental evidence. As a result, 216 human diseases,2659 human lncRNAs, and 9254 human lncRNA-disease interaction associations were obtained. Compared with lnc2cancer2.0 published in 2018, the number of diseases have increased by 51 and the number of lncRNAs increased by more than 1000, and the lncRNA-disease association nearly doubled. This allows us to collect enough data to learn the features of the latent space between lncRNAs and diseases in training the BiGAN.

#### *LncRNA-disease association*

According to the sorted dataset, the interaction information between diseases and lncRNAs was constructed into a matrix $A \in R^{nd \times nl}$, where the columns represent lncRNAs and the rows represent disease. If there was an experimentally verified lncRNA-disease association, the value of A in the matrix was set to 1. Otherwise, the value was set to 0, as shown in Fig. 5A.

#### *LncRNA sequence similarity*

An increasing number of studies have shown that similar pathologies between two different diseases may be linked with two similar lncRNAs. Therefore, one of the important characteristics of lncRNA-disease association prediction is the similarity between different lncRNAs. Between any two strings, the Levenshtein distance is the minimum cost required for a single word of one string to be converted to the other string after insertion, deletion, or replacement. To investigate the deeper similarity between lncRNAs, we used the Levenshtein distance to calculate the similarity between two lncRNAs. We set the editing cost as 2, and the cost for deletion and insertion as 1. The similarity between the *i*th lncRNA and the *j*th lncRNA is$L_{sim}(l_i, l_j) \in R^{nl \times nl}$, and it can be calculated as follows:

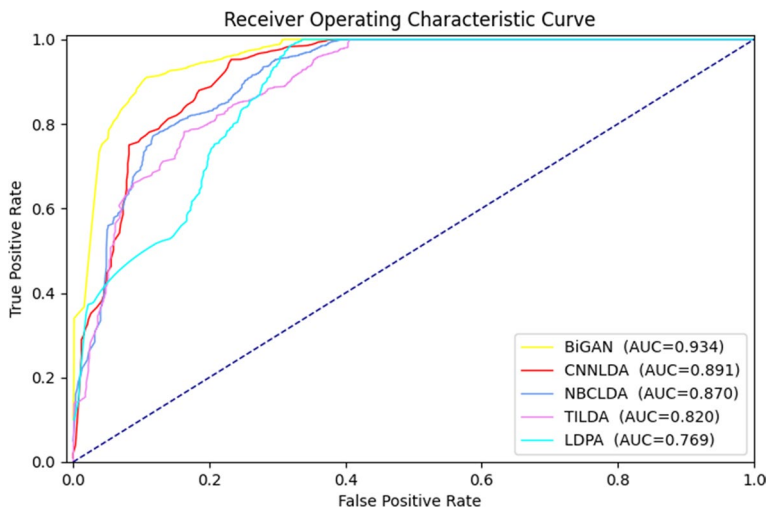


**Fig. 4** Ten-fold cross-validation ROC curves obtained by different methods on the MNDR dataset

$$L_{sim} = 1 - \frac{x}{len(l_i) + len(l_j)} \tag{3}$$

where $x$ represents the minimum cost required to convert one lncRNA sequence into another and *len* represents the sequence length of lncRNA.

#### *Disease semantic similarity*

In 2010, Schlicker et al. found that the more similar the disease phenotype was, the more similar the gene dysfunction [42]. Gene Ontology annotations provide a way to obtain the semantic similarity of genes [43]. Thus, some researchers employ directed acyclic graphs (*DAGs*) to represent diseases. Additionally, the Jaccard correlation coefficient has been used to calculate the functional similarity of diseases. We applied DAGs to this study to calculate semantic similarity scores for diseases. Let $D_{sim} \in R^{nd \times nd}$ be the disease similarity between the ith disease and the *j*th disease. It can be calculated as follows:

$$D_{sim}(d_i, d_j) = \frac{\sum_{x \in G_{d_i} \cap G_{d_j}} (SVD_i(x) + SVD_j(x))}{\sum_{x \in G_{d_i}} SVD_i(x) + \sum_{x \in G_{d_j}} SVD_j(x)} \tag{4}$$

where $G_{d_i}$ represents disease $d_i$ in *DAGs* , $G_{d_j}$represents $d_j$ disease in *DAGs* . Compare disease $i$ and disease $j$, $SVD_i(x)$ denotes the disease semantic value of $x \in G_{d_i}$ , and $SVD_j(x)$ denotes the disease semantic value of $x \in G_{d_j}$ .We can calculate the semantic value of a disease $d$ by using the following equation:

$$SVD(x) = \begin{cases} \max\{\mu \cdot SVD(d')\}, & \text{if} \quad x \neq d \\ 1, & \text{Otherwise} \end{cases} \tag{5}$$

where $d' \in$ children of $d$, and $\mu$ represents the factor of semantic contribution. According to previous research, we set it to 0.5 [44].

