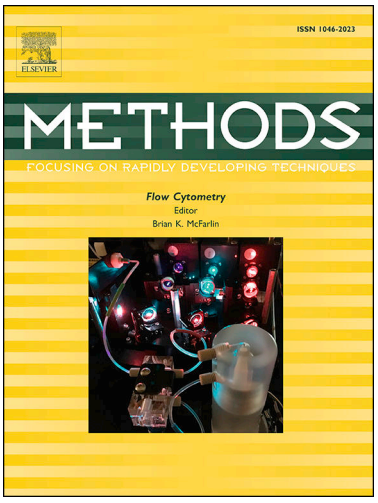
Journal Pre-proofs



SDLDA: lncRNA–disease association prediction based on singular value decomposition and deep learning

Min Zeng, Chengqian Lu, Fuhao Zhang, Yiming Li, Fang-Xiang Wu, Yaohang Li, Min Li

PII: S1046-2023(20)30001-3
DOI: https://doi.org/10.1016/j.ymeth.2020.05.002
Reference: YMETH 4894

To appear in: *Methods*

Received Date: 10 February 2020
Revised Date: 24 April 2020
Accepted Date: 2 May 2020

Please cite this article as: M. Zeng, C. Lu, F. Zhang, Y. Li, F-X. Wu, Y. Li, M. Li, SDLDA: lncRNA–disease association prediction based on singular value decomposition and deep learning, *Methods* (2020), doi: https://doi.org/10.1016/j.ymeth.2020.05.002

SDLDA: lncRNA–disease association prediction based on singular value decomposition and deep learning

Min Zeng[1], Chengqian Lu[1], Fuhao Zhang[1], Yiming Li[1],
Fang-Xiang Wu[2], Yaohang Li[3], Min Li[1,*]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China

[2]Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK S7N5A9, Canada

[3]Department of Computer Science, Old Dominion University, VA 23507, Norfolk, USA

* To whom correspondence should be addressed.
E-mail address: limin@mail.csu.edu.cn (Min Li).

Abstract

In recent years, accumulating studies have shown that long non-coding RNAs (lncRNAs) not only play an important role in the regulation of various biological processes but also are the foundation for understanding mechanisms of human diseases. Due to the high cost of traditional biological experiments, the number of experimentally verified lncRNA-disease associations is very limited. Thus, many computational approaches have been proposed to discover the underlying associations between lncRNAs and diseases. However, the associations between lncRNAs and diseases are too complicated to model by using only traditional matrix factorization-based methods. In this study, we propose a hybrid computational framework (SDLDA) for the lncRNA–disease association prediction. In our computational framework, we use singular value decomposition and deep learning to extract linear and non-linear features of lncRNAs and diseases, respectively. Then we train SDLDA by combing the linear and non-linear features. Compared to previous computational methods, the combination of linear and non-linear features reinforces each other, which is better than using only either matrix factorization or deep learning. The computational results show that SDLDA has a better performance over existing methods in the leave-one-out cross-validation. Furthermore, the case studies show that 28 out of 30 cancer-related lncRNAs (10 for gastric cancer, 10 for colon cancer and 8 for renal cancer) are verified by mining recent biomedical literature. Code and data can be accessed at https://github.com/CSUBioGroup/SDLDA.

Keywords: deep learning, singular value decomposition, lncRNA-disease association prediction, linear feature, non-linear feature.

1. Introduction

Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs (ncRNAs) with the length is larger than 200 nucleotides [1-3]. Because lncRNAs are typically expressed at a lower level, they were considered to be noise when first discovered in the early 1990s [4]. However, with the rapid development of high-throughput sequencing technology, researchers have found that lncRNAs play an important role in the various biological processes, including cell growth and development, regulation of gene expression, alternative splicing and nuclear organization [5]. In principle, lncRNAs can regulate the transcription by sequestering a protein target binding to DNAs; regulate transcriptional activities or pathways under a specific stimulation; act as platforms to host the formation of molecular complexes; and act as "miRNA sponge" to regulate the level of miRNAs and then affect the expression of miRNA's targets [5]. Accumulating studies have shown that lncRNAs can regulate gene expression in many ways, and the variation in gene expression is important in complex diseases. Thus lncRNAs are associated with various human diseases. For example, lncRNA PCA3 is treated as a potential biomarker of prostate cancer [6]. lncRNA BC200 expresses significantly higher in Alzheimer's disease tissue compared to normal tissues [7]. Identifying potential lncRNA–disease associations can enhance the study of human complex diseases including disease diagnosis, treatment, and prevention, which can help understand the disease mechanism at the lncRNA level. However, traditional biological experiments are expensive and time-consuming, which leads to the very limited number of experimentally verified lncRNA-disease associations. Thus it is urgent to develop accurate and effective computational methods to reveal potential lncRNA-disease associations.

In recent years, many computational methods have been proposed to predict potential lncRNA-disease associations and achieved the decent performance. These computational methods can be roughly divided into three categories: biological network-based, known lncRNA–disease associations-based, and machine learning-based methods. The first category integrates different types of networks, including lncRNA–disease association networks, disease similarity networks, lncRNA similarity networks, and other biological networks. Then some network algorithms (random walk and propagation algorithm) are adopted to predict potential lncRNA–disease associations in constructed heterogeneous networks [8-12]. The second category is based on the assumption that similar lncRNAs tend to have similar interactions or non-interactions with similar diseases. Using this assumption, the researchers could infer unknown lncRNA–disease associations based on known lncRNA–disease associations by constructing connections between them [13, 14]. The third category employs machine learning algorithms to model the associations between lncRNAs and diseases based on training samples (experimentally verified lncRNA–disease associations) and unlabeled samples (unknown lncRNA–disease pairs). Multiple machine learning algorithms have been applied to the prediction of lncRNA–disease associations, such as Naive Bayesian classifier [15], bagging support vector machine (SVM) [16] and matrix factorization-based methods [17-19]. Matrix factorization-based methods are the most popular methods among these machine learning-based methods. However, as mentioned above, lncRNAs can regulate gene expression in many ways. Thus the associations between lncRNAs and diseases are very complicated. Traditional matrix factorization-based methods predict potential lncRNA-disease associations by projecting a constructed matrix into a latent space to obtain their linear features. However, the associations between lncRNAs and diseases are too complicated;

only linear features extracted by matrix factorization techniques cannot model such complicated associations.

