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.2020.2987034, IEEE Journal of Biomedical and Health Informatics

miRTMC: A miRNA target prediction method based on matrix completion algorithm

Hui Jiang[†], Mengyun Yang[†], Xiang Chen, Min Li, Yaohang Li* and Jianxin Wang*

***Abstract*—microRNAs (miRNAs) are small non-coding RNAs which modulate the stability of gene targets and their rates of translation into proteins at transcriptional level and post-transcriptional level. miRNA dysfunctions can lead to human diseases because of dysregulation of their targets. Correct miRNA target prediction will lead to better understanding of the mechanisms of human diseases and provide hints on curing them. In recent years, computational miRNA target prediction methods have been proposed according to the interaction rules between miRNAs and targets. However, these methods suffer from high false positive rates due to the complicated relationship between miRNAs and their targets. The rapidly growing number of experimentally validated miRNA targets enables predicting miRNA targets with high precision via accurate data analysis. Taking advantage of these known miRNA targets, a novel recommendation system model (miRTMC) for miRNA target prediction is established using a new matrix completion algorithm. In miRTMC, a heterogeneous network is constructed by integrating the miRNA similarity network, the gene similarity network, and the miRNA-gene interaction network. Our assumption is that the latent factors determining whether a gene is the target of miRNA or not are highly correlated, i.e., the adjacency matrix of the heterogeneous network is low-rank, which is then completed by using a nuclear norm regularized linear least squares model under non-negative constraints. Alternating direction method of multipliers (ADMM) is adopted to numerically solve the matrix completion problem. Our results show that miRTMC outperforms the competing methods in terms of various evaluation metrics. Our software package is available at https://github.com/hjiangcsu/miRTMC.**

***Index Terms*—matrix completion, miRNA target prediction, recommendation algorithm.**

I. Introduction

MICRORNAS (miRNAs) are short and important non-coding RNAs that regulate gene translation or mediate gene degradation in animals and plants [1]. miRNA is a sequence of approximately 22 nucleotides (A, U, G, C). The target gene is also a sequence composed of (A, U, G, C). miRNA interacts with its target genes in a sequence manner through the observations of biological experiment samples [2]. miRNA targets are usually defined as the complementary sites within target genes pairing with the miRNA seed region. Due to the fact that one miRNA may target multiple genes in order to regulate gene expression, a complex many-to-many relationship exists between miRNAs and genes. Recent studies on miRNAs have shown that they play important roles in many biological processes, such as cell growth and differentiation [3], development [4], apoptosis [5], and others. The dysfunctions of miRNAs can dysregulate their targets and thus lead to many diseases including cancers [6]. Consequently, predicting miRNA targets correctly will bring insight into complex disease mechanisms.

Many biological experimental techniques, such as qRT-PCR, luciferase reporter assays, western blot, differential expression, and cross-linking immunoprecipitation (CLIP), have been developed to identify miRNA targets [7]. Biological experiment is the most reliable way to identify miRNA targets, but it is usually not only expensive and but also time-consuming. In recent years, dozens of computational methods for miRNA target prediction have been proposed [8]. These miRNA target prediction methods derive binding rules observed from biological experiments. These rules determine miRNA targets according to sequence complementarity scores, free energy of miRNA-target duplex, cross-species evolutionary conservation scores, site accessibility, and others [9]. Most of these methods firstly seek the potential binding sites on target mRNA (3'-UTR sequence) with the miRNA 5' end sequence by using one or two of the rules aforementioned. Then, scores measuring the binding affinity, such as minimal free energy, sequence match score, and so on, are calculated. Finally, filtering these scores against the pre-specific thresholds is used to determine whether the mRNA is the target of the miRNA or not. For example, miRanda [1] predicts miRNA target by using the features of seed pairing, free energy, and cross-species evolutionary conservation. miRanda algorithm calculates the match score of potential binding sites of each miRNA-mRNA pair by dynamic programming and then computes the minimal free energy. Finally, the results are processed by cross-species conservation analysis. After seeking the potential binding sites on the mRNA with miRNA, TargetScan [10] calculates the folding free energy of these binding sites, and then the Z-score of each

H. Jiang is with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China, and School of Computer, University of South China, Hengyang, 421001, China. E-mail: jianghui@csu.edu.cn.

M. Yang is with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China, and Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang, Hunan, 422000, China E-mail: mengyunyang@csu.edu.cn.

X. Chen is with is with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China. E-mail: chenxofhit@gmaill.com.

M. Li is with with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China. E-mail: limin@mail.csu.edu.cn.

Y. Li is with Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA. E-mail: yaohang@cs.odu.edu.

J. Wang is with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China, Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, 410083, China. E-mail: jxwang@mail.csu.edu.cn.

* Corresponding author. [†] These authors contributed equally to this work.

Manuscript received XXX XX, XXXX; revised XXX XX, XXXX.

miRNA-mRNA pair, where the miRNA-mRNA pairs scoring over a certain threshold is considered as true. TargetScanS [11] is an extension of TargetScan by introducing conservation features to replace the free energy feature to predict potential miRNA targets. RNAhybrid [12] predicts miRNA targets by calculating the minimal free energy of miRNA-mRNA pairs, assuming that miRNA binds to its target genes with an energy optimized pathway. PITA [13] finds the potential miRNA target sites and then builds a parameter-free model for calculating the thermodynamic scores of potential miRNA targets. MovingTarget [14] uses biological constraints, such as the number of binding sites and the sequence complementarity, to predict miRNA targets. Although these approaches can lead to miRNA target predictions with certain accuracy, they often suffer from relatively high false positive rates.

Machine learning methods, such as support vector machine (SVM) [15][16], ensemble learning [17], deep learning [18]–[20], and many others, have been incorporated into the sequence-based miRNA target prediction approaches to reduce false positive rates. Most of these are supervised learning methods, which are based on labeled training sets. However, the positive samples are usually much easier to obtain than the negative ones, which makes the method of collecting negative samples important. SVMicrO [16] selects positive samples from miRecords [21] and obtains the high quality negative samples from 20 miRNA over-expressed microarray data. A three-stage method based on seed match rules and SVM is proposed to predict miRNA targets. MiRTDL [18] predicts miRNA targets with a convolutional neural network (CNN) using 20 features extracted from each sample. The positive and negative samples used in MiRTDL are obtained from TarBase [22]. deepTarget [19] uses two autoencoders to extract the features of the miRNAs and 3'UTR sequences of mRNAs, respectively. Then, an RNN (recurrent neural networks) is trained to recognize the miRNA targets with negative samples generated by the Fisher-Yates shuffle algorithm [23]. DeepMirTar [20] is a miRNA target site level prediction method based on stacked denoising autoencoders. After selecting miRNA-target binding sites in mirMark [24] data and CLASH [25] data by using miRanda [1], a stacked denoising auto-encoder is trained to predict miRNA target binding sites in DeepMirTar method. In summary, the success of these supervised learning methods relies on extracting the effective sequence features that are capable of differentiating the positive and negative miRNA-gene association samples. However, the power of these supervised learning methods is limited when there is short of reliable negative miRNA-gene association samples in practice. Unlike supervised learning methods, miRTRS [26] does not require negative samples. miRTRS predicts miRNA targets based on a recommendation algorithm which focuses on network-based inference. miRTRS uses experimentally validated miRNA targets to construct a miRNA-gene interaction network and then the score of each miRNA-gene pair is calculated by a network-based inference method.

