Deep matrix factorization improves prediction of human circRNA-disease associations

Chengqian Lu, Min Zeng, Fuhao Zhang, Fang-Xiang Wu, Min Li, and Jianxin Wang*

***Abstract*—In recent years, more and more evidence indicates that circular RNAs (circRNAs) with covalently closed loop play various roles in biological processes. Dysregulation and mutation of circRNAs may be implicated in diseases. Due to its stable structure and resistance to degradation, circRNAs provide great potential to be diagnostic biomarkers. Therefore, predicting circRNA-disease associations is helpful in disease diagnosis. However, there are few experimentally validated associations between circRNAs and diseases. Although several computational methods have been proposed, precisely representing underlying features and grasping the complex structures of data are still challenging. In this paper, we design a new method, called DMFCDA (Deep Matrix Factorization CircRNA-Disease Association), to infer potential circRNA-disease associations. DMFCDA takes both explicit and implicit feedback into account. Then, it uses a projection layer to automatically learn latent representations of circRNAs and diseases. With multi-layer neural networks, DMFCDA can model the non-linear associations to grasp the complex structure of data. We assess the performance of DMFCDA using leave-one cross-validation and 5-fold cross-validation on two datasets. Computational results show that DMFCDA efficiently infers circRNA-disease associations according to AUC values, the percentage of precisely retrieved associations in various top ranks, and statistical comparison. We also conduct case studies to evaluate DMFCDA. All results show that DMFCDA provides accurate predictions.**

***Index Terms*—Circular RNA, Disease, Deep Matrix Factorization.**

I. Introduction

DERIVED from back-spliced precursor mRNAs, circular RNAs (circRNAs) with covalently closed loop are specific forms of single-stranded endogenous non-coding RNAs [1], [2]. Discovered in an electron microscopy-based research of viroids [3], circRNAs were once assumed to be aberrant resultants of incorrect or erratic RNA splicing due to their low levels of expression [4]. Recent years, accumulating evidences show that circRNAs have been identified in diverse biological processes [5]. According to their locations, they are categorized as intronic circRNAs, intergenic circRNAs, exon-intron circRNAs and exonic circRNAs [6]. In contrast to linear RNAs, circRNAs are more stable and resistant to exoribonucleases lacking the terminal structures (e.g., 3' poly A tail or 5' cap structure), consequently escaping normal RNA degradation [7]. Therefore, the abundance, multifariousness, and conservation of circRNAs signify that at least some circRNAs will perform biological functions. To date, it has been proved that circRNAs serve as intermediates in RNA processing reactions, regulators of transcription, miRNA sponges [8].

Partaking in a diversity of organic processes, the dysregulation and mutation of circRNAs may give rise to the progression of diseases, including atherosclerosis [9], Alzheimer's disease [10], [11], cancers [12] and so on. For instance, circZNF609 involved in tumorigenesis acts as a vying endogenous sponge and regulates translational repression of AKT3 in Hirschsprung's disease [13]. CircRNA hsa_circ_0000096 changes cell proliferation and metastasis of gastric cancer through suppressing the expression levels of cyclin-dependent kinase 6, matrix metalloproteinase (MMP)-2 and MMP-9 [14]. On account of tissue-specific expression patterns, highly conserved characteristics and various biological roles, circRNAs provide great potential to be diagnostic biomarkers and therapeutic targets of diseases. However, molecular mechanisms and functions of circRNAs during diseases initiation and progression are still far from understanding. Meanwhile, biological experiments to validate the relationships between circRNAs and diseases are costly and time-consuming. Hence, designing computational models to offer probable associations is not beneficial for learning biological mechanisms, but also for disease diagnosis.

In recent past, several computational methods, divided into two categories, were proposed to predict circRNA-disease associations. The first category applies the network-based method to infer potential associations. Fan *et al.* utilized the KATZ method to compute the possibility of associations between diseases and circRNAs in the constructed heterogeneous network, which is based on circRNA expression profiles, Gaussian interaction profile kernel similarity and disease phenotype similarity [15]. However, there are few recorded expression profiles of circRNAs that makes the constructed network sparse and impedes the prediction. Lei *et al*. designed a path weighted model to infer associations on the basis of the heterogeneous network, composed of circRNA-disease network, circRNA similarity network and disease similarity network [16]. Lack of known circRNA-related information, it is difficult to integrate various knowledge into a heterogeneous network in a proper way. The second category uses machine learning model. Yan *et al.* developed a regularized least squares method to predict associations based on Kronecker product kernel. It suffers from determining a proper number of neighbors, which is used to calculate an initial association

C. Lu, M. Zeng, F. Zhang, M. Li and J. Wang are with the School of Computer Science, Central South University, Changsha 410083, China. (E-mail: {chengqlu, zengmin, fhzhang}@csu.edu.cn, and {limin, jxwang}@mail.csu.edu.cn)

F. Wu is with the Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada. (E-mail: faw341@mail.usask.ca.)

* Corresponding author.

Manuscript received ***, 2020; revised ***, ***.

score [17]. Actually, matrix factorization methods have been successfully applied in predicting biomolecular associations. Based on the expression profiles of lncRNAs and mRNAs in ovarian cancer, Xiao *et al*. proposed a joint orthogonal non-negative matrix factorization to distinguish lncRNA-mRNA co-expression models [18]. Pan *et al*. used a self-weighted multi-kernel learning framework for recommending miRNA-disease associations [19]. Xiao *et al*. presented a non-negative matrix factorization with the graph regularization on heterogeneous omics data to extrapolate potential miRNA-disease associations [20]. Zitnik *et al*. designed a multi-level hierarchy to detect associations among diseases by means of fusing system biological data [21]. Fu *et al*. fused heterogeneous data to discover latent lncRNA-disease associations based on a matrix tri-factorization framework [22]. Zitnik *et al*. offered a collective matrix factorization to elicit different semantics and discover modules of gene-disease objects by fusing multi-modal biological data [23]. Reformulating association matrix by the constructed circRNA and disease similarity, Wei *et al*. proposed a graph regularization non-negative matrix factorization algorithm to predict associations [24]. After constructing a heterogeneous circRNA-disease bilayer network, Xiao *et al*. used a weighted dual-manifold regularization low-rank approximation algorithm to recommender associations [25]. Nonetheless, the performance of the model depended upon the hand-crafted feature extractors. In the traditional matrix factorization model, latent factors are manual features that are not very consistent with the model. In addition, associations are produced by a linear multiplication that is not sufficient to seize non-linear structures of data.

Since the great breakthrough in 2012 ImageNet competition [26], deep learning has been successfully applied in different domains including natural language processing, visual object recognition and bioinformatics [27]–[29]. Consisting of multiple neural layers, deep learning models can automatically grasp data representations through multiple levels of abstraction [30]. Inspired by the application of recommender system methods in association prediction [31], we propose a deep matrix factorization model (DMFCDA) to recommend potential circRNAs for investigated diseases. Implicit feedback, a.k.a. unknown associations, could improve the performance of prediction [32]. In this study, a deep learning framework is proposed to predict the potential cricRNA-disease associations. Specifically, we construct a cricRNA-disease matrix with all explicit and implicit feedback. A project layer is applied to capture dense non-linear representations of circRNAs and diseases. Dense representations are automatically learned and consistent with the model. Instead of the linear multiplication of latent features in the conventional matrix factorization, we utilize multi-layer neural networks to grasp complex associations. Experimental results imply that DMFCDA performs better than the-state-of-art computational methods.