To overcome the limitation, we design a hybrid computational framework called SDLDA to extract non-linear features of lncRNAs and diseases which improve the representation ability of linear features extracted by matrix factorization techniques. Recently, deep learning techniques have led to successes in many applications of bioinformatics, from medical text classification and protein function prediction to protein-protein interaction prediction [20-26]. Inspired by their powerful ability of feature representation, we use deep learning techniques to capture complex and useful non-linear features of lncRNAs and diseases. We believe that a combination of linear and non-linear linear features can reinforce each other to get high-quality features, and thus can improve the prediction performance. To make use of the advantages of traditional matrix factorization-based methods and deep learning techniques, we design a novel computational framework that combines linear and non-linear features to predict lncRNA-disease associations. Specifically, we employ two fully connected layers to learn non-linear features of lncRNAs and diseases. The singular value decomposition technique is an effective approach that projects a matrix into a low-dimensional space while preserving the linear features of lncRNAs and diseases. Then, we concatenate the linear and non-linear features of lncRNAs and diseases to form a vector, respectively. Finally, the two vectors which contain linear and non-linear features are fused to a new vector, and the new vector is fed into a fully connected layer to perform the last prediction task. Different from previous studies, SDLDA has two advantages: 1) deep learning techniques are powerful learning techniques with multiple levels of representation, which can learn non-linear and more complicated and useful features of lncRNAs and diseases than traditional matrix factorization-based methods; 2) the combination of the learned linear and non-linear linear features can reinforce each other to better predict potential lncRNA–disease associations.

To evaluate the effectiveness of our model, we compared SDLDA with four lncRNA-disease association prediction methods (SIMCLDA [19], MFLDA [18], TPGLDA [27] and LDAP [16]). According to the results of leave-one-out cross-validation, SDLDA obtains the highest AUC and AUPR, showing our model outperforms other prediction methods. Moreover, we conducted case studies for three cancers including gastric cancer, colon cancer, and renal cancer to further evaluate the real effects of our model. Case studies show the capability of our model for predicting potential lncRNA–disease associations. In summary, introducing deep learning techniques to a traditional matrix factorization model is useful for predicting potential lncRNAs–disease associations. All results corroborate the effectiveness of SDLDA.

2. Material and methods

2.1 Data sources

In this study, known lncRNA-disease associations were retrieved from three databases: LncRNADisease [28], GeneRIF [29], Lnc2Cancer [30]. By checking names of lncRNAs (according to Lncipedia, lncrnadb, HGNC, and NCBI) and diseases (according to Mesh, UMLS, and NCBI), all repeating records and entries are removed. The resultant dataset consists of 1583 associations among 577 lncRNAs and 272 diseases, whose interaction density is about 1.008%. The statistics of the resultant dataset are shown in Table 1.

Table 1. Statistics of the constructed dataset.

Dataset	No. of lncRNAs	No. of diseases	No. of known interactions	Interaction density[#] (%)
	577	272	1583	1.008

[#]The interaction density is defined as the ratio of the number of known lncRNA-disease interactions to the number of all possible lncRNA-disease interactions.

2.2 Problem formulation

As our previous studies [19], the lncRNA–disease association prediction problem is treated as a recommendation problem. We construct an interaction matrix $R \in R^{m \times n}$ to represent known lncRNA-disease associations, where m and n represent the numbers of lncRNAs and diseases, respectively. If an lncRNA i has a known association with a disease j, $R(i, j)$ is equal to 1, and otherwise 0. It's worth noting that 0 in the interaction matrix R does not mean that there is no association between corresponding lncRNA and disease; it means that the association between them is unobserved yet. The task of lncRNA-disease association prediction is to use known (1 in the interaction matrix R) and unobserved (0 in the interaction matrix R) lncRNA–disease associations to recalculate the unobserved associations.

2.3 Extracting linear features of lncRNAs and diseases with singular value decomposition method

Matrix factorization techniques have led to successes in the recommendation system. The singular value decomposition (SVD) method is the most popular matrix factorization method in the recommendation system [31]. Thus, the SVD technique is applied to our model to capture linear features of lncRNAs and diseases. The details of SVD are as follows. Let $R \in R^{m \times n}$ be the lncRNA-disease association matrix, the SVD of matrix R is a factorization of three matrices (U, $\sum$, *and* V^{T}) as follows.

$$R = U \sum V^T \tag{1}$$

where $U \in R^{m \times m}$ is a real matrix, $\sum \in R^{m \times n}$ is a diagonal matrix with non-negative square roots of the eigenvalues of the product $R^T R$ on the diagonal, and $V \in R^{n \times n}$ is a real matrix. The diagonal elements σ_i are called singular values of matrix R.

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0 \tag{2}$$

We can get an approximation representation of matrix R by keeping the k largest singular values.

$$R \approx U_k \sum_k (V_k)^T \tag{3}$$

where $U_k \in R^{m \times k}$ is a real matrix, $\sum_k \in R^{k \times k}$ is a diagonal matrix, and $V_k^T \in R^{k \times n}$ is a real matrix. Figure 1 gives the illustration of the SVD approximation representation of interaction matrix R. In our study, the U_i and V_j^T are considered as linear features of lncRNA i and disease j, respectively.

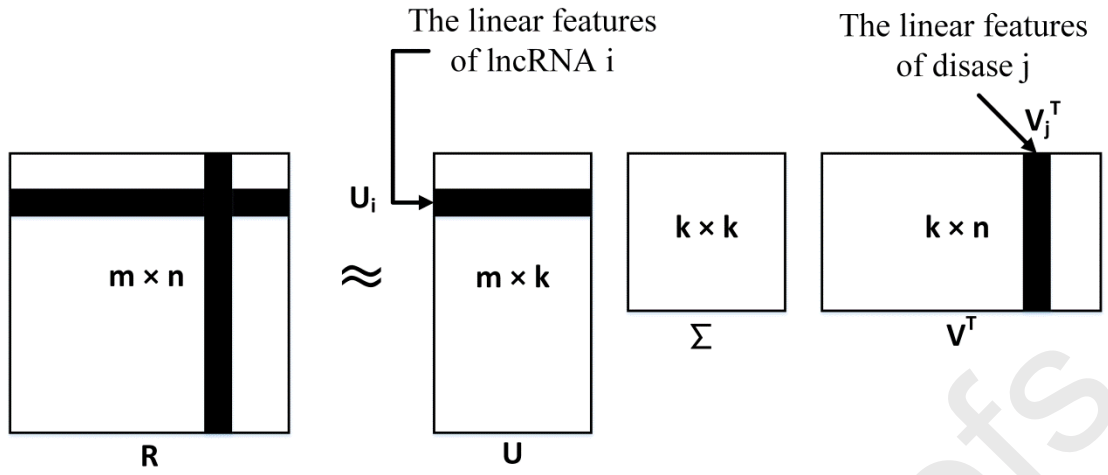


Figure 1. Illustration of the singular value decomposition of interaction matrix R into matrices U_k, $\sum_k$ and V_k^T. LncRNA i is represented by the linear feature vector U_i, i.e. row i in matrix U, disease j is represented by the linear feature vector V_j^T, i.e. column j in matrix V^T.

2.4 Extracting non-linear features of lncRNAs and diseases with deep learning techniques

In recent years, deep learning techniques have led to successes in many fields, including computer vision, natural language processing, and bioinformatics [32, 33]. It is straightforward to explore how to use deep learning techniques in the recommendation system. Inspired by their studies [34, 35], we design a novel computational framework to combine singular value decomposition and deep learning techniques for predicting lncRNA-disease associations.