#### *Gaussian interaction profile kernel similarity*

Similar lncRNAs may be associated with different diseases that have similar pathological characteristics, and vice versa. Based on this assumption, the kernel similarity between lncRNAs and diseases can be calculated by the Gaussian interaction profile (GIP). The GIP kernel similarities were computed based on the lncRNA-disease interaction matrix obtained from the LncRNADisease dataset. The GIP similarities $GKL(l_i, l_j)$ of lncRNAs can be computed as follows:

$$GKL(l_i, l_j) = \exp(-\lambda||A(l_i) - A(l_j)||^2) \tag{6}$$

where $A(l_i)$ and $A(l_j)$ represent the $i$th and $j$th columns information in the association matrix $A$. Let $\lambda$ be a parameter that can control the width of the kernel boundary and is represented by the average number of diseases associated with each lncRNA, which is defined as follows:

$$\lambda = \frac{1}{\frac{1}{nl}\sum_{i=1}^{nl}||A(l_i)||^2} \tag{7}$$

where $nl$ denotes the number of lncRNAs.

Similarly, we can obtain the GIP kernel similarity of disease $d_i$ and disease $d_j$ as follows:

$$GKD(d_i, d_j) = \exp(-\lambda||A(d_i) - A(d_j)||^2) \tag{8}$$

where $A(d_i)$ and $A(d_j)$ denote the $i$th and $j$th rows information in the lncRNA-disease association matrix $A$. Let $\lambda$ be a parameter that can control the width of the kernel boundary and is represented by the average number of lncRNAs associated with each disease, which can be calculated as follows:

$$\lambda = \frac{1}{\frac{1}{nl}\sum_{i=1}^{nd}||A(d_i)||^2} \tag{9}$$

where $nd$ denotes the number of diseases.

#### *Integrated similarity*

From the above,the lncRNAs sequence similarity, the semantic similarity of diseases, and the GIP kernel similarity of lncRNAs and diseases were gathered. We obtained the integrated similarity of lncRNA ($Ls$) and integrated similarity of diseases ($Ds$), (Fig. 5B), and the calculation formula is shown as follows:

$$Ls(l_i, i_j) = \frac{L_{sim}(l_i, l_j) + GKL(l_i, l_j)}{2} \tag{10}$$

$$Ds(d_i, d_j) = \frac{D_{sim}(d_i, d_j) + GKD(d_i, d_j)}{2} \tag{11}$$

The disease similarity vector for disease $d_i$ contains the similarity values of all other diseases to $d_i$. Additionally, the lncRNA similarity vector for lncRNA $l_i$ includes the similarity values of all other lncRNAs to $l_i$. Therefore, we concatenated these similarity vectors for the corresponding lncRNA-disease pair to generate large eigenvectors of size

$nd + nl$, where the number of diseases and lncRNAs was $nd$ and $nl$, as shown in Fig. 5C. There were $nd \times nl$ samples altogether, each corresponding to a lncRNA-disease pair.

## BiGAN

In 2018, Chen et al. proposed using linear-based principal component analysis (PCA) to obtain the traits of GIP kernel similarity [45]. However, the potential lncRNA-disease correlation features were difficult to mine. As a nonlinear generalization of PCA, an auto-encoder is an unsupervised neural network model that mainly includes an encoder and decoder. This special neural network has two advantages in dealing with the features of lncRNA-disease associations [46]. One is that auto-encoders are good at learning biological patterns that are annotated. Second, they can automatically recognize the comprehensive similarity characteristics of lncRNAs and diseases, eliminate noise, and reduce dimensions. This can solve the problem that features extracted from large datasets may produce considerable noise. To further study the model of unsupervised learning, we developed a novel generative adversarial network model inspired by the auto-encoder.