II. Materials and methods

A. Data description

In this study, we make use of two datasets to evaluate the effectiveness of DMFCDA. The first one is retrieved from CircR2Disease [33], including 739 associations between 676 circRNAs and 100 diseases. The second one is downloaded from LncRNADisease v2.0 [34], including 1,004 associations between 811 circRNAs and 112 diseases. Lack of standard nomenclature, circRNAs with the same sequence were given different names in different databases. In order to unify the names of circRNAs, we refer to standard databases, like circBase [35] and deepBase v2.0 [36]. We find sequences of circRNAs from circBase and deepBase, and then uniformly name circRNAs with the same sequence according to circBase. For diseases, there is a naming confusion problem. We refer to UMLS [37], OMIM [38] and NCBI, and find the description of disease symptoms. Diseases with the same symptoms are uniformly named according to UMLS. After that, we delete all the repeated associations and all the associations with unrecorded circRNAs or diseases. At last, we get 619 associations between 556 circRNAs and 80 diseases for dataset1. In contrast to dataset1, there are 76 added circRNAs, 9 added diseases and 125 added associations. For dataset2, we get 744 associations between 632 circRNAs and 89 diseases. The details of the dataset are shown in Table I.

TABLE I
Details of dataset

	# of circRNAs	# of diseases	# of associations
Dataset1	556	80	619
Dataset2	632	89	744

B. Problem formulation

Let $C = \{c_1, c_2, c_3, ..., c_m\}$ be the set of m circRNAs, and $D = \{d_1, d_2, d_3, ..., d_n\}$ be the set of n diseases. Let $\boldsymbol{A} \in \boldsymbol{R}^{m \times n}$ be a circRNA-disease association matrix. If the association between circRNA i and disease j has been experimentally verified, $\boldsymbol{A}_{ij}$ is 1; otherwise, then $\boldsymbol{A}_{ij}$ is 0.

$$\boldsymbol{A}_{ij} = \begin{cases} 1, \text{ if circRNA } i \text{ is associated with disease } j; \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

The problem of circRNA-disease association prediction is to infer unknown associations based on observed associations. It can be considered as a recommender system problem. Matrix factorization methods successfully applied in recommender systems can solve the problem. A matrix factorization model maps both features of circRNAs and diseases to a joint latent factor space of low-rank dimensionality in this system. Besides, brand-new recommendations could be made for circRNAs and diseases. Till now, limited associations are already experimentally verified. Therefore, it is a challenge to recommend circRNAs to the investigated diseases based on a sparse association matrix. In our model, we concentrate on the association matrix without importing extra biological knowledge to solve the problem under the general situation.

C. Deep matrix factorization

With the successful application of Netflix Prize, it is found that the preferences of users to movies are dominated by only a few latent factors of users and movies. Matrix factorization (MF), the most popular model in the recommender system, is based on the latent factor model [39]. MF intends to find latent representations of users and movies, respectively, and utilizes the inner product of the learned representations to approximate the preference. Latent factors of users and movies in a shared latent factor space of dimensionality d. More specifically, user i is related to a vector u_i, $u_i \in \boldsymbol{R}^d$, quantifying the extent to which the user possesses positive or negative factors; movie j is related to a vector v_j, $v_j \in \boldsymbol{R}^d$, quantifying the extent to which the movie possesses positive or negative factors. The inner product $u_i^T v_j$ approximates the observed rating r_{ij} of user i on movie j.

In order to train the latent models, MF usually minimizes a loss function $\mathcal{L}$, made up of sum-of-squared-error terms between the computed ratings and the real ratings and L_2 regularized terms obviating the over-fitting problem as the following:

$$\mathcal{L} = \sum_{i}^{m} \sum_{j}^{n} (r_{ij} - u_i^T v_j)^2 + \lambda_u \sum_{i}^{m} \|u_i\|^2 + \lambda_v \sum_{j}^{n} \|v_j\|^2, \tag{2}$$

where λ_u and λ_v are the parameters used to balance the regularization term of users and items, respectively.

Enlightened by the previous work using neural networks to matrix factorizations [32], we propose a deep matrix factorization model, called DMFCDA. DMFCDA extends the basic latent factor model for non-linear representations rather than linear representations. The projection composed of feed-forward neural networks maps raw representations including explicit and implicit feedback to dense representations. Dense representations are learned while optimizing the model. Besides, DMFCDA approximates the associations with multi-layer neural networks instead of the linear multiplication-inner product for non-linear structures of data. There are three steps in DMFCDA shown in Fig. 1. First, we extract row vectors or column vectors as raw representation of circRNAs or diseases, respectively. Each row vector or column vector contains association patterns for each circRNA or each disease, respectively. Then, each circRNA c_i is represented as a high-dimensional vector of A_{i*}, which corresponds to the associations of ith circRNA with all diseases. Each disease d_j is represented as a high-dimensional vector of A_{*j}, which corresponds to the jth disease's associations with all circRNAs. In the association matrix $\boldsymbol{A}$, there are explicit feedback represented as 1s, and implicit feedback represented as 0s. A value of 1 means that the association is experimentally verified now, while value of 0 means that the association is not verified so far, rather than it does not exist. Only explicit feedback, a.k.a. observed associations, are insufficient to make a good recommendation [40]. Implicit feedback made up of some association patterns can improve the performance [32], [41]. If the association is unknown, we mark a zero as an implicit feedback. Explicit and implicit feedback consist of vectors of circRNAs and diseases. Secondly, we feed the raw representations into a projection layer composed of three fully connected neural networks to project feature vectors. Stimulated by the latent factor model [42], we take use of a projection layer to learn non-linear representations of circRNAs and diseases. Formally, we use $\mathbf{p}$ and $\mathbf{q}$ to denote the input vectors of circRNAs and diseases, respectively. In each fully connected layer for circRNAs, W_{p1}, W_{p2}, W_{p3} denote the weight matrices, b_{p1}, b_{p2}, b_{p3} denote the bias terms, and O_{p1}, O_{p2}, O_{p3} denote the corresponding outputs. In each fully connected layer for diseases, W_{q1}, W_{q2}, W_{q3} denote the weight matrices, b_{q1}, b_{q2}, b_{q3} denote the bias terms, and O_{q1}, O_{q2}, O_{q3} denote the corresponding outputs. We use the *ReLU* as the activation function of hidden layers.