In our computational framework, we use the raw interaction matrix R as the input. Each row of the interaction matrix R is treated as the raw features of lncRNAs, each column of the interaction matrix R is treated as the raw features of diseases. The output is the intersection of a row and a column, i.e. the label of a pair of lncRNA and disease. In the interaction matrix R, the intersection of a row and a column is considered the label of a pair of corresponding lncRNA and disease. We should not directly use the label in the raw representation of lncRNAs and diseases. In our task, 1 in the interaction matrix R means a known interaction while and 0 in the interaction matrix R means an unobserved association. Thus, we use the unobserved value to mask the value in raw representation. For the non-linear features of lncRNAs and diseases, two fully connected layers with a non-linear activation function are applied. Formally, x and y denote the lncRNA and disease raw feature vectors, respectively. The two raw feature vectors are fed into the first fully connected layer, the outputs O_{x1} and O_{y1} are:

$$O_{x1} = \sigma(W_{x1}x + b_{x1}) \tag{4}$$

$$O_{y1} = \sigma(W_{y1}y + b_{y1}) \tag{5}$$

where σ is the non-linear activation function, W_{x1} and W_{y1} are weight matrices of the first fully connected layer, b_{x1} and b_{y1} are bias terms of the first fully connected layer.

Then the outputs O_{x1} and O_{y1} are fed into the second fully connected layer, the outputs O_{x2} and O_{y2} are considered as the final non-linear features:

$$O_{x2} = \sigma(W_{x2}O_{x1} + b_{x2}) \quad (6)$$

$$O_{y2} = \sigma(W_{y2}O_{y1} + b_{y2}) \quad (7)$$

where σ is the non-linear activation function, W_{x2} and W_{y2} are weight matrices of the second fully connected layer, b_{x2} and b_{y2} are bias terms of the second fully connected layer.

In each fully connected layer, the Rectified Linear Unit (ReLU) activation function is applied to extract non-linear features. The ReLU function is defined as follows:

$$ReLU(x) = max(0, x) \quad (8)$$

2.5 Combination of linear and non-linear features

So far we have obtained linear and non-linear features of lncRNAs and diseases by singular value decomposition and deep learning techniques. The next question is that how can we combine them into a computational framework in the prediction of lncRNAs and diseases. In our model, we first concatenate them to a vector.

$$Lnc_{new\_feature} = \begin{bmatrix} U_i \\ O_{x2} \end{bmatrix} \quad (9)$$

$$Dis_{new\_feature} = \begin{bmatrix} V_j^T \\ O_{y2} \end{bmatrix} \quad (10)$$

where [ ] is concatenation operation.

Then, the Hadamard product operation is applied to integrate the two vectors into a new vector.

$$vec_{new} = Hadamard\,product\,(Lnc_{new\_feature}, Dis_{new\_feature}) \quad (11)$$

Last, the integrated new vector is fed into a fully connected layer with a sigmoid function to perform the final prediction task.

$$sigmoid(x) = \frac{1}{1 + exp\ (-x)} \quad (12)$$

$$R_{pred} = sigmoid(vec_{new}) \quad (13)$$

Figure 2 illustrates our proposed computational framework.

The binary cross-entropy function is used in our model as the loss function, it is defined as follows:

$$Loss = \sum [Rlog(R_{pred}) + (1 - R)log\,(1 - R_{pred})] + \lambda(|\theta\,|\text{^}2) \quad (14)$$

where θ is the weight vector, λ is a weight to balance between the empirical risk and regularized term.

2.6 Evaluation metrics

Similar to previous studies, we use the leave-one-out cross-validation (LOOCV) based on the known lncRNA-disease associations to evaluate the performance of SDLDA. In LOOCV, each time we use one known lncRNA–disease association as the test sample and the remaining lncRNA–disease associations as the training samples. The new values in the interaction matrix can be obtained after our model is implemented. Then we can use the new interaction matrix to calculate the true positive rate (TPR) and false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, TN is the number of true negatives.

The receiver-operating characteristics (ROC) curve is a graphical plot that shows the diagnostic ability of a binary classifier. It plots the TPR against FPR at different threshold. The area under the ROC curve (AUC) is wildly used to evaluate the performance of a classifier. We use ROC and AUC to evaluate our model and other existing computational methods.

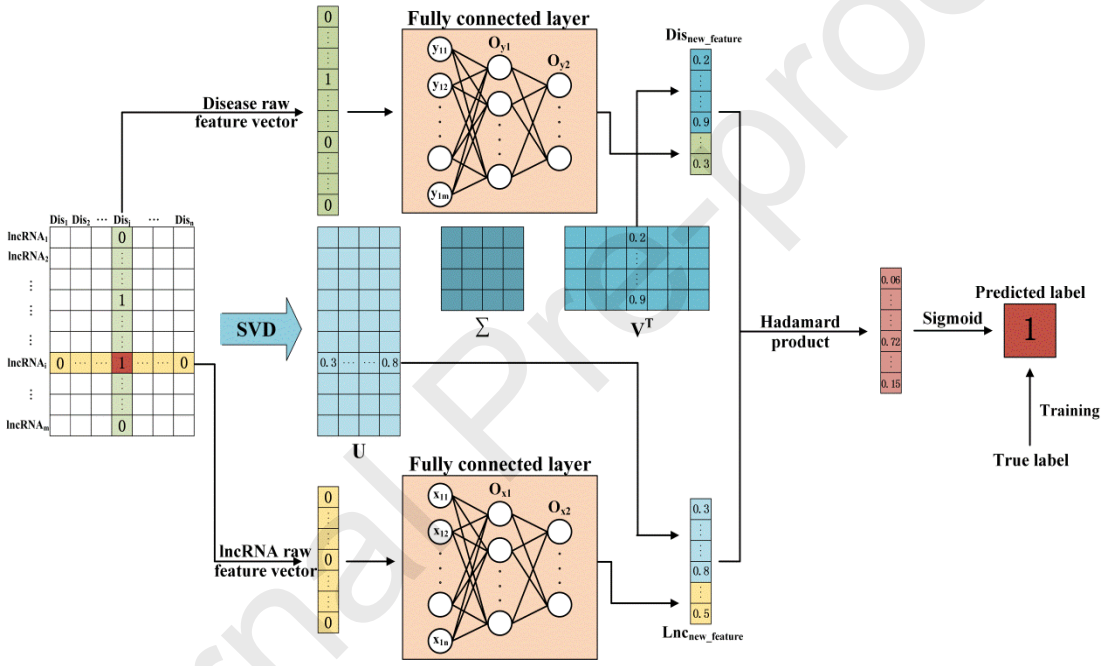


Figure 2. Illustration of our proposed computational framework. The input of the computational framework is the constructed lncRNA-disease interaction matrix *R*. The SVD technique is applied to decompose the interaction matrix *R* into three matrices *U*, ∑, and V^T. The row of *U* and the column of V^T represent the linear features of lncRNAs and diseases, respectively. In the deep learning part, the row and column in the interaction matrix *R* are considered as the raw feature vectors of corresponding lncRNA and disease, respectively. The two vectors are fed into two fully connected layers with the non-linear activation function to extract non-linear features. Then the linear features and non-linear features are concatenated to a vector, respectively. The Hadamard product operation is applied to fuse the two vectors. Last, a sigmoid activation function is applied to perform the final prediction task.