### *The main framework of the BiGAN*

In this study, we propose using the bidirectional generative adversarial network(BiGAN) model to complete the task of predicting the association of lncRNA-disease pairs. BiGAN consists of an encoder, a generator, and a discriminator, the main framework of which is shown in Fig. 6. The BiGAN encoder can map the original data point $x$ to the latent representation $z$. The BiGAN generator will capture the feature in the latent space to generate a new lncRNA-disease association. The BiGAN discriminator not only

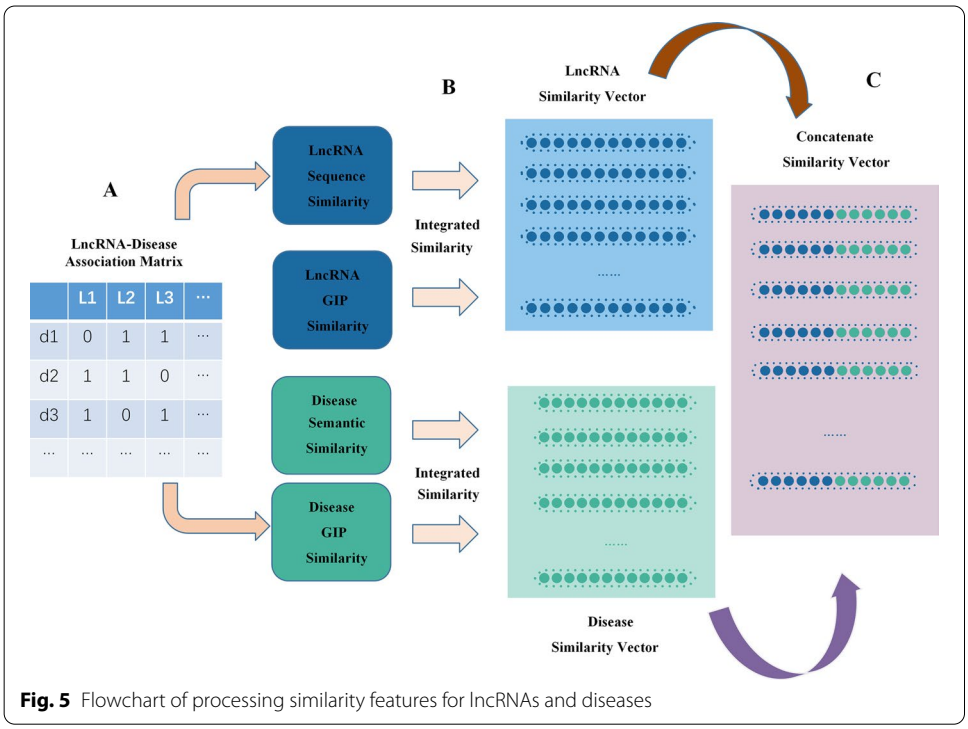


**Fig. 5** Flowchart of processing similarity features for lncRNAs and diseases

discriminates in the traditional data space ($x$ versus $G(z)$), but also discriminates in the joint data and latent space (($x$, $E(x)$) versus ($G(z)$, $z$)). The latent component is both an encoder output $E(x)$ and a generator input $z$.

We can clearly see that the encoder and the generator cannot "communicate" with each other directly. However, the encoder and generator will learn to reverse each other through the joint probability distribution. In other words, $E(G(z))$ and $G(E(x))$ can be computed to fool the BiGAN discriminator. In our model, an encoder $E : \Omega_X \rightarrow \Omega_Z$ and a generator $G : \Omega_Z \rightarrow \Omega_X$ are trained at the same time. The BiGAN encoder includes a distribution $P_E(Z|X) = \sigma(Z - E(X))$ mapping data points $x$ into a latent feature space of the generator. The BiGAN generator includes a distribution $Q_G(X|Z) = \sigma(X - G(Z))$ extracting randomly sampled noise from the encoder to generat new lncRNA-disease associations. The discriminator will take input from the latent space in to predict the distribution of $P_D(Y|X, Z)$, where the value of $Y$ is equal to 0 if $X$ is from the output of generator $(G(z), z \sim p_z)$, and the value of $Y$ is 1 if $X$ is sampled from the encoder data distribution $p_x$. Thus, we can define a minimax objective to replace the BiGAN training objective.

$$\min_{G,E} \max_{D} V(D, E, G) \tag{12}$$

where $V(D, E, G)$ can be computed based on the following formulas:

$$V(D, E, G) = E_{X \sim p_X}[logD(X, E(X))] + E_{Z \sim p_Z}[log(1 - D(G(Z), Z))] \tag{13}$$

$$logD(X, E(X)) = E_{Z \sim p_E(\cdot|X)}[logD(X, Z)] \tag{14}$$

$$log(1 - D(G(Z), Z)) = E_{X \sim p_G(\cdot|Z)}[log(1 - D(X, Z))] \tag{15}$$

In contrast to other advanced unsupervised computing models, the BiGAN can learn the gradient information perfectly, so as to ensure the correct weight allocation.