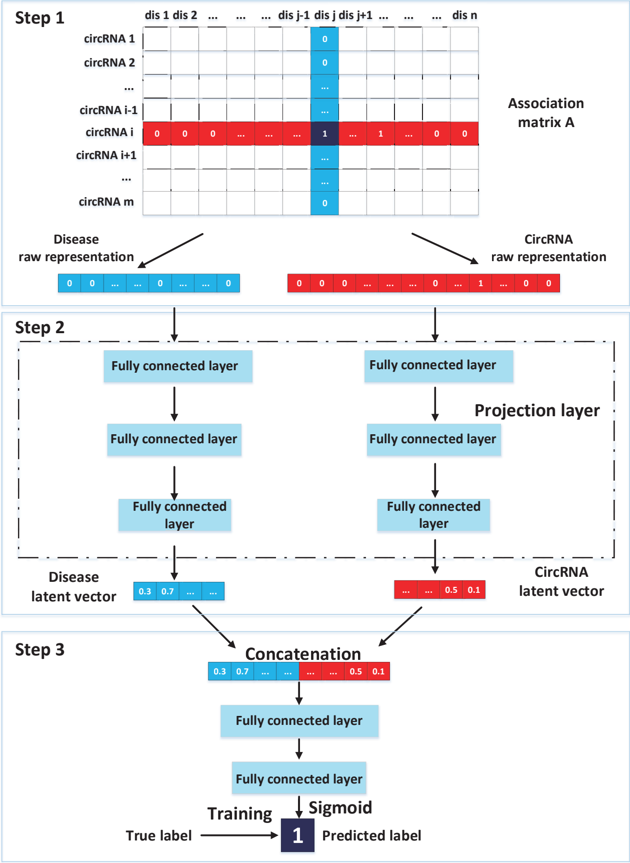


Fig. 1. Overview of DMFCDA model architecture. The input consists of two vectors, cricRNA i's vector and disease j's vector. These two vectors are raw representations of circRNA i and disease j. They are fed into a projection layer which consists of three fully connected layers. We concatenate the two projected latent vectors as a new vector, and use it to feed two fully connected layers. At last, we use sigmoid function to classify the label.

The outputs of the first fully connected layer can be calculated as:

$$O_{p1} = \mathrm{ReLU}(W_{p1}p + b_{p1}) \tag{3}$$
$$O_{q1} = \mathrm{ReLU}(W_{q1}q + b_{q1}) \tag{4}$$

The outputs of the second fully connected layer can be calculated as:

$$O_{p2} = \mathrm{ReLU}(W_{p2}O_{p1} + b_{p2}) \tag{5}$$
$$O_{q2} = \mathrm{ReLU}(W_{q2}O_{q1} + b_{q2}) \tag{6}$$

The outputs of the third fully connected layer can be calculated as:

$$O_{p3} = \text{ReLU}(W_{p3}O_{p2} + b_{p3}) \tag{7}$$

$$O_{q3} = \text{ReLU}(W_{q3}O_{q2} + b_{q3}) \tag{8}$$

where O_{p3} and O_{q3} are the latent feature vectors of circRNAs and diseases. For the purpose of averting the overfitting problem, we use the dropout on each fully connected layer. We set the dropout rate to 0.005 as default.

Thirdly, we concatenate O_{p3} and O_{q3} into a combined vector, and feed it to two fully connected layers. At last, we utilize sigmoid function to classify the label. Multi-layer neural networks replace linear multiplications to seize non-linear structures of data.

$$O_4 = \text{ReLU}(W_4 \cdot Concatenate(O_{p3}, O_{q3}) + b_4) \tag{9}$$

$$\hat{y} = sigmoid(\text{ReLU}(W_5O_4 + b_5) \tag{10}$$

where O_4, W_4 and b_4 are the output, weight matrix and bias of the fourth layers. W_5 and b_5 are the weight matrix and bias of the fifth layer. $\hat{y}$ is the predicted label.

We substitute the square loss function with a binary cross-entropy loss function. A binary cross-entropy loss function can evaluate the recoverability of explicit and implicit feedback rather than only explicit feedback. The loss functioin is depicted as follows:

$$Loss = -\left(\sum y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right) + \lambda(\|W\|_2)^2 \tag{11}$$

where y is the observed label, λ is the regularization coefficient, and $\| \boldsymbol{W} \|_2$ is the L_2 norm of weight matrix.

In the process of training, it is worth noting that the intersection of the two raw interaction representations is the true label of the interaction between a circRNA and a disease. To avoid using true labels in training and testing, prior knowledge should be removed. The value of the intersection is masked with 0. Then the true label is deleted from raw interaction representations of a circRNA and a disease. Furthermore, we take all observed interactions as positive samples and all unknown associations as negative samples. The number of negative samples is larger than the number of positive samples, which results in the imbalanced problem. The imbalanced data may mislead the model optimization. Thus, we randomly sample some positive samples from all observed interactions and select the same number of negative samples at each batch. Such a sampling method is not biased to any class in each training batch. This process is carried out to train the model. Calculating the gradient dominates the computational cost in each iteratior of DMFCDA. Based on the back-propagation, the main operation of calculating the gradient is matrix multiplication. The flop count of an $m \times n$ matrix multiplying a $n \times r$ matrix is $2mnr$. Thus, calculating the gradient requires $4m\left(f_1h_1 + h_1h_2 + \cdots + h_kn\right)$ flops, where f_1 is the dimension of the input vector, k is the number of hidden layers.

III. Results and Discussions

In this section, we firstly introduce evaluation metrics to assess the performance of DMFCDA. Secondly, we analyze the effects of parameters in the model. Thirdly, we compare the performance of DMFCDA with other methods. At last, we conduct case studies to verify the effectiveness on breast cancer and gastric cancer.

A. Evaluation metrics

To access the effectiveness of DMFCDA, we conduct leave-one out cross validation (LOOCV) and 5-fold cross validation (5-CV). In LOOCV, each positive sample is left out as the test sample in turn. Other positive samples and the same amount of randomly selected negative samples are considered as the training samples. In 5-CV, we divide all the observed samples into five folds. Each fold is considered as the test samples in turn. Other left folds and the same amount of extracted negative samples are considered as the training samples. After calculating the probabilities of all test samples, we rank the test samples by the probabilities in descending order. Then, we figure out the true positive rate (TPR) and the false positive rate (FPR) as follows:

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

where TP is the number of positive samples that are classified correctly, and FN is the number of negative samples that are classified incorrectly.

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

where FP is the number of negative samples that are classified incorrectly, and TN is the number of negative samples that are classified correctly.

TPR measures the proportion of actual positive samples that are correctly identified. FPR measures the proportion of actual negative samples that are incorrectly identified. The receiver operating characteristic (ROC) curve is used to show the predictive accuracy. The area under the ROC (AUC) is used to measure the overall performance of the prediction methods.

In addition, the percentage of correctly retrieved associations in various top rank provides more actual guidance for biologists. The higher the percentage is, the more accurate the recommendation is. In this study, we take it as an important metrics for assessing the efficiency.