3. Results

3.1 Implementation details

The SVD is implemented with the Scipy library, and a 64-dimensional vector is used to represent linear features of lncRNAs and diseases. The deep learning computational framework is implemented with Tensorflow [36], and two fully connected layers are used with ReLU activation function to extract non-linear features of lncRNAs and diseases. The numbers of neurons in the first and second fully connected layers are 48 and 32, respectively. We used the dropout rate of 0.05 and regularization parameter λ of 0.001 to avoid overfitting. In the training process, the batch size is set to 32; the adaptive moment estimation (Adam) optimizer is used as the optimizer; the initial learning rate is 0.001.

3.2 Comparison with existing computational methods

We compared SDLDA with four existing computational methods (SIMCLDA [19], MFLDA [18], TPGLDA [27] and LDAP [16]). These four methods are machine learning-based or matrix factorization-based methods. SIMCLDA uses the inductive matrix completion to estimate the potential lncRNA–disease associations by integrating prior knowledge of lncRNAs and diseases. MFLDA uses a matrix tri-factorization technique to decompose the constructed matrices of heterogeneous data into low-rank matrices which can exploit the intrinsic and shared structure of heterogeneous data. TPGLDA integrates gene-disease associations with lncRNA-disease associations to predict potential lncRNA-disease associations based on an allocation algorithm. LDAP fuses lncRNA similarity and disease similarity data to predict potential lncRNA-disease associations with a bagging SVM.

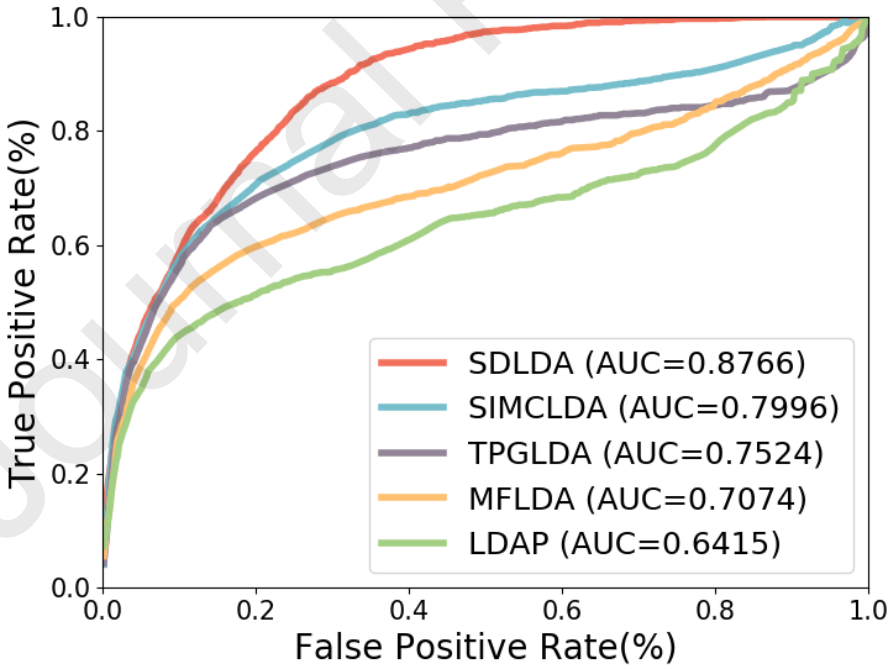


Figure 3. The ROC curves of SDLDA and other computational methods

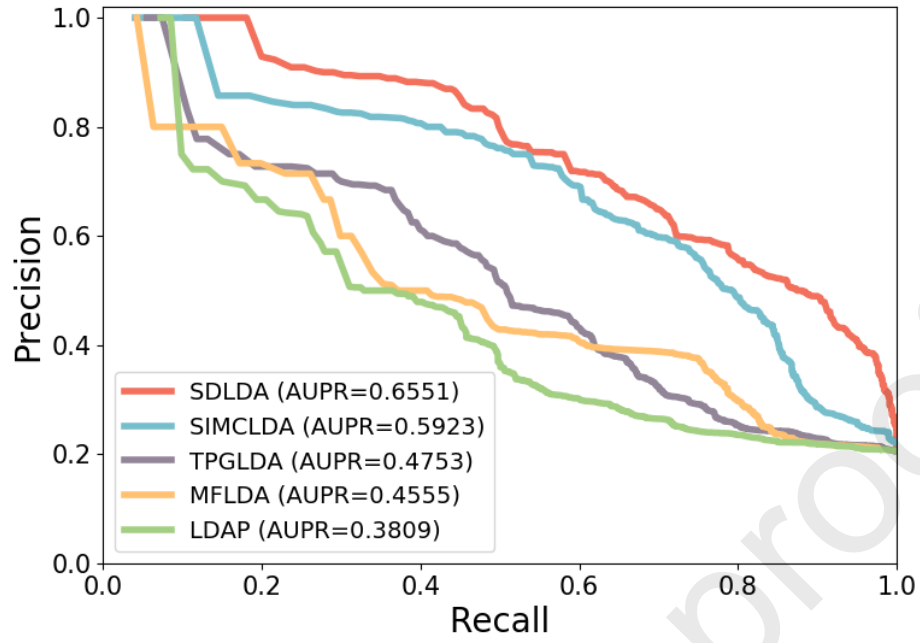


Figure 4. The PR curves of SDLDA and other computational methods

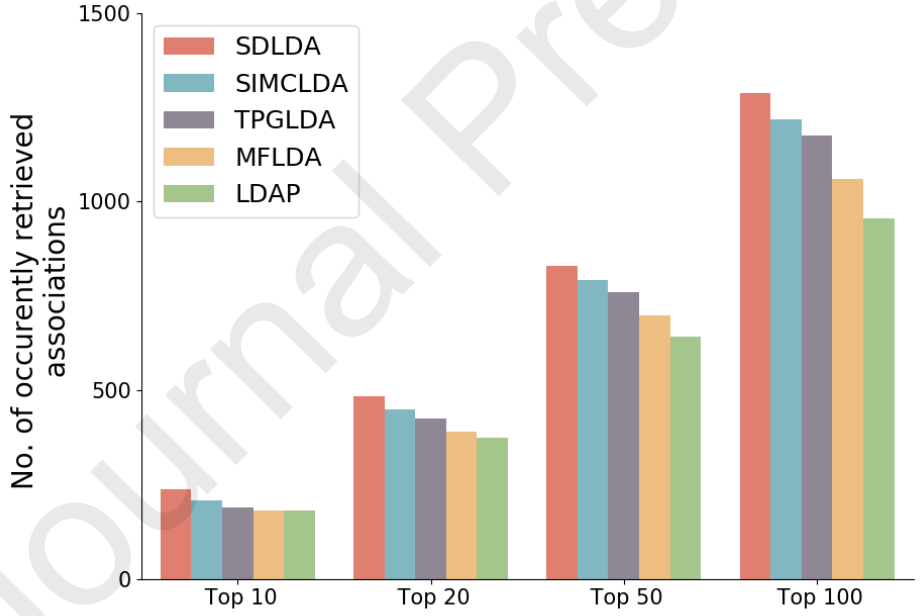


Figure 5. The number of correctly retrieved known lncRNA–disease associations for specified rank thresholds