At the same time, matrix completion algorithms have been successfully applied to predicting lncRNA-disease associations [27][28], drug-disease associations [29]–[33], and miRNA-disease associations [34]–[41] . DLRMC [34] uses a matrix completion method with a dual Laplacian regularization term on miRNA functional similarity and disease sematic similarity to predict miRNA-disease associations. IMCMDA [35] calculates miRNA similarity and disease similarity based on miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel, and then applies an inductive matrix completion method to predict miRNA-disease associations. MCLPMDA [36] uses a low-rank matrix completion method to fill out the disease semantic similarity matrix and miRNA functional similarity matrix, incorporates them with two other similarity matrices, and finally applies a label propagation method to predict miRNA-disease associations. NCMCMDA [37] integrates neighborhood constraints with a matrix completion model to predict miRNA-disease associations and adopts a fast iterative shrinkage-thresholding algorithm to solve the corresponding optimization problem. Most of the above matrix completion algorithms for predicting miRNA-related associations focus on filling out the incomplete association matrix using different features. The novelty of our method is summarized as follows: 1) miRTMC builds a heterogeneous network between miRNAs and genes by integrating miRNA similarity and gene similarity networks and miRNA-gene association networks and then completes the adjacency matrix of the heterogeneous network. The computations of miRNA similarities and gene similarities are often inaccurate, which bring noise to the matrix completion model. The advantage of completing the adjacency matrix of the heterogeneous network is that the miRNA similarities and gene similarities are flexible and adjustable during optimization, which allows the miRTMC method to tolerate the potential noise in the miRNA and gene similarity matrices. 2) Most of the existing matrix completion methods for miRNA related association prediction are base on Singular Value Thresholding (SVT) algorithm [42]. miRTMC adopts a noise model instead to further address the noise issue and adds a non-negative constraint to make the predicted values biologically explainable. 3) miRTMC adopt a fast SVD approximation method named R^4SVD [43] to fast compute the dominating singular values and their corresponding singular vectors. This enables miRTMC to efficiently handle large adjacency matrix from the miRNA-gene heterogeneous network.

The miRNA target prediction methods aforementioned have some limitations, including high false positive rate and the requirement of high-quality negative samples. As more and more miRNA targets are identified by biological experiments, more and more articles about miRNA targets are published. The miRNA targets in these articles are extracted by text mining-based methods such as miRTarBase [44], miRWalk [45], and TarMiner [46]. Several experimentally validated miRNA target databases are now available online, such as miRTarBase and TarBase [47]. The number of experimentally validated miRNA targets is increasing significantly every year. These large number of new miRNA targets enable particularly effective recommendation models to predict miRNA targets.

In this study, we propose a miRNA target prediction method, named miRTMC, by using a matrix completion method which has demonstrated success in collaborative filtering recommendation system applications. More specifically, a heterogeneous network is firstly constructed according to the experimentally validated miRNA targets, miRNA seed sequence similarity, and gene sequence similarity. Assuming that there is limited number of independent factors governing miRNA-gene interactions, the adjacency matrix of the heterogeneous network is of low rank. Therefore, this matrix can be completed by solving a nuclear norm regularized least squares model with non-negative constraints. Alternating direction method of multipliers (ADMM) [48] is adopted to numerically solve this problem. In order to computationally efficiently deal with the large matrices generated from large number of experimentally validated miRNA targets, a recycling rank-revealing randomized singular value decomposition algorithm (R^{4}SVD) [43] is used for fast and adaptively approximating the dominant singular values and their corresponding singular vectors. Our experiment results show that miRTMC outperforms four competing miRNA target prediction methods and one state-of-the-art matrix completion methods in terms of area under receiver operating characteristic (ROC) curve (AUC) and precision. The matrix completion model is able to capture the global pattern of the miRNA-gene association to reduce the false positive rate. Moreover, the construction of heterogeneous network takes advantage of the relationships among miRNAs as well as those among genes, where negative samples are not absolutely necessary. Our web service is available at http://bioinformatics.csu.edu.cn/miRTMC.

II. Methods

In this study, we propose a miRNA-target prediction method based on a matrix completion algorithm. First of all, a heterogeneous network is constructed by integrating the miRNA similarity network, the gene similarity network, and the miRNA-gene interaction network. Then, the association matrix of the heterogeneous network is filled out by a matrix completion algorithm. Finally, we get a recovered matrix which contains the recommendation score of each miRNA-gene pair.

A. Construction of the heterogeneous network

The prediction models based on matrix completion predict miRNA-target associations with the way of filling out the unknown elements in the association matrix. Most of these methods have difficulty in dealing with the cold start problems for novel miRNAs or genes. Previous studies have shown that integrating different kinds of miRNA and gene features can not only address the cold start problem, but also improve the accuracy of association prediction. Accordingly, we construct a heterogeneous network that integrates miRNA sequence similarity data, gene sequence similarity data, and miRNA-gene interaction data, and complete the adjacency matrix of this network to predict miRNA targets.

1) Construction of miRNA-gene interaction network: Let $M = \{m_1, m_2, \cdots, m_p\}$ be a set of miRNAs and $T = \{t_1, t_2, \cdots, t_q\}$ be a set of genes. The miRNA-gene interaction network can be presented as a bipartite graph *G(M,T,E)*, where the set of edges $E = \{(m_i, t_j) | t_j \in T \ and \ m_i \in M\}$ represents the known miRNA-target interactions. Let $A_{TM} = \{a_{ij}\}_{q\times p}$ be the adjacency matrix of the miRNA-gene interaction network, where a_{ij}=*1* if there exist biological experiment evidences showing that gene t_i is the target gene of miRNA m_j and a_{ij}=*0* for unknown indications. The experimentally validated miRNA-target associations are extracted from the available databases, such as miRTarBase and others. Consider a few miRNA targets in miRTarBase database as examples. HIF1A, the miRNA target of hsa-miR-20a-5p, has been verified by five biological experiments including Luciferase reporter assay; CXCR4, the miRNA target of hsa-miR-146a-5p, has been validated by four biological experiments including qRT-PCR. As a result, in miRNA-gene network construction, there is one edge between hsa-miR-20a-5p and its miRNA target HIF1A and one between hsa-miR-146a-5p and CXCR4.