B. Effects of parameters

1) Number of projection layers: For projecting raw representation of circRNAs and diseases to dense representation, we design a projection layer. The number of projection layers plays an important role in learning feature vectors. We perform the cross validation and grid search for the optimal number by automatically training the model from 2 layers to 6 layers, in steps of 1. After analyzing the effects of different numbers of projection layers on two datasets, we set its value to 3 as default. The detailed results are shown in Table II.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2999638, IEEE Journal of Biomedical and Health Informatics

TABLE II
NUMBER OF PROJECTION LAYERS

# of layers	2	**3**	4	5	6
AUC on dataset1	0.859	**0.8679**	0.857	0.825	0.818
AUC on dataset2	0.8824	**0.8861**	0.8843	0.8839	0.8832

TABLE III
DIMENSIONALITY OF LATENT FEATURE VECTOR

# of dimensionality	8	16	**32**	48	64
AUC on dataset1	0.764	0.825	**0.8679**	0.857	0.834
AUC on dataset2	0.8745	0.8796	**0.8861**	0.8812	0.8756

TABLE IV
THE EFFECT OF λ

λ	0.1	0.01	**0.001**	0.002	0.003	0.005	0.007
AUC on dataset1	0.694	0.8312	**0.8679**	0.862	0.856	0.853	0.8496
AUC on dataset2	0.6721	0.7403	**0.8861**	0.8713	0.8697	0.8566	0.8501

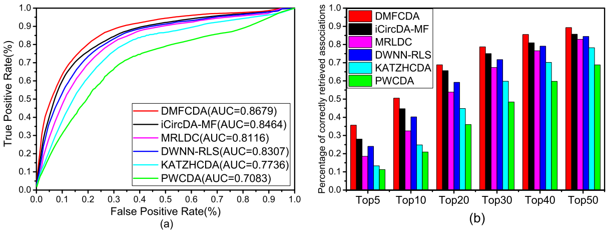


Fig. 2. Comparison of predicting methods on dataset1. (a) Performance of all methods in terms of ROC curve using LOOCV. (b) Percentage of correctly retrieved associations in various top rank in LOOCV.

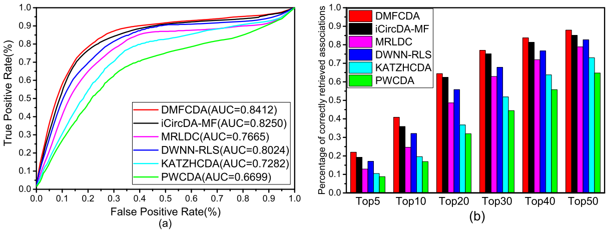


Fig. 3. Comparison of predicting methods on dataset1. (a) Performance of all methods in terms of ROC curve using 5-CV. (b) Percentage of correctly retrieved associations in various top rank in 5-CV.

2) Dimensionality of latent feature vector: The relevance of circRNAs and diseases is produced by latent feature vectors in the common low-dimensional space. Therefore, the dimensionality of latent feature vectors is vital to the method. It is determined by the number of last fully connected layers in the projection layers. We perform the cross validation and grid search for the optimal dimension range from 8 to 64. We choose 32 as the default number which yields the highest accuracy. The results are shown in Table III.

3) The effect of λ: In Equation (11), the parameter λ is utilized to weight the binary cross-entropy loss and the regularization term. It can alleviate the overfitting problem. We perform the cross validation and grid search for the optimal dimension range from 0.001 to 0.1. We choose 0.001 as default number which yields the highest accuracy. The results are shown in Table IV.

C. Comparison of methods

1) Accuracy Comparison: To measure the performance of prediction, we compare DMFCDA with five the-state-of-art computational methods on two datasets: iCircDA-MF [24], MRLDC [25], DWNN-RLS [17], KATZHCDA [15] and PWCDA [16]. Fig. 2(a) displays that the ROC curves of DMFCDA (red), iCircDA-MF (black), MRLDC (magenta), DWNN-RLS (blue), KATZHCDA (cyan) and PWCDA (green) on dataset1 obtained with LOOCV. DMFCDA achieves an AUC value of 0.8679, which performs better than others (iCircDA-MF 0.8464, MRLDC 0.8116, DWNN-RLS 0.8307, KATZHCDA 0.7736, PWCDA 0.7083). It means that DMFCDA can provide more accurate prediction than others. From Fig. 2(b), DMFCDA obtains higher percentage of correctly retrieved associations than others in top 5, 10, 20, 30, 40 and 50 with LOOCV. It signifies that DMFCDA gives more accuracy than others. Fig. 3(a) displays that DMFCDA achieves an AUC value of 0.8412, which outperforms others (iCircDA-MF 0.8250, MRLDC 0.7665, DWNN-RLS 0.8024, KATZHCDA 0.7282, PWCDA 0.6699) on dataset1 with 5-CV. Fig. 3(b) presents that DMFCDA can retrieve more true associations than others on dataset1 with 5-CV. Fig. 4(a) shows that the ROC curves on dataset2 with LOOCV, DMFCDA with an AUC value of 0.8861 outperforms other methods (iCircDA-MF 0.8582, MRLDC 0.8251, DWNN-RLS 0.8454, KATZHCDA 0.7881, PWCDA 0.7171). To further assess the effectiveness of DMCLDA, we conduct the experiments in comparing the percentage of correctly retrieved associations in various top ranks. It is obviously found that DMFCDA outperforms than others on datset2 with higher percentage of correctly retrieved associations in top 5, 10, 20, 30, 40 and 50 shown in Fig. 4(b). Fig. 5(a) reveals that DMFCDA with an AUC value of 0.8588 performs better than others (iCircDA-MF 0.8354, MRLDC 0.8010, DWNN-RLS 0.8180, KATZHCDA 0.7752, PWCDA 0.7036). From Fig. 5(b), DMFCDA provides higher percentage of correctly retrieved associations.

In addition, we replace the traditional cosine similarity approximation method with multi-layer neural networks to be general. In order to test the validity of the change, we add a comparison with the method DMF that utilizes cosine similarity measurement [32]. The details of numeric comparison are shown in Table V. We can see that DMFCDA is more stable with lowest standard deviations. It is worth noting that our method has improved the accuracy of prediction. In summary, DMFCDA achieves more accurate prediction than other methods.

2) Statistical Comparison: To statistically evaluate the quality of the model, we make use of the Friedman test to detect the significant difference and Nemenyi post-hoc to find which pairs are significantly different [43]. Specifically, we

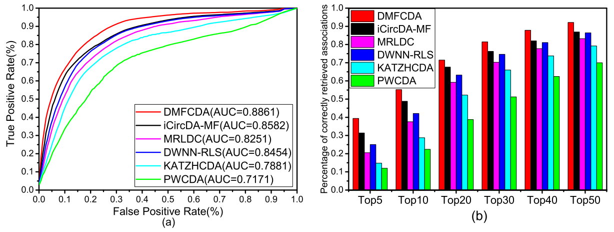


Fig. 4. Comparison of predicting methods on dataset2. (a) Performance of all methods in terms of ROC curve using LOOCV. (b) Percentage of correctly retrieved associations in various top rank in LOOCV.

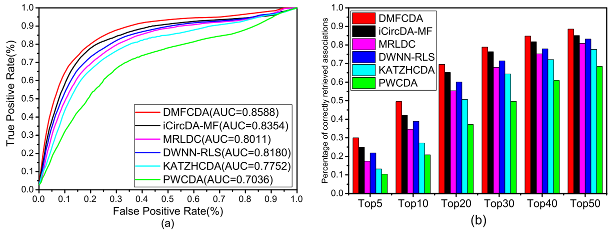


Fig. 5. Comparison of predicting methods on dataset2. (a) Performance of all methods in terms of ROC curve using 5-CV. (b) Percentage of correctly retrieved associations in various top rank in 5-CV.