Figure 3 plots the ROC curves of SDLDA and other computational methods. The AUC of SDLDA is 0.8766, which is significantly higher than those of SIMCLDA (0.7996), TPGLDA (0.7524), MFLDA (0.7074) and LDAP (0.6415). The PR curves of SDLDA and other computational methods are shown in Figure 4. The AUPR of SDLDA is 0.6551, which is higher than those of SIMCLDA (0.5923), TPGLDA (0.4753), MFLDA (0.4555) and LDAP (0.3809).

Table 2 shows the accuracy of SDLDA and other computational methods. From Table 2, we find that the accuracy of SDLDA is higher than other computational methods. Therefore, we can see the improved performance of SDLDA in predicting lncRNA-disease associations. Furthermore, we calculated the numbers of correctly retrieved lncRNA–disease associations for all computational methods. Specifically, for a specific rank threshold *k*, if the rank of predicted association is higher than *k*, the association is regarded as a correctly retrieved association. Figure 5 shows the number of correctly retrieved known lncRNA–disease associations in the top 10, 20, 50, and 100. From Figure 5, we can see that SDLDA can retrieve more correct associations than other computational methods. In summary, all results corroborate the effectiveness of SDLDA.

Table 2. Accuracy of SDLDA and other computational methods

	SDLDA	SIMCLDA	TPGLDA	MFLDA	LDAP
Accuracy	0.913	0.854	0.815	0.798	0.776

3.3 The effects of hyper-parameters

In our model, some hyper-parameters have different effects on experimental performance. Here, we focus on two specific hyper-parameters, i.e. the number of neurons in the last fully connected layer and the regularization parameter λ in the loss function. The number of neurons in the last fully connected layer can determine the dimension of non-linear features; the regularization parameter λ is used to balance the empirical risk and regularized term. We changed the number of neurons in the last layer from 8 to 48 (8, 16, 32 and 48) to find the best parameter. The results are shown in Table 3. We trained our model with the different parameters of 0.0001, 0.0003, 0.001, 0.003, 0.01 and 0.03 for λ to find the best parameter with LOOCV. The results are shown in Table 4. In summary, we set neurons = 32 and λ = 0.001 in SDLDA.

Table 3. Performance for models with different values of neurons in the last fully connected layer

# of neurons	8	16	32	48
AUC	0.8613	0.8710	0.8766	0.8735
AUPR	0.6458	0.6505	0.6551	0.6514
Accuracy	0.9056	0.9111	0.9130	0.9118

Table 4. Performance for models with different values of regularization parameter λ

λ	0.0001	0.0003	0.001	0.003	0.01	0.03
AUC	0.8747	0.8745	0.8766	0.8755	0.8698	0.8539
AUPR	0.6532	0.6530	0.6551	0.6540	0.6492	0.6401
Accuracy	0.9123	0.9121	0.9130	0.9125	0.9101	0.9003

3.4 Case studies

To further evaluate the real effects of SDLDA, we applied SDLDA to predict three human cancers: gastric cancer, colon cancer, and renal cancer. For a specific cancer, we utilize the well-trained model to predict new values for those lncRNAs that do not have interactions with specific cancer. Then the top 10 candidate lncRNAs of this cancer are selected as our predicted disease-related lncRNAs. Last we check them by manually mining recent biomedical literature.

Table 5. SDLDA predicted lncRNAs associated with gastric cancer (top 10) with the corresponding references.

Rank	LncRNA	Reference
1	H19	Yan et al. (2017)
2	MALAT1	Li et al. (2017)
3	CDKN2B-AS1	Riquelme et al. (2016)
4	NEAT1	Ma et al. (2016)
5	PVT1	Zhao et al. (2018)
6	TUG1	Baratieh et al. (2017)
7	MEG3	Peng et al. (2015)
8	GAS5	Guo et al. (2015)
9	KCNQ1OT1	Sunamura et al. (2016)
10	BCYRN1	Ren et al. (2018)

Gastric cancer is the fourth most common cancer and the second leading cause of cancer death worldwide [37]. SDLDA is applied to predict potential gastric cancer-related lncRNAs. As shown in Table 5, the top 10 predicted lncRNAs have been all validated by manually mining recent biomedical literature. LncRNA H19 promotes gastric cancer via FADD/Caspase 8/Caspase 3 signaling pathway [38]. LncRNA MALAT1 correlates with human gastric cancer vasculogenic mimicry density [39]. Riquelme et al. pointed out that LncRNA CDKN2B-AS1 has a higher expression in human gastric cancer tissues [40]. The expression of lncRNA NEAT1 is enhanced in gastric cancer [41]. LncRNA PVT1 promotes angiogenesis via activating the STAT3/VEGFA axis in gastric cancer [42]. LncRNA TUG1 is a potential biomarker for gastric cancer [43]. LncRNA MEG3 functions as a competing endogenous RNA to regulate gastric cancer progression [44]. LncRNA GAS5 plays an important role in the molecular etiology of gastric cancer [45]. Sunamura

et al. found that the accumulation of nuclear β-catenin induced dysregulation of lncRNA KCNQ1OT1 transcription in gastric cancer cells [46]. The upregulation of lncRNA BCYRN1 promotes tumor progression and enhances EpCAM expression in gastric carcinoma [47].

Colon cancer, also known as colorectal cancer, is the third most commonly diagnosed cancer in males and the second in females worldwide [48]. SDLDA is applied to predict potential colon cancer-related lncRNAs. As shown in Table 6, the top 10 predicted lncRNAs have been all validated by manually mining recent biomedical literature. LncRNA PVT1 functions as an oncogene in colon cancer through miR-30d-5p/RUNX2 axis [49]. Chen et al. have pointed out that lncRNA CDKN2B-AS1 has been experimentally confirmed [50]. LincRNA-p21 enhances the sensitivity of radiotherapy in colorectal cancer by targeting the Wnt/β-catenin signaling pathway [51]. LncRNA NEAT1 can directly sponge miR-662 to promote invasion and migration of colon cancer cells [52]. LncRNA GAS5 is commonly downregulated in colorectal cancer tissues [53]. LncRNA XIST is a prognostic factor in colorectal cancer [54]. LncRNA TUSC7 can act as a tumor suppressor in colorectal cancer [55]. The overexpression of LncRNA HOTTIP serves as an unfavorable prognosis predictor for colorectal cancer patients [56]. LncRNA CRNDE promotes colorectal cancer cell proliferation [57]. LncRNA SPRY4-IT1 promotes the malignant development of colorectal cancer [58].