2) Construction of miRNA similarity network: According to the rules of miRNA interacting with its targets, the seed region of miRNA and the 3′ UTR of mRNA play an important role in miRNA target prediction. Hence, in this method, the seed region similarities between each pair of miRNAs are calculated, and the miRNA similarity network is constructed. The Needleman-Wunsch algorithm [49] is used to calculate the miRNA sequence similarity score by global alignment. Let $SM \in \mathcal{R}^{p\times p}$ be the matrix of miRNA sequence similarity score and then the normalized seed region similarity score $M_s(i,j)$ between m_i and m_j becomes

$$M_s(i,j) = \frac{SM(i,j)}{Max(SM)}, \tag{1}$$

where $M_s(i,j)$ is the normalized seed region similarity score between m_i and m_j, $SM(i,j)$ is the seed region similarity score between m_i and m_j from the Needleman-Wunsh global alignment algorithm, and $Max(SM)$ is the maximum sequence similarity score among all pairs of miRNAs in SM.

3) Construction of gene similarity network: Since the 3′UTR sequences are relatively long, calculating the similarity of large number of mRNA pairs is time consuming. In addition, miRNAs usually bind to a small area of 3′UTR. Therefore, the sequence similarity of 3′UTR region of paired mRNAs is calculated by Smith-Waterman local alignment algorithm [50]. Let $ST \in \mathcal{R}^{q\times q}$ be the matrix of gene sequence similarity score. The normalized similarity score $T_s(i,j)$ between t_i and t_j is calculated as

$$T_s(i,j) = \frac{ST(i,j)}{Max(ST)}, \tag{2}$$

where $T_s(i,j)$ is the normalized gene sequence similarity score between t_i and t_j, $ST(i,j)$ is the gene sequence similarity score between t_i and t_j from the Smith-Waterman local alignment algorithm, and $Max(ST)$ is the maximum sequence similarity score among all pairs of genes in ST.

Finally, we connect the miRNA similarity network and the gene similarity network using a biologically experimentally validated miRNA-gene interaction network, which results in a heterogeneous network. Figure 1 is a hypothetical example

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.2020.2987034, IEEE Journal of Biomedical and Health Informatics

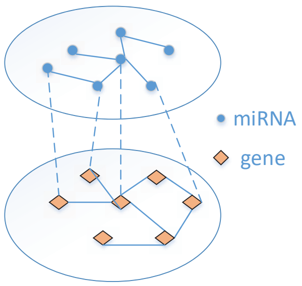

Fig. 1. The diagram of miRNA-gene heterogeneous network.

of this heterogeneous network. The association matrix of the heterogeneous network is defined as:

$$H = \begin{bmatrix} T_s & A_{TM} \\ A_{TM}^T & M_s \end{bmatrix}, \tag{3}$$

where A_{TM}^T represents the transpose of A_{TM}. Because the three subnetworks in the heterogeneous network are undirected networks, the adjacency matrix H is symmetric. Due to the fact that the similarities between miRNAs, the similarities between genes, and the elements in miRNA-gene association matrix are non-negative, the adjacency matrix H is a semi-positive definite matrix, where the eigenvalues of H are real, positive and are equal to the singular values. The zeros located in the submatrices A_{TM} and A_{TM}^T represent the unknown interaction which we attempt to predict. As a result, the aim of the miRNA targets prediction problem is considered as filling out the unknown elements in matrix H.

B. Prediction using low-rank matrix completion

Based on the assumption that similar miRNAs tend to regulate similar genes, the potential factors that dominate the association between miRNAs and their targets are highly related, which leads to correlation in the corresponding data matrix. The miRNA target prediction algorithm which we propose in this paper, is based on generating an r-rank matrix H_r to approximate the *(p+q)×(p+q)* adjacency matrix H of the miRNA-gene heterogeneous network aforementioned, and the value of r is far less than the dimensionality of H, under the low-rank assumption. Let Ω be the set of indices of all known elements in H, which includes miRNA similarities in matrix M_s, gene similarities in matrix T_s, and experimentally validated miRNA targets in matrix A_{TM} and the transpose of A_{TM}. Ideally, the low-rank matrix X is obtained by minimizing the rank model as follows:

$$\begin{aligned} &\min \ \ rank(X) \\ &s.t. \ \ P_\Omega(X) = P_\Omega(H), \end{aligned} \tag{4}$$

where function $rank(\cdot)$ measures the rank of the input matrix and $P_\Omega(\cdot)$ is a projection function defined as follows:

$$(P_\Omega(X))_{ij} = \left\{ \begin{array}{ll} X_{ij}, & if \ (i,j) \in \Omega \\ 0, & otherwise. \end{array} \right. \tag{5}$$

Unfortunately, the above rank minimization problem is a well-known NP-hard problem [51]. This problem can be relaxed as minimizing the nuclear norm of matrix X such that

$$\begin{aligned} &\min \|X\|_* \\ &s.t. \quad P_\Omega(X) \ = P_\Omega(H), \end{aligned} \tag{6}$$

where $\|X\|_*$ denotes the nuclear norm of X. In order to make model (6) more suitable for handling noisy biological data, we adopt a noise model such that

$$\begin{aligned} &\min_X \lambda\|X\|_* + \tfrac{1}{2}\left\|P_\Omega(X) - P_\Omega(H)\right\|_F^2 \\ &s.t. \, X \geq 0, \end{aligned} \tag{7}$$

where a non-negative constraint is incorporated to ensure the predicted values are positive or 0 and λ is the harmonic parameter. Here we adopt an ADMM-based method to solve the nuclear norm regularized linear least squares model with nonnegative constraint. We introduce an auxiliary matrix variable Y and this problem becomes

$$\begin{aligned} &\min_{X,Y} \lambda\|X\|_* + \tfrac{1}{2}\left\|P_\Omega(Y) - P_\Omega(H)\right\|_F^2 \\ &s.t. \quad \begin{array}{l} Y = X \\ Y \geq 0. \end{array} \end{aligned} \tag{8}$$

The augmented Lagrangian of model (8) is

$$\begin{aligned} \mathcal{L}(X,Y,\Lambda) = \lambda\|X\|_* + \tfrac{1}{2}\left\|P_\Omega(Y) - P_\Omega(H)\right\|_F^2 \\ + < \Lambda, Y - X > + \tfrac{\alpha}{2}\left\|Y - X\right\|_F^2, \end{aligned} \tag{9}$$

where $\Lambda \in R^{(p+q)\times(p+q)}$ is a Lagrangian multiplier and α >0 is a penalty parameter. The iterative scheme of ADMM for model (8) becomes

$$Y_{k+1} = \arg\min_{Y\geq 0} \mathcal{L}(X_k, Y, \Lambda_k) \tag{10}$$

$$X_{k+1} = \arg\min_X \mathcal{L}(X, Y_{k+1}, \Lambda_k) \tag{11}$$

$$\Lambda_{k+1} = \Lambda_k + \gamma\alpha(Y_{k+1} - X_{k+1}), \tag{12}$$

where γ is the learning rate.