TABLE V
COMPARISON WITH OTHER METHODS IN LOOCV AND 5-CV

Comparison Methods	LOOCV		5-CV	
	dataset1	dataset2	dataset1	dataset2
DMFCDA	0.8679	0.8861	0.8412±0.017	0.8588±0.008
DMF	0.8510	0.8681	0.8300±0.029	0.8411±0.019
iCircDA-MF	0.8464	0.8582	0.8250±0.046	0.8354±0.026
MRLDC	0.8116	0.8251	0.7665±0.044	0.8011±0.039
DWNN-RLS	0.8307	0.8454	0.8024±0.051	0.8180±0.044
KATZHCDA	0.7736	0.7881	0.7282±0.068	0.7752±0.057
PWCDA	0.7083	0.7171	0.6699±0.074	0.7036±0.062

randomly select parts of the positive samples as test samples and utilize the optimized models to calculate their values. For the results, we use statistical methods to find significant differences. The statistic for Friedman test that is a non-parametric analysis of variance is a Chi-square with 5 (number of methods -1) degrees of freedom. If the Chi-square is bigger than 11.07 (the p-value is 0.05), there is a significant difference among methods. Then, we use Nemenyi post-hoc to determine which pairs of models are significantly different with the critical distance (CD) as a standard,

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6M}} \tag{14}$$

where q_α is a fixed critical value 2.85 while $\alpha = 0.05$, k is the degrees of freedom and M is the number of data sets. The value of CD is 3.65 considered as the significance level. For dataset1, the value of Chi-square of Friedman is 21.78 bigger than the threshold of 11.07, which means that there is a significant difference. From Table VI, we can see that the critical distances between DMFCDA and other methods are bigger than 3.65, indicating that DMFCDA is significantly different from others. For dataset2, the value of Chi-square of Friedman is 24.53 bigger than the threshold of 11.07, which means that there is a significant difference. From Table VII, we can see that the critical distances between DMFCDA and other methods are bigger than 3.65, indicating that DMFCDA is significantly different from others.

TABLE VI
CRITICAL DISTANCE ON DATASET1

	DMFCDA	iCircDA-MF	MRLDC	DWNN-RLS	KATZHCDA	PWCDA
DMFCDA	NA	4.71	4.90	5.37	6.80	7.14
iCircDA-MF	4.71	NA	1.75	2.35	4.21	4.62
MRLDC	4.90	1.75	NA	3.43	4.02	5.17
DWNN-RLS	5.37	2.35	3.43	NA	5.63	5.3
KATZHCDA	6.80	4.21	4.02	5.63	NA	1.06
PWCDA	7.14	4.62	5.17	5.3	1.06	NA

TABLE VII
CRITICAL DISTANCE ON DATASET2

	DMFCDA	iCircDA-MF	MRLDC	DWNN-RLS	KATZHCDA	PWCDA
DMFCDA	NA	3.71	4.20	4.58	5.34	6.83
iCircDA-MF	3.71	NA	1.95	1.89	4.74	4.83
MRLDC	4.20	1.95	NA	3.63	4.41	5.34
DWNN-RLS	4.58	1.89	3.63	NA	5.12	4,75
KATZHCDA	5.34	4.74	4.41	5.12	NA	1.27
PWCDA	6.83	4.83	5.34	4.75	1.27	NA

D. *Validation across datasets*

In order to evaluate the effectiveness of DMFCDA, we perform prediction on dataset1 and verify the predicted results with the experimentally validated associations in dataset2. Owing to differences in the recorded circRNAs and diseases between two datasets, the predicted associations whose ends existing in both datasets can be validated. There are 41 added associations in dataset2 that meet the condition. DMFCDA can predict 32 experimentally validated associations with dataset1. The validated results in Table VIII show DMFCDA has the potential to provide the accurate prediction.

E. *Case studies*

To illustrate the capability of DMFCDA, we conduct case studies on dataset2. We fed the model with the known associations as training samples to predict potential associations for colorectal cancer, hepatocellular carcinoma and lung cancer. We then sorted the predicted associations according to calculated possibilities. Top-10 ranking results are manually validated by mining existing literatures. If the predicted associations are confirmed in the existing literature, we provide the corresponding reference. In the same way, we get all the result lists.

Colorectal cancer is the second leading cause of cancer death [44]. It is attractive to assure associations between

TABLE VIII
THE DETAILS OF VALIDATION ON DATASET2

	Added associations meet the condition in Dataset2		Validated
	CircRNAs	Diseases	
1	hsa_circ_0000520	stomach cancer	YES
2	hsa_circ_0001727	stomach cancer	YES
3	hsa_circ_100782	non-small cell lung carcinoma	YES
4	hsa_circ_0082582	non-small cell lung carcinoma	YES
5	hsa_circ_0061265	liver cancer	YES
6	hsa_circ_0000096	colorectal cancer	YES
7	hsa_circ_0023404	urinary bladder cancer	YES
8	hsa_circ_0001313	colorectal cancer	YES
9	hsa_circ_0004771	stomach cancer	YES
10	hsa_circ_0072088	hepatocellular carcinoma	YES
11	hsa_circ_0001006	stomach cancer	YES
12	hsa_circ_0003221	urinary bladder cancer	YES
13	hsa_circ_0041103	stomach cancer	YES
14	hsa_circ_103595	cancer	YES
15	hsa_circ_0000437	hypertension	NO
16	hsa_circ_0000437	stomach cancer	YES
17	hsa_circ_0075829	stomach cancer	YES
18	hsa_circ_0000745	colon cancer	YES
19	circ-ZFR	lung cancer	YES
20	hsa_circ_0004712	Triple Negative Breast Neoplasms	NO
21	hsa_circ_0003570	stomach cancer	NO
22	hsa_circ_0000284	cancer	YES
23	hsa_circ_0000284	urinary bladder cancer	YES
24	hsa_circ_0000284	colon cancer	YES
25	hsa_circ_0001451	esophageal cancer	NO
26	CDR1-AS	cancer	YES
27	CDR1-AS	cholangiocarcinoma	YES
28	hsa_circ_0067934	stomach cancer	YES
29	hsa_circ_0001649	urinary bladder cancer	YES
30	circ-ITCH	bladder carcinoma	YES
31	circ-ITCH	urinary bladder cancer	YES
32	hsa_circ_0002768	urinary bladder cancer	YES
33	hsa_circ_0004214	cardiovascular system disease	NO
34	hsa_circ_0001821	acute myeloid leukemia	NO
35	hsa_circ_0001821	urinary bladder cancer	NO
36	hsa_circ_0005273	hepatocellular carcinoma	YES
37	hsa_circ_0007158	lung cancer	YES
38	hsa_circ_0000118	colon cancer	YES
39	hsa_circ_0000140	colorectal cancer	YES
40	circANRIL	cardiovascular system disease	NO
41	hsa_circ_0005075	urinary bladder cancer	NO