Table 6. SDLDA predicted lncRNAs associated with colon cancer (top 10) with the corresponding references.

Rank	LncRNA	Reference
1	PVT1	Yu et al. (2018)
2	CDKN2B-AS1	Chen et al. (2016)
3	lincRNA-p21	Chen et al. (2019)
4	NEAT1	Song et al. (2017)
5	GAS5	Li et al. (2018)
6	XIST	Xiao et al. (2017)
7	TUSC7	Ren et al. (2017)
8	HOTTIP	Ren et al. (2015)
9	CRNDE	Ding et al. (2017)
10	SPRY4-IT1	Gao et al. (2016)

Renal cancer is one of the ten most common cancers. SDLDA is applied to predict potential renal cancer-related lncRNAs. As shown in Table 7, 8 out of the top 10 predicted lncRNAs have been all validated by manually mining recent biomedical literature. Down-regulated lncRNA H19 inhibits carcinogenesis of renal cell carcinoma [59]. LncRNA CDKN2B-AS1 is associated with the progression of renal cell carcinoma [60]. The upregulation of lncRNA MIAT regulates LOXL2 expression in clear cell renal cell carcinoma [61]. LncRNA neat1 enhances epithelial-to-mesenchymal transition and chemoresistance in renal cell carcinoma [62]. The relative level of lncRNA TUG1 is significantly higher in renal cell carcinoma tissues [63]. LncRNA GAS5 expression level is significantly lower in renal cell carcinoma [64]. LncRNA PVT1 functions as

competing endogenous RNA to regulate clear cell renal cell carcinoma progression [65]. The expression of lncRNA RN7SK paralogs in tumors of renal cell carcinoma [66].

Table 7. SDLDA predicted lncRNAs associated with renal cancer (top 10) with the corresponding references.

Rank	LncRNA	Reference
1	H19	Wang et al. (2015)
2	CDKN2B-AS1	He et al. (2016)
3	MIAT	Qu et al. (2018)
4	NEAT1	Liu et al. (2017)
5	TUG1	Wang et al. (2017)
6	GAS5	Seles et al. (2016)
7	PVT1	Yang et al. (2017)
8	BCYRN1	Unknown
9	RN7SK	Zhang et al. (2011)
10	DISC2	Unknown

In summary, 28 out of 30 cancer-related lncRNAs (10 for gastric cancer, 10 for colon cancer and 8 for renal cancer) are checked in the recent biomedical literature. Two lncRNAs (BCYRN1 and DISC2) are not found in the recent literature, the associations of the two lncRNAs are unknown which is deserved for biologists to validate their functions via wet-lab experiments.

3.5 Ablation study

In our method, we combine linear and non-linear features to predict potential lnRNA-disease associations. In order to investigate whether the combination of linear and non-linear features is helpful to predict lncRNA-disease associations, we have conducted an ablation study by removing the individual part in our method. Specifically, we tested the performances of models by using only linear features, only non-linear features or the combination of linear and non-linear features. The linear features are extracted by SVD technique and the non-linear features are extracted by deep learning techniques. From the results presented in Table 8, AUCs of models with only linear features, only non-linear features, and the combination of linear and non-linear features are 0.8168, 0.8393, and 0.8766, respectively. AUPRs of models with only linear features, only non-linear features, and the combination of linear and non-linear features are 0.6551, 0.6324, and 0.6086, respectively. Accuracies of models with only linear features, only non-linear features, and the combination of linear and non-linear features are 0.9130, 0.8892, and 0.8704, respectively. In summary, we can get better performance with combining linear and nonlinear features than the only linear features or only non-linear features.

Table 8. An ablation study on different models

Model	AUC	AUPR	Accuracy
Only using linear features	0.8168	0.6086	0.8704
Only using non-linear features	0.8393	0.6324	0.8892
Linear features & non-linear feature	0.8766	0.6551	0.9130

4. Conclusions

Identifying potential lncRNA–disease associations can enhance the study of human complex diseases at the lncRNA level. However, traditional biological experiments are expensive, time-consuming and laborious, which lead the number of experimentally verified lncRNA-disease associations is very limited. Consequently, a lot of computational methods have been proposed in recent years. In this study, we develop a novel computational framework (SDLDA) by combining singular value decomposition and deep learning techniques for predicting lncRNA–disease associations. The singular value decomposition technique is applied to extract linear features of lncRNAs and diseases. We use a neural network with two fully connected layers to learn non-linear features of lncRNAs and diseases. The linear and non-linear features are concatenated into a vector for the final prediction. In order to illustrate the effectiveness of SDLDA, four computational methods (SIMCLDA, MFLDA, TPGLDA and LDAP) are compared. The LOOCV results demonstrate the improved performance of SDLDA in predicting lncRNA-disease associations. To further evaluate the performance of SDLDA, three case studies of gastric cancer, colon cancer and renal cancer are performed. 28 out of 30 cancer-related lncRNAs are verified by mining recent biomedical literature. Results have shown that our method could be a useful tool for predicting lncRNA-disease associations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant (No. U19A2064), the 111 Project (No. B18059), Hunan Provincial Science and Technology Program (2018WK4001), the Fundamental Research Funds for the Central Universities of Central South University (No. 2019zzts281). Part of this paper is published in the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2019 [67].

Author contributions

MZ and ML conceived and designed the experiments. MZ, CL, FZ and YL (Yiming Li) performed the experiments. MZ, CL, FXW and ML drafted the manuscript. MZ and YL (Yiming Li) drafted the figures. MZ, FXW, YL (Yaohang Li) and ML revised the manuscript. All authors approved the final manuscript.

Declaration of interest

The authors declare that they have no competing interests.

References

[1] P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermüller, I.L. Hofacker, RNA maps reveal new RNA classes and a possible function for pervasive transcription, Science, 316 (2007) 1484-1488.

[2] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: insights into functions, Nature reviews genetics, 10 (2009) 155.

[3] M. Guttman, P. Russell, N.T. Ingolia, J.S. Weissman, E.S. Lander, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins, Cell, 154 (2013) 240-251.

[4] M. Esteller, Non-coding RNAs in human disease, Nature Reviews Genetics, 12 (2011) 861.

[5] J.K. DiStefano, The emerging role of long noncoding RNAs in human disease, Disease Gene Identification, (Springer, 2018), pp. 91-110.

[6] H. van Poppel, A. Haese, M. Graefen, A. de la Taille, J. Irani, T. de Reijke, M. Remzi, M. Marberger, The relationship between Prostate CAncer gene 3 (PCA3) and prostate cancer significance, BJU international, 109 (2012) 360-366.

[7] W. Lukiw, P. Handley, L. Wong, D.C. McLachlan, BC200 RNA in normal human neocortex, non-Alzheimer dementia (NAD), and senile dementia of the Alzheimer type (AD), Neurochemical research, 17 (1992) 591-597.