In order to obtain Y_{k+1}, we need to solve

$$\begin{aligned} &\min_Y \tfrac{1}{2}\left\|P_\Omega(Y) - P_\Omega(H)\right\|_F^2 + \tfrac{\alpha}{2}\left\|Y - (X_k - \tfrac{1}{\alpha}\Lambda_k)\right\|_F^2 \\ &s.t. Y \geq 0, \end{aligned} \tag{13}$$

which can be decomposed into two subproblems:

$$\begin{aligned} &\min_Y \tfrac{1}{2}\left\|P_\Omega(Y) - P_\Omega(H)\right\|_F^2 + \tfrac{\alpha}{2}\left\|P_\Omega(Y) - P_\Omega(X_k - \tfrac{1}{\alpha}\Lambda_k)\right\|_F^2 \\ &s.t. P_\Omega(Y) \geq 0 \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\min \left\|P_{\bar\Omega}(Y) - P_{\bar\Omega}(X_k - \tfrac{1}{\alpha}\Lambda_k)\right\|_F^2 \\ &s.t. P_{\bar\Omega}(Y) \geq 0, \end{aligned} \tag{15}$$

where $\bar\Omega$ is the complement of Ω. Then, the solution of (10) will be obtained by computing the above two parts, respectively:

$$(Y_{k+1})_\Omega = Q_+(\frac{1}{\alpha+1} P_\Omega(H + \alpha X_k - \Lambda_k)) \tag{16}$$

and

$$(Y_{k+1})_{\bar\Omega} = Q_+(P_{\bar\Omega}(X_k - \frac{1}{\alpha}\Lambda_k)), \tag{17}$$

where Q_+ is the projection onto the non-negative matrix subspace, i.e.

$$(Q_+(X))_{ij} = \begin{cases} X_{ij}, & if\ X_{ij} > 0 \\ 0, & otherwise. \end{cases} \tag{18}$$

X_{k+1} can be obtained by using linearized Bregman iteration [52] such that

$$\begin{aligned} X_{k+1} &= \arg\min \lambda\|X\|_* + \tfrac{\alpha}{2}\left\|X - (Y_{k+1} + \tfrac{1}{\alpha}\Lambda_k)\right\|_F^2 \\ &= D_{\frac{\lambda}{\alpha}}(Y_{k+1} + \tfrac{1}{\alpha}\Lambda_k), \end{aligned} \tag{19}$$

where $D_{\frac{\lambda}{\alpha}}(X)$ is the Singular Value Thresholding (SVT) operator [42][53] defined as follows:

$$D_{\frac{\lambda}{\alpha}}(X) = \sum_{j=1}^{\sigma_j \geq \frac{\lambda}{\alpha}} (\sigma_j - \frac{\lambda}{\alpha}) u_j v_j^T, \tag{20}$$

where σ_j is the jth singular value not less than $\frac{\lambda}{\alpha}$ and u_j and v_j are the corresponding left and right singular vectors, respectively.

In summary, the proposed ADMM for (8) generates (Y_{k+1}, X_{k+1}, Λ_{k+1}) by the following iterative framework:

$$\begin{aligned} &(Y_{k+1})_\Omega = Q_+(\tfrac{1}{\alpha+1} P_\Omega(H + \alpha X_k - \Lambda_k)) \\ &(Y_{k+1})_{\bar{\Omega}} = Q_+(P_{\bar{\Omega}}(X_k - \tfrac{1}{\alpha}\Lambda_k)) \\ &X_{k+1} = D_{\frac{\lambda}{\alpha}}(Y_{k+1} + \tfrac{1}{\alpha}\Lambda_k) \\ &\Lambda_{k+1} = \Lambda_k + \gamma\alpha(Y_{k+1} - X_{k+1}). \end{aligned} \tag{21}$$

Computing the singular values of X is required at each iteration step during the matrix completion based on ADMM algorithm. The direct way to compute the singular values of a matrix is to compute singular value decomposition (SVD). However, repeatedly computing the full SVD of a large-scale matrix, such as the adjacency matrix of a large miRNA-gene heterogeneous network, is time-consuming and memory-intensive. Actually, $D_{\frac{\lambda}{\alpha}}(X)$ only requires the singular values in X that are greater than $\frac{\lambda}{\alpha}$. Hence, we adopt a fast SVD algorithm that focuses on approximating the dominant singular values to reduce the computation costs. The underlying idea of fast SVD algorithm is the randomized SVD algorithms. The large matrix X is condensed into a small, dense matrix by projecting X onto a sampling matrix as an approximate basis while retaining the important information of X. After that, the top singular values/vectors of X are approximated by performing a deterministic SVD on the small, dense matrix with the relatively low computation cost and high confidence. Based on the basis idea aforementioned, a rank-revealing randomized singular value decomposition algorithm (R^{3}SVD) [54] is proposed for conducting partial SVD to fast approximate $D_{\frac{\lambda}{\alpha}}(X)$ adaptively. By projecting X onto a small Gaussian matrix and applying power iterations, R^{3}SVD fast approximates the SVT operator of X. A low-rank QB decomposition based on orthogonal Gaussian projection is built up incrementally in R^{3}SVD. Then the low-rank SVD is derived. A recycling rank-revealing randomized SVD(R^{4}SVD) [43] has been proposed subsequently by taking advantage of the singular vectors obtaining from the previous iterations. In this study, R^{4}SVD is integrated into miRTMC method for reducing the computational costs of the ADMM method, which is referred to as ADMM-R^{4}SVD in this paper.