#### *More details of the encoder, generator, and discriminator*

*Encoder* In the similarity eigenvectors, each lncRNA contains the similarity information and position information of all other lncRNAs. Likewise, each disease contains information about the similarity and position of all the other diseases. As mentioned above, the BiGAN encoder is one of the two parts of an auto-encoder. The main functions of the encoder are to compress data, eliminate noise, and learn the features of the latent space. We take the similarity feature vectors of the samples as input so that the encoder can fully learn the parameters of the similarity vectors. In this way, the encoder can effectively map the data points into the latent feature space. The structure of BiGAN encoder is shown in Fig. 7A. The encoder is composed of three fully connected layers of the neural network. We can compute the output of each layer with the following formula:

$$E(x) = W^E x + b^E \tag{16}$$

where $x$ denotes the similarity features of lncRNA-disease pairs. $W^E$ and $b^E$ represent the encoder weights and bias, respectively.

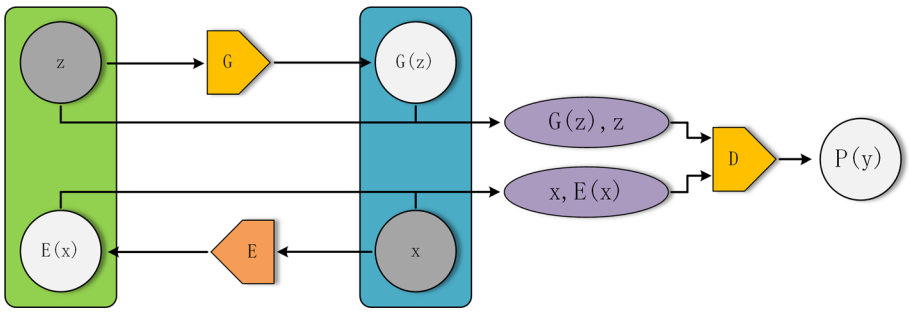


**Fig. 6** The main framework of BiGAN

The dimension of the similarity eigenvectors between the lncRNA and disease will be compressed into a low-dimensional vector after passing through each layer in the encoder. A trained encoder can predict the feature representations of data by capturing semantic attributes. The dense information of compressed low-dimensional vectors is more conducive to learning the mapping relationship of the latent space. To mine the representation of latent space more effectively, we decided to set the number of neurons in the final layer to 100. We employed ReLU as the activation function in the BiGAN model, and it can be defined as follows:

$$ReLU(y) = \begin{cases} y & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{17}$$

In addition, the encoder will randomly sample noise $z$ in distribution $P_E(Z|X) = \sigma(Z - E(X))$ and output latent features $E(x)$ during training. Ultimately, we can obtain many data pairs $(x, E(x))$.