TABLE IX
TOP TEN CANDIDATE CIRCRNAS FOR COLORECTAL CANCER

Rank	Name of circRNAs	References
1	hsa_circ_0001649	[45]
2	CDR1-AS	[46]
3	hsa_circ_0000615	Circad [47]
4	hsa_circ_0023404	[48]
5	circPVT1	[49]
6	hsa_circ_103809	Unknown
7	hsa_circ_0000437	Unknown
8	hsa_circ_0003221	Unknown
9	hsa_circ_103595	Unknown
10	hsa_circ_0005075	[50]

TABLE X
TOP TEN CANDIDATE CIRCRNAS FOR HEPATOCELLULAR CARCINOMA

Rank	Name of circRNAs	References
1	hsa_circ_0001451	Unknown
2	hsa_circ_0001821	Unknown
3	hsa_circ_0000745	[52]
4	hsa_circ_0082582	Unknown
5	hsa_circ_0072088	[53]
6	hsa_circ_103595	Unknown
7	hsa_circ_0000437	Unknown
8	circPTK2	Unknown
9	hsa_circ_0000118	MalaCards [54]
10	hsa_circ_0002768	[55]

circRNAs and colorectal cancer. We check whether 10 most likely associations between inferred circRNAs and colorectal cancer are verified by mining existing literatures. The 6 inferred circRNAs in the top-10 rank have been checked, indicating that associations between them and colorectal cancer are verified (Table IX, hsa_circ_0001649, 1st, CDR1-AS, 2nd, hsa_circ_0000615, 3rd, hsa_circ_0023404, 4th, circPVT1, 5th, hsa_circ_0005075, 10th). The expression level of circRNA hsa_circ_0001649 is down-regulated in tissue and serum samples from colorectal cancer patients [45]. CircRNA CDR1-AS (ciRS-7) is significantly up-regulated in colorectal cancer tissues. Its over-expression is related to poor patient survival [46]. The association between hsa_circ_0000615 and colorectal cancer is recorded in a published database Circad [47]. Knockdown of hsa_circ_0023404 significantly promotes expression level of miR-36, which has effects on suppressing colorectal cancer [48]. CircRNA circPVT1 is one of top 10 dysregulated circRNAs in colorectal cancer [49]. By activating Wnt/β-catenin pathway, hsa_circ_0005075 acts as tumor-promotive oncogene in colorectal cancer [50].

With affecting more than 500,000 people, hepatocellular carcinoma (HCC) is the third leading cause of cancer deaths. The abnormal expression levels of circRNAs may be related to the occurrence of hcc [51]. The 5 inferred circRNAs in the top-10 rank have been verified as shown in Table X (hsa_circ_0000745, 3rd, hsa_circ_0072088, 5th, hsa_circ_0000118 9th and hsa_circ_0002768 10th). CircRNA hsa_circ_0000745 is involved in HCC [52]. CircRNA hsa_circ_0072088 associates with miR-620 and restrain the propagation and aggression of HCC [53]. The association between hsa_circ_0000118 and HCC is recorded in a published database MalaCards [54]. By regulating cytoskeleton, MYLK (hsa_circ_0002768) furthers the progression of HCC to augment epithelial-mesenchymal transition [55].

Lung cancer is one cause of cancer death in both men and women, with more than 1 million cases diagnosed each year. It is necessary to discover the associations between circRNAs and lung cancer. The 4 inferred circRNAs have been validated

TABLE XI
TOP TEN CANDIDATE CIRCRNAS FOR LUNG CANCER

Rank	Name of circRNAs	References
1	hsa_circ_0000284	[56]
2	hsa_circ_0001313	Unknown
3	CDR1-AS	[57]
4	hsa_circ_0001649	Unknown
5	hsa_circ_0001445	Unknown
6	hsa_circ_0067934	[58]
7	hsa_circ_0061265	Unknown
8	hsa_circ_0003221	[59]
9	hsa_circ_0041150	Unknown
10	hsa_circ_0072088	Unknown

as shown in Table XI (hsa_circ_0000284, 1st, CDR1-AS, 3rd, hsa_circ_0067934, 6th, hsa_circ_0003221, 8th). CircRNA hsa_circ_0000284 (circHIPK3) functions as an endogenous miR-338-3p sponge by regulating FMT and inhibits miR-338-3p activity, which is related to non-small cell lung cancer by targeting IRS3 [56]. The expression level of CDR1-AS is robustly increased with the progression of non-small-cell lung cancer by down-regulating miR-7 [57]. Over-expression of hsa_circ_0067934 promotes proliferation and tumorigenesis of lung adenocarcinoma in related tissues [58]. In non-small-cell lung cancer, circRNA hsa_circ_0003221 (circPTK2) inhibits TGF-β-induced epithelial-mesenchymal transition and metastasis [59].

Nevertheless, there is not enough experimentally validated associations to be used for training the model. What's worse, the lack of sufficient knowledge impedes the subsequent study and investigation of all predicted associations. However, DMFCDA can provide the accurate prediction from the experimental results. It means that deep matrix factorization grasp the complex structure of data and is powerful to infer associations.

IV. CONCLUSION

Acting as regulators of transcription, intermediates in RNA processing reactions, and miRNA sponges, circRNAs take in various biological processes. Dyregulation and mutation of circRNAs may lead to diseases. Identifying associations between circRNAs and diseases is helpful in disease diagnosis. However, there are few associations that have been validated. It is a challenge to infer potential associations. Traditional matrix factorization models force vectors of circRNAs and diseases to map to a common space, and approximate associations with the inner product. It is hard to learn latent representations and grasp complex structures of data. In this study, we propose a deep matrix factorization to recommend circRNAs for queried diseases. DMFCDA utilizes a projection layer to learn representations, and multi-layer neural networks to capture non-linear associations. So complex that the process of disease is related to multiple biomolecules. The deep matrix factorization methods have been applied in the related field. For miRNA-disease, Li *et al*. utilizes a neural inductive matrix completion method with a graph convolutional network to predict miRNA-disease associations [60]. For lncRNA-disease, Hu *et al*. proposes a method combing traditional matrix factorization and deep learning to predict lncRNA-disease associations [61]. All the methods improve accuracy by integrating miRNA-related and lncRNA-related source data. Related biological data to miRNA and lncRNA can assist in predicting biomolecular associations. In contrast to circRNAs, miRNA and lncRNA are linear non-coding RNA. There is limited biological information for circRNAs due to its short-term study. DMFCDA makes use of explicit and implicit feedback, and feeds the neural network with raw vectors from interaction profiles instead of random initialization. Compared with state-of-the-art methods, DMFCDA performs outstanding according to AUC value with both LOOCV and five-fold cross validation. Moreover, DMFCDA outperforms the-state-of-art methods in the percentage of correctly retrieved true associations in various top ranks. DMFCDA provides accurate recommendation in the case study.

However, some improvements should be made in the future. Firstly, more and more biological information are useful for predicting associations with the development of techniques. DMFCDA could integrate more biological information to learn more data-consistent representation. Second, linear feature vectors could offer beneficial feature representation. Combining linear and non-linear feature vectors may provide a way to boost the performance. Third, the attention mechanism may be incorporated to focus important stuffs and ignore or diminish others. We would boost the framework to learn feature representation and get accurate prediction.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China under Grant No. 61732009, No.61972423, and No.61972185. The code can be found at https://github.com/bioinfomaticsCSU/DMFCDA.