C. Two-step method for miRNA target prediction

Selecting a suitable rank r for matrix completion is important to achieve good prediction performance while avoiding overfitting and reducing computational cost. Here, a two-step method is adopted in miRTMC. The first step is to determine the optimal rank by designating a validation set, which is constructed by randomly selecting 10% known miRNA targets from the miRNA-gene interaction matrix. Then the 10% known miRNA targets are predicted iteratively by using the ADMM-R^{4}SVD algorithm, and meanwhile we record the sum of the AUC value and the maximum value of precision with respect to all ranks (we define this sum as tradeoff value in the following text) in each iteration so that the AUC value and the maximum value of precision are balanced. Rank r corresponding to the maximum tradeoff value is the desired rank. At second step, the ADMM-R^{4}SVD algorithm runs on the whole known miRNA-target interactions iteratively until rank r is reached or the maximum number of iterations is reached. At last, the completed miRNA-gene association matrix is obtained from the matrix H^* and the recommendations are conducted by sorting the predicted scores of the miRNA-gene pairs.

Algorithm 1 depicts the two-step recommendation method for miRNA target prediction. Function ADMM-R^{4}SVD$(\cdot)$ uses R^{4}SVD to accelerate the ADMM algorithm for matrix completion. The AUC value and precision of validation set in the first step are calculated by function calc_AUC_pre$(\cdot)$ in our algorithm. Function calc_residual$(\cdot)$ calculates the residual between the original matrix and the completed matrix on the training data.

III. Materials

We download the human miRNA sequence from miRBase (release 21) [55] and extract the seed regions of these miRNA sequences. The $3'$ UTR sequences of mRNAs are extracted from NCBI RefSeq, GRCh38/hg38 database by using hgTables from the web site genome.ucsc.edu. To facilitate comparison with competing methods, we prepare two datasets, named D1 and D2, for the 10-fold cross validation experiments and five independent datasets, named IDS1, IDS2 and so on, for independent datasets experiments. D1 and D2 are constructed as follows:

1) D1: We download the biologically experimentally validated human miRNA targets from miRTarBase (version 6.1) from its website [56]. After removing the duplicated miRNA targets, it contains 322,160 human miRNA-target interactions between 2,618 human miRNAs and 14,814 human genes. After mapping the genes to NCBI gene ids and selecting the $3'$ UTR sequences from hg38 dataset and mapping the miRNA names to miRBase (release 21), we get 319,172 miRNA-target interactions between 2,588 miRNAs and 14,499 genes (36,065 mRNAs).

2) D2: We download the biologically experimentally validated human miRNA targets from miRTarBase (version 7) from the website [44]. After removing the duplicated miRNA targets, it contains 380,639 human miRNA-target interactions between 2,599 miRNAs and 15,064 human genes. After mapping the genes to NCBI gene ids and selecting the $3'$ UTR

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.2020.2987034, IEEE Journal of Biomedical and Health Informatics

Algorithm 1: matrix completion method for miRNA target prediction

```
Input: gene 3′ UTR sequence similarity matrix $\mathbf{T_s}$ and its indices set $\Omega_{\mathrm{TT}}$, miRNA seed sequence similarity matrix $\mathbf{M_s}$ and its indices set $\Omega_{\mathrm{MM}}$, miRNA-gene association matrix $\mathbf{A}_{\mathrm{TM}}$ and its indices set $\Omega_{\mathrm{TM}}$.
Output: Completed miRNA-gene association matrix $\mathbf{A}^*_{\mathrm{TM}}$.
Set the value of parameters: $\lambda$; $\gamma$ ; $\alpha$; $tol$;
/* Step I: find the optimal rank */
10% of indices of $\Omega_{\mathrm{TM}}$ are selected randomly as the verification set $\Omega^v_{\mathrm{TM}}$. Hence $\mathbf{A}_{\mathrm{TM}} = \mathbf{A}'_{\mathrm{TM}} + \mathbf{A}^v_{\mathrm{TM}}$, $\Omega_{\mathrm{TM}} = \Omega'_{\mathrm{TM}} \cup \Omega^v_{\mathrm{TM}}$;
$\mathbf{H}^* \leftarrow \begin{bmatrix} \mathbf{T_s} & \mathbf{A}'_{\mathrm{TM}} \\ \mathbf{A_{\mathrm{TM}}}'^T & \mathbf{M_s} \end{bmatrix}$;
$\Omega \leftarrow \Omega_{\mathrm{MM}} \cup \Omega_{\mathrm{TT}} \cup \Omega'_{\mathrm{TM}} \cup \Omega'_{\mathrm{MT}}$;
$best\_rank \leftarrow 0$; $r \leftarrow 0$; $max\_aucpre \leftarrow 0$;
while the maximum number of iterations is not reached do
    $\left(\begin{bmatrix} \mathbf{T}^*_{\mathbf{s}} & \mathbf{A}^*_{\mathrm{TM}} \\ \mathbf{A}^*_{\mathrm{TM}}{}^T & \mathbf{M}^*_{\mathbf{s}} \end{bmatrix}, r\right) \leftarrow$ ADMM-R$^4$SVD($\mathbf{H}^*, \Omega, \lambda, \gamma, \alpha$);
    /* calculate AUC and precision by using function calc_AUC_pre */
    $aucpre \leftarrow$ calc_AUC_pre($\mathbf{A}^*_{\mathrm{TM}}, \mathbf{A}^v_{\mathrm{TM}}, \Omega^v_{\mathrm{TM}}$);
    $\mathbf{H}^{*'} \leftarrow \begin{bmatrix} \mathbf{T}^*_{\mathbf{s}} & \mathbf{A}^*_{\mathrm{TM}} \\ \mathbf{A}^*_{\mathrm{TM}}{}^T & \mathbf{M}^*_{\mathbf{s}} \end{bmatrix}$
    if $aucpre > max\_aucpre$ then
        $max\_aucpre \leftarrow aucpre; best\_rank \leftarrow r$;
        /* optimal rank */
    end
    /* calculate the residual by using function calc_residual */
    $res \leftarrow calc\_residual(\mathbf{H}^*, \mathbf{H}^{*'}, \mathbf{\Omega}_{\mathrm{train}})$
    if $r \geq (m+n)$ or $res \leq tol$ then break;
    /* m,n represent the number of miRNA and gene respectively */
    $\mathbf{H}^* \leftarrow \begin{bmatrix} \mathbf{T}^*_{\mathbf{s}} & \mathbf{A}^*_{\mathrm{TM}} \\ \mathbf{A}^*_{\mathrm{TM}}{}^T & \mathbf{M}^*_{\mathbf{s}} \end{bmatrix}$;
end
/* Step II: predict miRNA target by matrix completion */
$\mathbf{H}^* \leftarrow \begin{bmatrix} \mathbf{T_s} & \mathbf{A_{TM}} \\ \mathbf{A^T_{TM}} & \mathbf{M_s} \end{bmatrix}$;
$\Omega \leftarrow \Omega_{\mathrm{MM}} \cup \Omega_{\mathrm{TT}} \cup \Omega_{\mathrm{TM}} \cup \Omega_{\mathrm{MT}}$;
while the maximum number of iterations is not reached do
    $\left(\begin{bmatrix} \mathbf{T}^*_{\mathbf{s}} & \mathbf{A}^*_{\mathrm{TM}} \\ \mathbf{A}^*_{\mathrm{TM}}{}^T & \mathbf{M}^*_{\mathbf{s}} \end{bmatrix}, r\right) \leftarrow$ ADMM-R$^4$SVD($\mathbf{H}^*, \Omega, \lambda, \gamma, \alpha$);
    if $r \geq best\_rank$ then break;
    $\mathbf{H}^* \leftarrow \begin{bmatrix} \mathbf{T}^*_{\mathbf{s}} & \mathbf{A}^*_{\mathrm{TM}} \\ \mathbf{A}^*_{\mathrm{TM}}{}^T & \mathbf{M}^*_{\mathbf{s}} \end{bmatrix}$;
end
return $\mathbf{A}^*_{\mathrm{TM}}$;
```