[8] J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu, M. Zhou, Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network, Molecular BioSystems, 10 (2014) 2074-2081.

[9] Q. Yao, L. Wu, J. Li, L. guang Yang, Y. Sun, Z. Li, S. He, F. Feng, H. Li, Y. Li, Global prioritizing disease candidate lncRNAs via a multi-level composite network, Scientific reports, 7 (2017) 39516.

[10] X. Chen, KATZLDA: KATZ measure for the lncRNA-disease association prediction, Scientific reports, 5 (2015) 16840.

[11] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou, J. Sun, Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network, Molecular bioSystems, 11 (2015) 760-769.

[12] J. Zhang, Z. Zhang, Z. Chen, L. Deng, Integrating multiple heterogeneous networks for novel lncRNA-disease association inference, IEEE/ACM transactions on computational biology and bioinformatics, (2017).

[13] M.-X. Liu, X. Chen, G. Chen, Q.-H. Cui, G.-Y. Yan, A computational framework to infer human disease-associated long noncoding RNAs, PloS one, 9 (2014) e84408.

[14] X. Chen, Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA, Scientific reports, 5 (2015) 13186.

[15] T. Zhao, J. Xu, L. Liu, J. Bai, C. Xu, Y. Xiao, X. Li, L. Zhang, Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features, Molecular BioSystems, 11 (2015) 126-136.

[16] W. Lan, M. Li, K. Zhao, J. Liu, F.-X. Wu, Y. Pan, J. Wang, LDAP: a web server for lncRNA-disease association prediction, Bioinformatics, 33 (2016) 458-460.

[17] X. Chen, G.-Y. Yan, Novel human lncRNA–disease association inference based on lncRNA expression profiles, Bioinformatics, 29 (2013) 2617-2624.

[18] G. Fu, J. Wang, C. Domeniconi, G. Yu, Matrix factorization-based data fusion for the prediction of lncRNA–disease associations, Bioinformatics, 34 (2017) 1529-1537.

[19] C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li, J. Wang, Prediction of lncRNA-disease associations based on inductive matrix completion, Bioinformatics, 1 (2018) 8.

[20] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9 coding via deep transfer learning, Neurocomputing, 324 (2019) 43-50.

[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems2012), pp. 1097-1105.

[22] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems2013), pp. 3111-3119.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, (2013).

[24] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, J. Wang, Automated ICD-9 Coding via A Deep Learning Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics, (2018) 1-1.

[25] M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, J. Wang, A deep learning framework for identifying essential proteins by integrating multiple types of biological information, IEEE/ACM transactions on computational biology and bioinformatics, (2019).

[26] F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, M. Li, DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions, Proteomics, (2019) 1900019.

[27] L. Ding, M. Wang, D. Sun, A. Li, TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph, Scientific reports, 8 (2018) 1065.

[28] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, Q. Cui, LncRNADisease: a database for long-non-coding RNA-associated diseases, Nucleic acids research, 41 (2012) D983-D986.

[29] Z. Lu, K. BRETONNEL COHEN, L. Hunter, GeneRIF quality assurance as summary revision, Biocomputing 2007, (World Scientific, 2007), pp. 269-280.

[30] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, L. Wang, Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers, Nucleic Acids Research, 44 (2016) D980-D985.

[31] D. Billsus, M.J. Pazzani, Learning Collaborative Information Filters, Icml1998), pp. 46-54.

[32] M. Zeng, M. Li, F.-X. Wu, Y. Li, Y. Pan, DeepEP: a deep learning framework for identifying essential proteins, BMC Bioinformatics, 20 (2019) 506.

[33] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, M. Li, Protein–protein interaction site prediction through combining local and global features with deep neural networks, Bioinformatics, (2019).

[34] H.-J. Xue, X. Dai, J. Zhang, S. Huang, J. Chen, Deep Matrix Factorization Models for Recommender Systems, IJCAI2017), pp. 3203-3209.

[35] M. van Baalen, Deep Matrix Factorization for Recommendation, (2016).

[36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a system for large-scale machine learning, OSDI2016), pp. 265-283.

[37] E. Van Cutsem, X. Sagaert, B. Topal, K. Haustermans, H. Prenen, Gastric cancer, The Lancet, 388 (2016) 2654-2664.

[38] J. Yan, Y. Zhang, Q. She, X. Li, L. Peng, X. Wang, S. Liu, X. Shen, W. Zhang, Y. Dong, Long noncoding RNA H19/miR-675 axis promotes gastric cancer via FADD/Caspase 8/Caspase 3 signaling

pathway, Cellular Physiology and Biochemistry, 42 (2017) 2364-2376.
[39] Y. Li, Z. Wu, J. Yuan, L. Sun, L. Lin, N. Huang, J. Bin, Y. Liao, W. Liao, Long non-coding RNA MALAT1 promotes gastric cancer tumorigenicity and metastasis by regulating vasculogenic mimicry and angiogenesis, Cancer letters, 395 (2017) 31-44.
[40] I. Riquelme, C. Ili, J.C. Roa, P. Brebi, Long non-coding RNAs in gastric cancer: mechanisms and potential applications, Oncotarget, 5 (2016).
[41] Y. Ma, L. Liu, F. Yan, W. Wei, J. Deng, J. Sun, Enhanced expression of long non-coding RNA NEAT1 is associated with the progression of gastric adenocarcinomas, World journal of surgical oncology, 14 (2016) 41.
[42] J. Zhao, P. Du, P. Cui, Y. Qin, J. Wu, Z. Zhou, W. Zhang, L. Qin, G. Huang, LncRNA PVT1 promotes angiogenesis via activating the STAT3/VEGFA axis in gastric cancer, Oncogene, 37 (2018) 4094.
[43] Z. Baratieh, Z. Khalaj, M.A. Honardoost, M. Emadi-Baygi, H. Khanahmad, M. Salehi, P. Nikpour, Aberrant expression of PlncRNA-1 and TUG1: potential biomarkers for gastric cancer diagnosis and clinically monitoring cancer progression, Biomarkers in medicine, 11 (2017) 1077-1090.
[44] W. Peng, S. Si, Q. Zhang, C. Li, F. Zhao, F. Wang, J. Yu, R. Ma, Long non-coding RNA MEG3 functions as a competing endogenous RNA to regulate gastric cancer progression, Journal of experimental & clinical cancer research, 34 (2015) 79.
[45] X. Guo, K. Deng, H. Wang, J. Xia, T. Shan, Z. Liang, L. Yao, S. Jin, GAS5 inhibits gastric cancer cell proliferation partly by modulating CDK6, Oncology research and treatment, 38 (2015) 362-366.
[46] N. Sunamura, T. Ohira, M. Kataoka, D. Inaoka, H. Tanabe, Y. Nakayama, M. Oshimura, H. Kugoh, Regulation of functional KCNQ1OT1 lncRNA by β-catenin, Scientific reports, 6 (2016) 20690.
[47] H. Ren, X. Yang, Y. Yang, X. Zhang, R. Zhao, R. Wei, X. Zhang, Y. Zhang, Upregulation of LncRNA BCYRN1 promotes tumor progression and enhances EpCAM expression in gastric carcinoma, Oncotarget, 9 (2018) 4851.
[48] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R.E.M. Pirozzi, F. Corcione, Worldwide burden of colorectal cancer: a review, Updates in surgery, 68 (2016) 7-11.
[49] X. Yu, J. Zhao, Y. He, Long non-coding RNA PVT1 functions as an oncogene in human colon cancer through miR-30d-5p/RUNX2 axis, J BUON, 23 (2018) 48-54.
[50] X. Chen, C.C. Yan, X. Zhang, Z.-H. You, Long non-coding RNAs and complex diseases: from experimental results to computational models, Briefings in bioinformatics, 18 (2016) 558-576.
[51] L. Chen, D. Yuan, Y. Yang, M. Ren, LincRNA-p21 enhances the sensitivity of radiotherapy for gastric cancer by targeting the β-catenin signaling pathway, Journal of cellular biochemistry, 120 (2019) 6178-6187.
[52] B. Song, J. Yan, C. Liu, H. Zhou, Long non-coding RNA NEAT1 promotes metastasis via enhancing ZEB2 by sponging miR-662 in colorectal cancer, Int J Clin Exp Pathol, 10 (2017) 4470-4478.
[53] Y. Li, Y. Li, S. Huang, K. He, M. Zhao, H. Lin, D. Li, J. Qian, C. Zhou, Y. Chen, Long non-coding RNA growth arrest specific transcript 5 acts as a tumour suppressor in colorectal cancer by inhibiting interleukin-10 and vascular endothelial growth factor expression, Oncotarget, 8 (2017) 13690.
[54] Y. Xiao, U.A. Yurievich, S.V. Yosypovych, Long noncoding RNA XIST is a prognostic factor in colorectal cancer and inhibits 5-fluorouracil-induced cell cytotoxicity through promoting thymidylate synthase expression, Oncotarget, 8 (2017) 83171.