REFERENCES

[1] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer *et al.*, "Circular rnas are a large class of animal rnas with regulatory potency," *Nature*, vol. 495, no. 7441, p. 333, 2013.

[2] B. Chen and S. Huang, "Circular rna: An emerging non-coding rna as a regulator and biomarker in cancer," *Cancer letters*, vol. 418, pp. 41–50, 2018.

[3] H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt, "Viroids are single-stranded covalently closed circular rna molecules existing as highly base-paired rod-like structures," *Proceedings of the National Academy of Sciences*, vol. 73, no. 11, pp. 3852–3856, 1976.

[4] C. Cocquerelle, B. Mascrez, D. Hetuin, and B. Bailleul, "Mis-splicing yields circular rna molecules." *The FASEB Journal*, vol. 7, no. 1, pp. 155–160, 1993.

[5] W. R. Jeck and N. E. Sharpless, "Detecting and characterizing circular rnas," *Nature biotechnology*, vol. 32, no. 5, p. 453, 2014.

[6] F. Wang, A. J. Nazarali, and S. Ji, "Circular rnas as potential biomarkers for cancer diagnosis and therapy," *American journal of cancer research*, vol. 6, no. 6, p. 1167, 2016.

[7] Q. Vicens and E. Westhof, "Biogenesis of circular rnas," *Cell*, vol. 159, no. 1, pp. 13–14, 2014.

[8] E. Lasda and R. Parker, "Circular rnas: diversity of form and function," *Rna*, vol. 20, no. 12, pp. 1829–1842, 2014.

[9] C. E. Burd, W. R. Jeck, Y. Liu, H. K. Sanoff, Z. Wang, and N. E. Sharpless, "Expression of linear and novel circular forms of an ink4/arf-associated non-coding rna correlates with atherosclerosis risk," *PLoS genetics*, vol. 6, no. 12, p. e1001233, 2010.

[10] W. Lukiw, "Circular rna (circrna) in alzheimer's disease (ad)," *Frontiers in genetics*, vol. 4, p. 307, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2999638, IEEE Journal of Biomedical and Health Informatics

[11] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, and J. Wang, "Applications of deep learning to mri images: A survey," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 1–18, 2018.

[12] J. He, Q. Xie, H. Xu, J. Li, and Y. Li, "Circular rnas and cancer," *Cancer Letters*, vol. 396, pp. 138–144, 2017.

[13] L. Peng, G. Chen, Z. Zhu, Z. Shen, C. Du, R. Zang, Y. Su, H. Xie, H. Li, X. Xu *et al.*, "Circular rna znf609 functions as a competitive endogenous rna to regulate akt3 expression by sponging mir-150-5p in hirschsprung's disease," *Oncotarget*, vol. 8, no. 1, p. 808, 2017.

[14] P. Li, H. Chen, S. Chen, X. Mo, T. Li, B. Xiao, R. Yu, and J. Guo, "Circular rna 0000096 affects cell growth and migration in gastric cancer," *British journal of cancer*, vol. 116, no. 5, p. 626, 2017.

[15] C. Fan, X. Lei, and F.-X. Wu, "Prediction of circrna-disease associations using katz model based on heterogeneous networks," *International journal of biological sciences*, vol. 14, no. 14, p. 1950, 2018.

[16] X. Lei, Z. Fang, L. Chen, and F.-X. Wu, "Pwcda: path weighted method for predicting circrna-disease associations," *International journal of molecular sciences*, vol. 19, no. 11, p. 3410, 2018.

[17] C. Yan, J. Wang, and F.-X. Wu, "Dwnn-rls: regularized least squares method for predicting circrna-disease associations," *BMC bioinformatics*, vol. 19, no. 19, p. 520, 2018.

[18] Q. Xiao, J. Luo, C. Liang, G. Li, J. Cai, P. Ding, and Y. Liu, "Identifying lncrna and mrna co-expression modules from matched expression data in ovarian cancer," *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[19] Z. Pan, H. Zhang, C. Liang, G. Li, Q. Xiao, P. Ding, and J. Luo, "Self-weighted multi-kernel multi-label learning for potential mirna-disease association prediction," *Molecular Therapy-Nucleic Acids*, vol. 17, pp. 414–423, 2019.

[20] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microrna-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, 2018.

[21] M. Žitnik, V. Janjić, C. Larminie, B. Zupan, and N. Pržulj, "Discovering disease-disease associations by fusing systems-level molecular data," *Scientific reports*, vol. 3, no. 1, pp. 1–9, 2013.

[22] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncrna–disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, 2018.

[23] M. Zitnik and B. Zupan, "Jumping across biomedical contexts using compressive data fusion," *Bioinformatics*, vol. 32, no. 12, pp. i90–i100, 2016.

[24] H. Wei and B. Liu, "icircda-mf: identification of circrna-disease associations based on matrix factorization," *Briefings in bioinformatics*, 2019.

[25] Q. Xiao, J. Luo, and J. Dai, "Computational prediction of human disease-associated circrnas based on manifold regularization learning framework," *IEEE journal of biomedical and health informatics*, 2019.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein-protein interaction site prediction through combining local and global features with deep neural networkseriginal paper," *Bioinformatics*, p. DOI: 10.1093/bioinformatics/btz699, 2019.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[31] C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li, and J. Wang, "Prediction of lncrna–disease associations based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 19, pp. 3357–3364, 2018.

[32] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems." in *IJCAI*, 2017, pp. 3203–3209.

[33] C. Fan, X. Lei, Z. Fang, Q. Jiang, and F.-X. Wu, "Circr2disease: a manually curated database for experimentally supported circular rnas associated with various diseases," *Database*, vol. 2018, 2018.

[34] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, "Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases," *Nucleic acids research*, vol. 47, no. D1, pp. D1034–D1037, 2018.

[35] P. Glažar, P. Papavasileiou, and N. Rajewsky, "circbase: a database for circular rnas," *Rna*, vol. 20, no. 11, pp. 1666–1670, 2014.

[36] L.-L. Zheng, J.-H. Li, J. Wu, W.-J. Sun, S. Liu, Z.-L. Wang, H. Zhou, J.-H. Yang, and L.-H. Qu, "deepbase v2. 0: identification, expression, evolution and function of small rnas, lncrnas and circular rnas from deep-sequencing data," *Nucleic acids research*, vol. 44, no. D1, pp. D196–D202, 2015.

[37] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[38] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[39] R. Jenatton, N. L. Roux, A. Bordes, and G. R. Obozinski, "A latent factor model for highly multi-relational data," in *Advances in Neural Information Processing Systems*, 2012, pp. 3167–3175.

[40] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 2008, pp. 263–272.

[41] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 2017, pp. 173–182.

[42] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 2333–2338.

[43] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[44] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, and M. J. Thun, "Cancer statistics, 2006," *CA: a cancer journal for clinicians*, vol. 56, no. 2, pp. 106–130, 2006.