sequences from hg38 dataset and then mapping the miRNA names to miRBase (release 21), finally, we get 377,236 interactions between 2,588 miRNAs and 14,742 genes (36,565 mRNAs).

Table 1 summarizes the datasets for 10-fold cross validation experiments.

TABLE I
THE DESCRIPTION OF THE DATASETS FOR 10-FOLD CROSS VALIDATION EXPERIMENTS

	miRNAs	genes	mRNAs	Interactions	sparsity
D1	2,588	14,499	36,065	319,172	0.0085
D2	2,588	14,742	36,565	377,236	0.0099

Sparsity is the ratio between the number of known miRNA-target pairs and the number of all possible miRNA pairs

In order to further evaluate miRTMC and the competing methods, two independent datasets are constructed based on D1 and D2. We select 2,588 miRNAs and 14,499 genes from D1 and extract 374,566 miRNA-target pairs from D2. We name this dataset IDS1. Compared with D1, 54,955 new interactions are included in IDS1. We select 2,588 miRNAs from D1 and extract 377,236 miRNA target pairs consisting of 2,588 miRNAs from D2. We name this dataset IDS2. Compared with D1, 243 new genes and 58,064 new interactions are included in IDS2. Table 2 shows the datasets for independent datasets experiments. All miRNA-target pairs in D1 are treated as positive samples in the training dataset. The interactions in IDS1 not in D1 are considered as the test samples of IDS1. The test samples of IDS2 are constructed in the same way. We also construct three independent datasets by different categories of biologically experimental methods based on D2. We select miRNA targets validated by Luciferase report assay to form IDS3 as an independent test dataset. Similarly, we build IDS4 using those validated by Western blot. IDS5 is a dataset consisting of miRNA targets validated by both Luciferase report and Western blot. The rest of the samples serves as the training set. Table 3 shows these datasets for detailed information.

TABLE II
THE DESCRIPTION OF THE DATASET FOR INDEPENDENT DATASET EXPERIMENTS

	miRNAs	genes	Interactions
D1	2,588	14,499	319,172
IDS1	2,588	14,499	374,566
IDS1-D1	0	0	55,394
D1	2,588	14,499	319,172
IDS2	2,588	14,742	377,236
IDS2-D1	0	243	58,064

”IDS1-D1” indicates that miRNAs, genes and interactions in IDS1 and not in D1, and so on.

IV. EXPERIMENTS AND RESULTS

Cross-validation method and independent dataset experiments are used to evaluate the proposed method in this paper.

TABLE III
THE DESCRIPTION OF THE DATASETS FOR INDEPENDENT DATASET EXPERIMENTS CONSTRUCTED BY LUCIFERASE REPORTER ASSAY AND/OR WESTERN BLOT EXPERIMENTS.

	LRA (IDS3)	WB (IDS4)	LRA and WB (IDS5)
number of testing miRNA targets	7,158	5,816	8,068
number of training miRNA targets	370,078	371,420	369,168
Total	377,236		

LRA is short for Luciferase reporter assay. WB is short for Western blot.

These evaluation methods have also been used in many works [57]–[59]. Three state-of-the-art methods miRTRS [26], deepTarget [19], GMCLDA [27], as well as two sequence-based methods miRanda [1] and TargetScan [60] are selected as the competing methods. Precision-Recall curves, ROC curves, and the AUC evaluation metrics are used to evaluate the performance of our method and the competing methods. In this section, we introduce the evaluation metrics and method firstly and then compare the results of our method and competing methods.

A. Validation methods and metrics

In order to evaluate the performance of miRTMC and the competing methods, we conduct 10-fold cross validation experiments on D1 and D2. In addition, we use five independent datasets to evaluate our method and competing methods. In 10-fold cross validation experiments, the experimentally validated miRNA targets are randomly divided into 10 parts with approximately same sizes, where in each fold, one is considered as the testing dataset alternately, while the rest 9 parts of them are used as training dataset. In independent dataset experiments, for datasets IDS1-D1 and IDS2-D1, all the experimentally validated miRNA targets in D1 are used to construct a prediction model for the prediction of newly added experimentally validated miRNA targets in IDS1 and IDS2, respectively. For IDS3, IDS4, and IDS5, table 3 describes the training and testing datasets. In the meanwhile, we also design a $de\ novo$ experiment to evaluate the performance of miRTMC and the competing methods in predicting miRNA targets for new genes. After predicting miRNA targets using these prediction methods, we obtain the score of each miRNA-gene pair. For each gene t_i, all the scores of unknown miRNA-gene pairs (miRNA m_j ($1 \leq j \leq p$, A_{ij}=0) that are not related to gene t_i) are sorted in descending order (ascending order for TargetScan). The numbers of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) are counted in each ranking threshold. The miRNA-gene pair is considered as a true positive if it appears in the testing set and is ranked higher than the threshold. The miRNA-gene pair is considered as a false positive if it does not appear in the testing set and is ranked higher than the threshold. TP and TN represent the numbers of correctly identified positive samples and negative samples, respectively. FP and FN represent the numbers of incorrectly identified positive samples and negative samples, respectively. The true positive rate (TPR), the false positive rate (FPR) and the precision are calculated by varying the rank threshold. The ROC curves are drawn by plotting TPR against FPR. The AUC values of the ROC curves are calculated to evaluate the performance of our method and the competing methods. The precision-recall curves are drawn by plotting precision against recall (TPR). In order to make our results statistically meaningful, we repeat the above 10-fold cross validation procedure 10 times and the average values are reported as the final results.