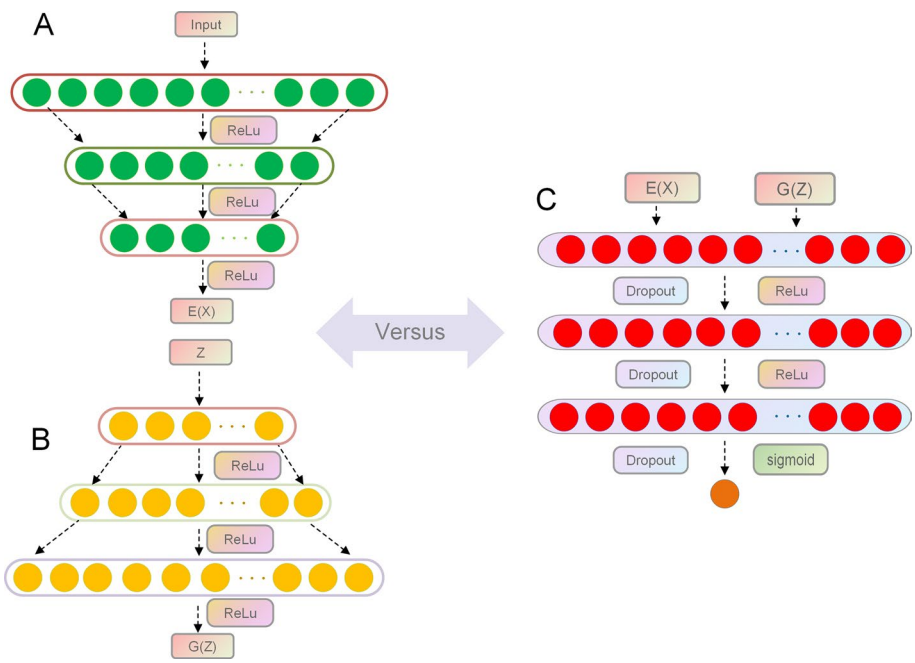


**Fig. 7** The structure of encoder, generator, and discriminator

*Generator* In most generative adversarial network(GAN) models, the role of the generator is to learn the features of the original data and generate new data based on the learned characteristics. However, in the BiGAN model, the generator takes randomly sampled noise as input. As shown in Fig. 7B, the generator is similar to the encoder in that it has the same network structure. The output of the generator is calculated as follows:

$$G(z) = W^G z + b^G \tag{18}$$

where $z$ is the feature of the latent space. $W^G$ and $b^G$ denote the weights and bias of the generator, respectively.

However, each layer in the generator increases the dimension of the potential representation and the final output dimension is the same as the original similarity feature vector dimension. Next, the representation with noise is decoded by the generator, and new lncRNA-disease associations are generated. Then, we can obtain a series of data pairs(G(z),z).

*Discriminator* The two data pairs mentioned above are taken as inputs to fool the discriminator. The discriminator discrimines whether the input data are real. If the discriminator thinks the data pairs come from the encoder, will be set as 1. If the discriminator thinks data pairs come from the generator, it will be set as 0. The structure of the discriminator is shown in Fig. 7C, where the sigmoid function is defined as follows:

$$sigmoid(\theta) = \frac{1}{1 - e^{(-\theta)}} \tag{19}$$

where $\theta$ is the input of the sigmoid function.

The BiGAN encoder has a strong representation learning ability to learn the latent association between lncRNAs and diseases. The BiGAN generator will extract the features of the joint data and latent space to generate new lncRNA—disease associations. Finally, $z = E(G(z))$ and $x = G(E(x))$ are determined through a union probability distribution to arrive at a bidirectional structure. And you can see the concrete proof in the study of Jeff et al. According to our experiment, the BiGAN is an unsupervised feature learning model with strong robustness and representational learning ability. Compared with other computing models, the BiGAN performs remarkably well.

**Abbreviations**

BiGAN: Bidirectional generative adversarial network; LRLS: Laplacian regularized least square method; LNCSIM: Functional similarity models of lncRNAs; DAGs: Directed acyclic graphs; TPR: True positive rate; FPR: False positive rate; ROC: Receiver operating characteristic; AUC: Areas under ROC curve; AUPR: Areas under precision-recall curve.

**Acknowledgements**

We are grateful for the anonymous suggestions that helped to improve the paper in quantity.

**Authors' contributions**

QY conceived the study. QY and XKL developed the method. QY implemented the algorithms and collected the data. QY and XKL performed the data analyses. QY wrote the manuscript. Both authors have read and approved the manuscript.

**Funding**

The project is supported by the National Natural Science Foundation of China (Nos. 81273649, 61501132, 61672181), the Natural Science Foundation of Heilongjiang Province (Nos. LH2019F049, LH2019A029), the China Postdoctoral Science Foundation (No. 2019M650069), the Research Funds for the Central Universities (No. 3072019CFT0603), the Fund for Young Innovation Team of Basic Scientific Research in Heilongjiang Province (No. RCYJTD201805), the Heilongjiang Basic Scientific Research and Technological Innovation Fund (No. KJCX201805), the Young Eagles Plan (2020CYJBGX0057, 2020CYJBGX0353), the Foundation Items: Innovation fund for smes (No. 2017FF1GJ023), and the Patent Advantage Demonstration Enterprise Fund (No. 2017YBQCZ029).The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Availability of data and materials**
All the data used are collected from the public datasets below. The LncRNAdisease database can be downloaded from http://www.cuilab.cn/lncrnadisease. The Lnc2Cancer database can be downloaded from http://bio-bigdata.hrbmu.edu.cn/lnc2cancer/download.html. The MNDR database can be download from http://www.rna-society.org/mndr/download.html. The source code is available at https://github.com/TomasYang001/BiGAN-lncRNA-disease-associations-prediction.git

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
We declare that we have no conflicts of interests.