[55] W. Ren, S. Chen, G. Liu, X. Wang, H. Ye, Y. Xi, TUSC7 acts as a tumor suppressor in colorectal cancer, American journal of translational research, 9 (2017) 4026.
[56] Y.-K. Ren, Y. Xiao, X.-B. Wan, Y.-Z. Zhao, J. Li, Y. Li, G.-S. Han, X.-B. Chen, Q.-Y. Zou, G.-C. Wang, Association of long non-coding RNA HOTTIP with progression and prognosis in colorectal cancer, International journal of clinical and experimental pathology, 8 (2015) 11458.
[57] J. Ding, J. Li, H. Wang, Y. Tian, M. Xie, X. He, H. Ji, Z. Ma, B. Hui, K. Wang, Long noncoding RNA CRNDE promotes colorectal cancer cell proliferation via epigenetically silencing DUSP5/CDKN1A expression, Cell death & disease, 8 (2017) e2997.
[58] D. Cao, Q. Ding, W. Yu, M. Gao, Y. Wang, long noncoding rna SPRY4-IT1 promotes malignant development of colorectal cancer by targeting epithelial–mesenchymal transition, OncoTargets and therapy, 9 (2016) 5417.
[59] L. Wang, Y. Cai, X. Zhao, X. Jia, J. Zhang, J. Liu, H. Zhen, T. Wang, X. Tang, Y. Liu, Down-regulated long non-coding RNA H19 inhibits carcinogenesis of renal cell carcinoma, Neoplasma, 62 (2015) 412-418.
[60] H.-T. He, M. Xu, Y. Kuang, X.-Y. Han, M.-Q. Wang, Q. Yang, Biomarker and competing endogenous RNA potential of tumor-specific long noncoding RNA in chromophobe renal cell carcinoma, OncoTargets and therapy, 9 (2016) 6399.
[61] Y. Qu, H. Xiao, W. Xiao, Z. Xiong, W. Hu, Y. Gao, Z. Ru, C. Wang, L. Bao, K. Wang, Upregulation of MIAT regulates LOXL2 expression by competitively binding MiR-29c in clear cell renal cell carcinoma, Cellular Physiology and Biochemistry, 48 (2018) 1075-1087.
[62] F. Liu, N. Chen, Y. Gong, R. Xiao, W. Wang, Z. Pan, The long non-coding RNA NEAT1 enhances epithelial-to-mesenchymal transition and chemoresistance via the miR-34a/c-Met axis in renal cell carcinoma, Oncotarget, 8 (2017) 62927.
[63] P. Wang, Y. Wu, X. Zhong, B. Liu, G. Qiao, Prognostic significance of overexpressed long non-coding RNA TUG1 in patients with clear cell renal cell carcinoma, Eur Rev Med Pharmacol Sci, 21 (2017) 82-86.
[64] M. Seles, G. Hutterer, T. Kiesslich, K. Pummer, I. Berindan-Neagoe, S. Perakis, D. Schwarzenbacher, M. Stotz, A. Gerger, M. Pichler, Current insights into long non-coding RNAs in renal cell carcinoma, International journal of molecular sciences, 17 (2016) 573.
[65] T. Yang, H. Zhou, P. Liu, L. Yan, W. Yao, K. Chen, J. Zeng, H. Li, J. Hu, H. Xu, lncRNA PVT1 and its splicing variant function as competing endogenous RNA to regulate clear cell renal cell carcinoma progression, Oncotarget, 8 (2017) 85353.
[66] X. Zhang, M.A. Hildebrandt, Y. Horikawa, X. Pu, C.G. Wood, J. Gu, The expression of RN7SK paralogs in tumors of renal cell carcinoma is associated with patient survival, (AACR2011).
[67] M. Zeng, C. Lu, F. Zhang, Z. Lu, F. Wu, Y. Li, M. Li, LncRNA–disease association prediction through combining linear and non-linear features with matrix factorization and deep learning techniques, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)2019), pp. 577-582.

Min Zeng: Conceptualization, Methodology, Software, Validation, Writing - Original draft preparation, Writing - Reviewing and Editing.
Chengqian Lu: Data curation, Software, Writing - Original draft preparation.
Fuhao Zhang: Software.
Yiming Li: Visualization.

Fang-Xiang Wu: Writing - Original draft preparation, Writing - Reviewing and Editing.
Yaohang Li: Writing - Reviewing and Editing.
Min Li: Supervision, Conceptualization, Writing - Original draft preparation, Writing - Reviewing and Editing.

Highlights:

- A novel hybrid computational framework (SDLDA) is proposed for lncRNA–disease association prediction.
- It is the first time to combine traditional matrix factorization and deep learning techniques to extract linear and non-linear features in the prediction of lncRNA–disease associations.
- Results show that SDLDA outperforms existing computational methods. In addition, case studies on three diseases illustrate the capability of SDLDA.