B. Results on dataset D1 by 10-fold cross validation

Our method is compared with miRTRS, deepTarget, miRanda, TargetScan and GMCLDA. miRTRS predicts miRNA targets by using a network-based inference approach on the bipartite network of miRNA-gene interaction. deepTarget predicts miRNA targets by training a recurrent neural networks with the training datasets. GMCLDA adopts geometric matrix completion algorithm to predict lncRNA-disease association. miRanda and TargetScan predict miRNA targets based on the binding rules between miRNAs and their targets. In this experiment, 9 parts of validated miRNA-targets are used to construct the miRNA-gene interaction network in miRTRS, miRTMC, GMCLDA. For deepTarget, the 9 parts of validated miRNA-targets are considered as positive training samples and the same number of negative samples are generated by using the algorithm used in deepTarget. The parameters which deepTarget needs are set as default as specified in its original paper. Since deepTarget is a sequence-based machine-learning miRNA target prediction method, its inputs are miRNA sequence and $3^{'}$ UTR sequence while its outputs are the scores of miRNA-mRNA pairs. The interactions in miRTarBase are the relationships between miRNAs and genes and, therefore, the maximum score between miRNA m_i and the transcripts (mRNA) of the same gene is selected as the score of this miRNA-gene pair. And the result of miRanda is processed in the same way. Since the targets with the lowest scores are the most representative miRNA-gene scores in TargetScan, the minimal score between miRNA m_i and the transcripts (mRNA) of the same gene is selected as the score of this miRNA-gene pair.

Hyperparameter γ is the learning rate of miRTMC, according to the experience [61], where we set $\gamma = 1.618$ in the experiments. λ and α are two main parameters in the miRTMC method. These two parameters need to be determined in the experiments. According to the experience and the scale of dataset, the value of λ for miRTMC is set to 10. Then we obtain the range of α by calculating the singular values of the to-be-complemented matrix in the first step of ours algorithm. Finally, by searching in this range, the value of α that maximizes the values of AUC is determined. According to the parameter determination method aforementioned, we set $\lambda = 10$, and since the SVT operator uses the singular values that are larger than $\frac{\lambda}{\alpha}$, we set $\alpha \in \{10/10000, 10/20000, ..., 10/90000, 10/100000\}$ in this experiment in order to let the cutoff singular value be in the set $\{10000, 20000, ..., 100000\}$. Searching in the range of

10 α values, we find that the AUC values become maximum and stable when λ=10, α=10/10000 and $tol = 1e - 4$.

Figure 2 shows the ROC curves of miRTMC and the other methods, where one can find that miRTMC outperforms the competing methods in terms of AUC values. The average AUC values of miRTMC, miRTRS, deepTarget, miRanda, TargetScan, GMCLDA are 0.9285±0.001, 0.8954±0.002, 0.8034±0.0058, 0.6550±0.0017, 0.7194±0.0028, and 0.7979±0.0017, respectively. miRTMC yields high TPRs at low FPRs. The paired t-test is adopted to further analyze the performance of these methods. The paired t-test is performed on the results of ten-fold cross validation. The significant difference between miRTMC and competing methods (miRTRS, deepTarget, miRanda, TargetScan, GMCLDA), reports p-values of 1.94E-23, 1.1E-26, 3.58E-44, 6.43E-38, and 8.22E-37, respectively. Figure 3 shows the precision-recall curves among miRTMC and the competing methods, indicating that the performance of miRTMC is the best among the five. miRTMC achieves the best accuracy 0.115, which indicates that it can predict 11.5% miRNA targets correctly when ranked in the first place.

In the 10-fold cross validation experiments, 31,917 miRNA-target pairs are selected as the testing set. Figure 4 shows the bar chart of the accumulated numbers of miRNA targets predicted correctly at top 10 among miRTMC and competing methods. As shown in figure 4, after predicting miRNA targets by using these models, miRTMC, miRTRS, deepTarget, miRanda, TargetScan, and GMCLDA have correctly predicted averagely 13,048, 9,728, 3,772, 633, 1,023, and 4,879 pairs in top-10 rankings, respectively. Moreover, miRTMC, miRTRS, and deepTarget have predicted 9,936, 7,346, and 3,678 pairs correctly in top-5 rankings on average, respectively. miRTMC is able to predict 31.3% and 40.9% miRNA targets correctly in top-5 ranking and top-10 ranking on average, respectively, which are significantly more than the correct predictions generated by the other methods.

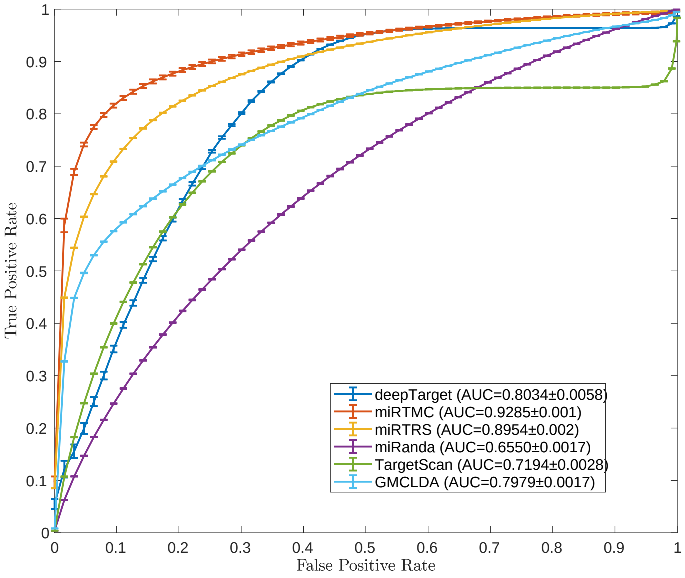

Fig. 2. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by ROC curves with the error bar on dataset D1.**