**Author details**
[1]School of Electronic Engineering, Heilongjiang University, Harbin 150080, China. [2]Postdoctoral Program of Heilongjiang Hengxun Technology Co., Ltd., Harbin 150090, China.

Received: 5 March 2021 Accepted: 15 June 2021
Published online: 30 June 2021

**References**

1. Yanofsky C. Establishing the triplet nature of the genetic code. Cell. 2007;128(5):815–8. https://doi.org/10.1016/j.cell.2007.02.029.
2. Kapranov P, Cheng J, Dike S, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007;316(5830):1484–8. https://doi.org/10.1126/science.1138341.
3. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: regulators of disease. J Pathol. 2010;220(2):126–39. https://doi.org/10.1002/path.2638.
4. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–51. https://doi.org/10.1016/j.cell.2013.06.009.
5. Bertone P. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004;306(5705):2242–6. https://doi.org/10.1126/science.1103388.
6. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482(7385):339–46. https://doi.org/10.1038/nature10887.
7. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43(6):904–14. https://doi.org/10.1016/j.molcel.2011.08.018.
8. Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071–6. https://doi.org/10.1038/nature08975.
9. Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA. 2009;106(28):11667–72. https://doi.org/10.1073/pnas.0904715106.
10. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–7. https://doi.org/10.1038/nature07672.
11. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009;23(13):1494–504. https://doi.org/10.1101/gad.1800909.
12. Chakravarty D, Sboner A, Nair SS, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun. 2014;5:5383. https://doi.org/10.1038/ncomms6383.
13. He X, Tan X, Wang X, et al. C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion. Tumour Biol. 2014;35(12):12181–8. https://doi.org/10.1007/s13277-014-2526-4.
14. Tan L, Yu JT, Hu N, Tan L. Non-coding RNAs in Alzheimer's disease. Mol Neurobiol. 2013;47(1):382–93. https://doi.org/10.1007/s12035-012-8359-5.
15. Klattenhoff CA, Scheuermann JC, Surface LE, et al. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell. 2013;152(3):570–83. https://doi.org/10.1016/j.cell.2013.01.003.
16. Zhang EB, Yin DD, Sun M, et al. P53-regulated long non-coding RNA TUG1 affects cell proliferation in human non-small cell lung cancer, partly through epigenetically regulating HOXB7 expression. Cell Death Dis. 2014;5(5):e1243. https://doi.org/10.1038/cddis.2014.201.
17. Wang L, Cai Y, Zhao X, et al. Down-regulated long non-coding RNA H19 inhibits carcinogenesis of renal cell carcinoma. Neoplasma. 2015;62(3):412–8. https://doi.org/10.4149/neo_2015_049.
18. 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. https://doi.org/10.1039/c3mb70608g.
19. Yao Q, Wu L, Li J, et al. Global prioritizing disease candidate lncRNAs via a multi-level composite network. Sci Rep. 2017;7:39516. https://doi.org/10.1038/srep39516.
20. Zhou M, Wang X, Li J, et al. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Mol BioSyst. 2015;11(3):760–9. https://doi.org/10.1039/c4mb00511b.