[45] W. Ji, C. Qiu, M. Wang, N. Mao, S. Wu, and Y. Dai, "Hsa_circ_0001649: A circular rna and potential novel biomarker for colorectal cancer," *Biochemical and biophysical research communications*, vol. 497, no. 1, pp. 122–126, 2018.

[46] W. Weng, Q. Wei, S. Toden, K. Yoshida, T. Nagasaka, T. Fujiwara, S. Cai, H. Qin, Y. Ma, and A. Goel, "Circular rna cirs-7a promising prognostic biomarker and a potential therapeutic target in colorectal cancer," *Clinical Cancer Research*, vol. 23, no. 14, pp. 3918–3928, 2017.

[47] R. Mercy, "Circad: a manually curated database of circular rnas associated with diseases," *In Communication*, 2017.

[48] J. Zhang, X. Zhao, J. Zhang, X. Zheng, and F. Li, "Circular rna hsa_circ_0023404 exerts an oncogenic role in cervical cancer through regulating mir-136/tfcp2/yap pathway," *Biochemical and biophysical research communications*, vol. 501, no. 2, pp. 428–433, 2018.

[49] Z. Wang, M. Su, B. Xiang, K. Zhao, and B. Qin, "Circular rna pvt1 promotes metastasis via mir-145 sponging in crc," *Biochemical and biophysical research communications*, vol. 512, no. 4, pp. 716–722, 2019.

[50] Y. Jin, Y. Ren, Y. Gao, L. Zhang, and Z. Ding, "Hsa_circ_0005075 predicts a poor prognosis and acts as an oncogene in colorectal cancer via activating wnt/β-catenin pathway." *European review for medical and pharmacological sciences*, vol. 23, no. 8, pp. 3311–3319, 2019.

[51] L. Fu, Z. Jiang, T. Li, Y. Hu, and J. Guo, "Circular rna s in hepatocellular carcinoma: Functions and implications," *Cancer medicine*, vol. 7, no. 7, pp. 3101–3109, 2018.

[52] M. Huang, Y.-R. He, L.-C. Liang, Q. Huang, and Z.-Q. Zhu, "Circular rna hsa_circ_0000745 may serve as a diagnostic marker for gastric cancer," *World journal of gastroenterology*, vol. 23, no. 34, p. 6330, 2017.

[53] X. Li and M. Shen, "Circular rna hsa_circ_103809 suppresses hepatocellular carcinoma proliferation and invasion by sponging mir-620." *European review for medical and pharmacological sciences*, vol. 23, no. 2, pp. 555–566, 2019.

[54] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. Iny Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, and D. Lancet, "Malacards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search," *Nucleic acids research*, vol. 45, no. D1, pp. D877–D887, 2016.

[55] J. Lin, Y. He, L. Chen, X. Chen, S. Zang, and W. Lin, "Mylk promotes hepatocellular carcinoma progression through regulating cytoskeleton to enhance epithelial–mesenchymal transition," *Clinical and experimental medicine*, vol. 18, no. 4, pp. 523–533, 2018.

[56] J.-x. Zhang, J. Lu, H. Xie, D.-p. Wang, H.-e. Ni, Y. Zhu, L.-h. Ren, X.-x. Meng, and R.-l. Wang, "circhipk3 regulates lung fibroblast-to-myofibroblast transition by functioning as a competing endogenous rna," *Cell death & disease*, vol. 10, no. 3, p. 182, 2019.

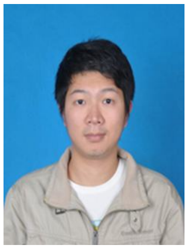
[57] X. Zhang, D. Yang, and Y. Wei, "Overexpressed cdr1as functions as an oncogene to promote the tumor progression via mir-7 in non-small-cell lung cancer," *OncoTargets and therapy*, vol. 11, p. 3979, 2018.

[58] M. Qiu, W. Xia, R. Chen, S. Wang, Y. Xu, Z. Ma, W. Xu, E. Zhang, J. Wang, T. Fang *et al.*, "The circular rna circprkci promotes tumor growth in lung adenocarcinoma," *Cancer research*, vol. 78, no. 11, pp. 2839–2851, 2018.

[59] L. Wang, X. Tong, Z. Zhou, S. Wang, Z. Lei, T. Zhang, Z. Liu, Y. Zeng, C. Li, J. Zhao *et al.*, "Circular rna hsa_circ_0008305 (circptk2) inhibits tgf-β-induced epithelial-mesenchymal transition and metastasis by controlling tif1γ in non-small cell lung cancer," *Molecular cancer*, vol. 17, no. 1, p. 140, 2018.

[60] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction," *Bioinformatics*, 2020.

[61] J. Hu, "Deep learning enables accurate prediction of interplay between lncrna and disease," *Frontiers in genetics*, vol. 10, p. 937, 2019.



Chengqian Lu received the B.S. degree from Xiangtan University in 2009, and the M.S. degree from Yunnan University in 2011. He received the Ph.D. degree from the School of Computer Science and Engineering, Central South University, China, in 2019. His current research interests include bioinformatics, machine learning and deep learning.

Min Zeng received the B.S. degree from Lanzhou University in 2013, and the M.S. degree from Central South University in 2016. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His research interests include bioinformatics, machine learning and deep learning.

Fuhao Zhang received his BSc degrees in Chongqing University of Posts and Telecommunications, China in 2014. He is currently a PhD student in Bioinformatics at Central South University. His current research interests include bioinformatics, network representation learning, and deep learning.

Fang-Xiang Wu (M'06-SM'11) received the B.Sc. degree and the M.Sc. degree in applied mathematics, both from Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first Ph.D. degree in control theory and its applica-tions from Northwestern Polytechnical University, Xian, China, in 1998, and the second Ph.D. degree in biomedical engineering from University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a Postdoctoral Fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a Professor of the Division of Biomedical Engineering and the Department of Mechanical Engineering at the U of S. His current research interests include Artificial Intelligence, Machine/Deep Learning, Computational Biology and Bioinformatics, Medical Image Analytics and Complex Network Analytics. Dr. Wu is serving as the editorial board member of five international journals, the guest editor of several international journals, and as the program committee chair or member of several inter-national conferences.

Min Li received the Ph.D. degree from the School of Computer Science and Engineering, Central South University, China, in 2008. She is currently a Professor at the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her research interests include computational biology, systems biology and bioinformatics. She has published more than 80 technical papers in refereed journals such as Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Proteomics, and conference proceedings such as BIBM, GIW and ISBRA. According to Google scholar, her paper citations is more than 3000 and H-index is 28.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the Ph.D. degree in computer science from Central South University, China, in 2001. He is the dean and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include bioinformatics and computer algorithm. He published over 200 papers in the field of computational biology and bioinformatics in journals including Bioinformatics, TCBB and conferences including ISMB and BIBM. His research has been cited over 6700 times (Google Scholar) with H-index=36. He is a senior member of the IEEE, the Chair of ACM Sigbio China and a senior member of China Computer Federation. He serves as editor of TCBB, IJBRA, IJDMB, NMSHIB, IJDSN, CPPS, PPL, CBIO and guest editor of BMC Bioinformatics, TCBB. He served as program chair, co-chair or committee in several international conferences including FAW, ISBRA and BIBM.