21. Lan W, Li M, Zhao K, et al. LDAP: a web server for lncRNA-disease association prediction. Bioinformatics. 2017;33(3):458–60. https://doi.org/10.1093/bioinformatics/btw639.
22. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013;29(20):2617–24. https://doi.org/10.1093/bioinformatics/btt426.
23. Li Y, Kuwahara H, Yang P, Song L, Gao X. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. Biorxiv. 2019. https://doi.org/10.1101/532226
24. Chen G, Wang Z, Wang D, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41(Database issue):D983–6. https://doi.org/10.1093/nar/gks1099.
25. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. Sci Rep. 2015;5:16840. https://doi.org/10.1038/srep16840.
26. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32(12):i121–7. https://doi.org/10.1093/bioinformatics/btw255.
27. Suk HI, Lee SW, Shen D. Alzheimer's disease neuroimaging initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct. 2015;220(2):841–59. https://doi.org/10.1007/s00429-013-0687-3.
28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747.
29. 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. https://doi.org/10.1371/journal.pone.0118432.
30. Yu J, Ping P, Wang L, Kuang L, Li X, Wu Z. A novel probability model for LncRNA-disease association prediction based on the Naïve Bayesian classifier. Genes (Basel). 2018;9(7):345. https://doi.org/10.3390/genes9070345.
31. Xuan P, Cao Y, Zhang T, Kong R, Zhang Z. Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. Front Genet. 2019;10:416. https://doi.org/10.3389/fgene.2019.00416.
32. Ping P, Wang L, Kuang L, Ye S, Iqbal MFB, Pei T. A novel method for LncRNA-disease association prediction based on an lncRNA-disease association network. IEEE/ACM Trans Comput Biol Bioinform. 2019;16(2):688–93. https://doi.org/10.1109/TCBB.2018.2827373.
33. Wan G, Mathur R, Hu X, et al. Long non-coding RNA ANRIL (CDKN2B-AS) is induced by the ATM-E2F1 signaling pathway. Cell Signal. 2013;25(5):1086–95. https://doi.org/10.1016/j.cellsig.2013.02.006.
34. Cui M, Chen M, Shen Z, Wang R, Fang X, Song B. LncRNA-UCA1 modulates progression of colon cancer through regulating the miR-28-5p/HOXB3 axis. J Cell Biochem. 2019. https://doi.org/10.1002/jcb.27630.
35. Zhai HY, Sui MH, Yu X, et al. Overexpression of long non-coding RNA TUG1 promotes colon cancer progression. Med Sci Monit. 2016;22:3281–7. https://doi.org/10.12659/msm.897072.
36. Zhang H, Li W, Gu W, Yan Y, Yao X, Zheng J. MALAT1 accelerates the development and progression of renal cell carcinoma by decreasing the expression of miR-203 and promoting the expression of BIRC5. Cell Prolif. 2019;52(5):e12640. https://doi.org/10.1111/cpr.12640.
37. Pan Y, Wu Y, Hu J, Shan Y, Ma J, Ma H, Qi X, Jia L. Long Noncoding RNA HOTAIR promotes renal cell carcinoma malignancy through alpha-2, 8-Sialyltransferase 4 by sponging MicroRNA-124. Cell Prolif. 2018;51(6):e12507. https://doi.org/10.1111/cpr.12507.
38. Pan Y, Wu Y, Hu J, et al. Long noncoding RNA HOTAIR promotes renal cell carcinoma malignancy through alpha-2, 8-sialyltransferase 4 by sponging microRNA-124 [published correction appears in Cell Prolif. 2020 Jul;53(7):e12873]. Cell Prolif. 2018;51(6):e12507. https://doi.org/10.1111/cpr.12507
39. Ning L, Cui T, Zheng B, et al. MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Res. 2021;49(D1):D160–4. https://doi.org/10.1093/nar/gkaa707.
40. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. 2019;47(D1):D1034–7. https://doi.org/10.1093/nar/gky905.
41. Gao Y, Wang P, Wang Y, et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. Nucleic Acids Res. 2019;47(D1):D1028–33. https://doi.org/10.1093/nar/gky1096.
42. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of gene ontology terms. Bioinformatics. 2010;26(18):i561–7. https://doi.org/10.1093/bioinformatics/btq384.
43. Xu Y, Guo M, Shi W, Liu X, Wang C. A novel insight into gene ontology semantic similarity. Genomics. 2013;101(6):368–75. https://doi.org/10.1016/j.ygeno.2013.04.010.
44. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010;26(13):1644–50. https://doi.org/10.1093/bioinformatics/btq241.
45. Lu C, Yang M, Luo F, et al. Prediction of lncRNA-disease associations based on inductive matrix completion. Bioinformatics. 2018;34(19):3357–64. https://doi.org/10.1093/bioinformatics/bty327.
46. Deepthi K, Jereesh AS. An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network. Gene. 2020;762:145040. https://doi.org/10.1016/j.gene.2020.145040.


## Publisher's Note