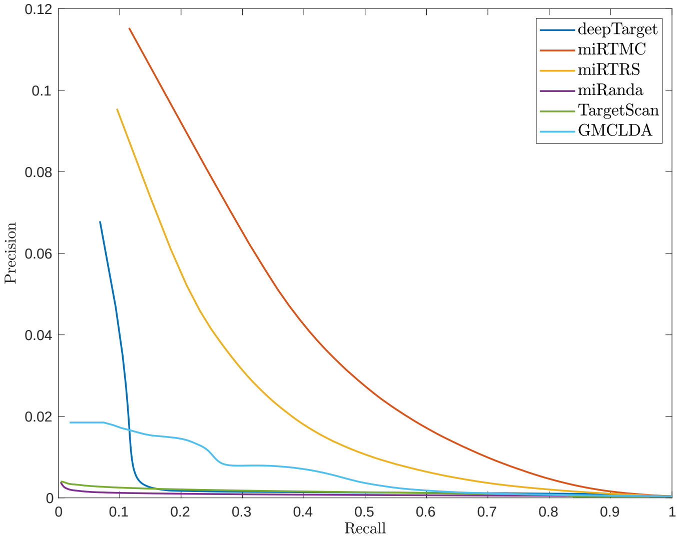

Fig. 3. **Comparison among miRTMC, miRTRS deepTarget, miRanda, TargetScan and GMCLDA by precision-recall curves on dataset D1.**

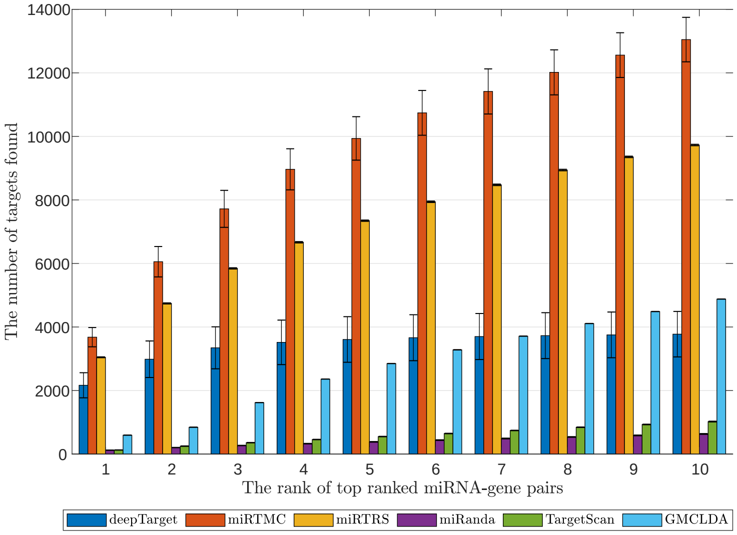

Fig. 4. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by bar chart with the error bar on dataset D1.**

C. Results on dataset D2 by 10-fold cross validation

We also compare miRTMC with three state-of-the-art prediction methods and two sequence-based miRNA target prediction methods aforementioned on D2 by 10-fold cross validation experiment. Similar cross-validation setup and parameters for these methods as those in D1 are adopted.

Figure 5 depicts the ROC curves of miRTMC and competing methods. miRTMC achieves average AUC value of 0.9216±0.0077, in comparison, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA yield the average AUC values of 0.8885±0.0012, 0.8040±0.0077, 0.6551±0.0015,

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.2020.2987034, IEEE Journal of Biomedical and Health Informatics

0.7303±0.0013 and 0.8004±0.0013, respectively. The paired t-test is performed on these results, the significant difference between miRTMC and competing methods (miRTRS, deepTarget, miRanda, TargetScan, GMCLDA) yields p-values of 2.18E-14, 1.2E-21, 1.06E-30, 4.64E-28, and 3.64E-24, respectively. Figure 6 shows the precision-recall curves of miRTMC and the competing methods. Similar to that of D1, the performance of miRTMC is better than the four competing methods. miRTMC achieves the best accuracy of 0.096. The bar chart in Figure 7 shows the accumulated miRNA targets predicted correctly at top 10 rank. In the 10-fold cross validation experiments, ten percent (37,724 pairs) of known miRNA targets need to be identified in each fold. The average number of correctly predicted miRNA-gene pairs for miRTMC is 13,850 by counting their appearances on the top 10, compared to 10,665 in miRTRS, 3,716 in deepTarget, 810 in miRanda, 1,261 in TargetScan and 6,053 in GMCLDA. In summary, miRTMC also outperforms the competing methods in terms of AUC and accuracy in experiments on dataset D2.

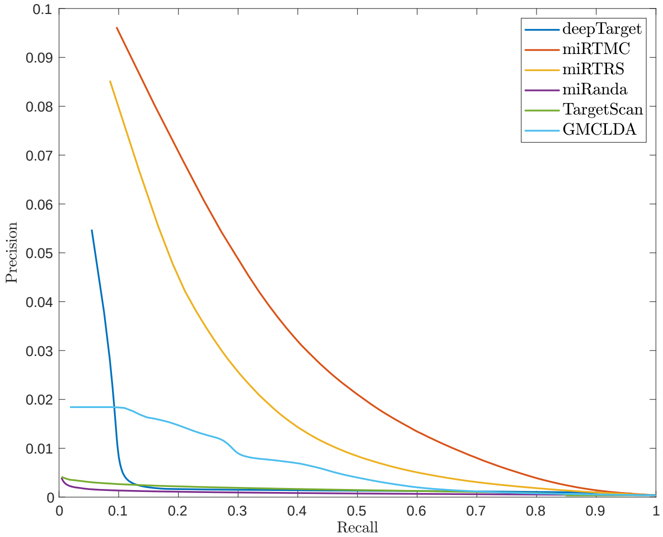

Fig. 6. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by precision-recall curves on dataset D2.**

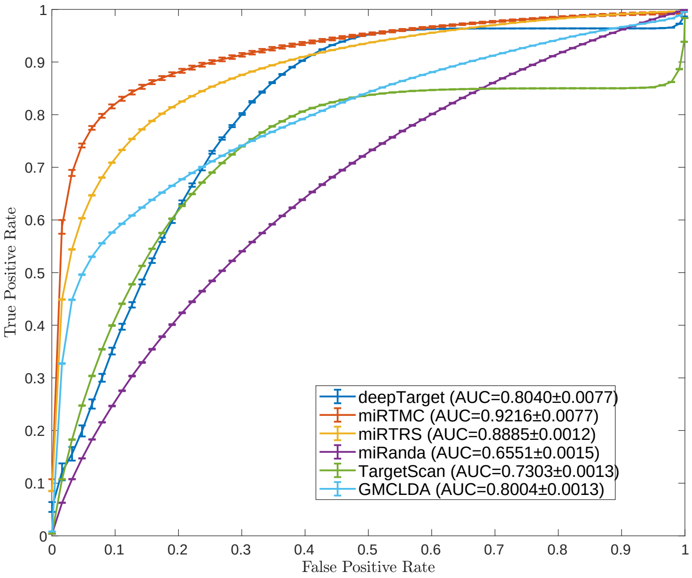

Fig. 5. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by ROC curves with the error bar on dataset D2.**

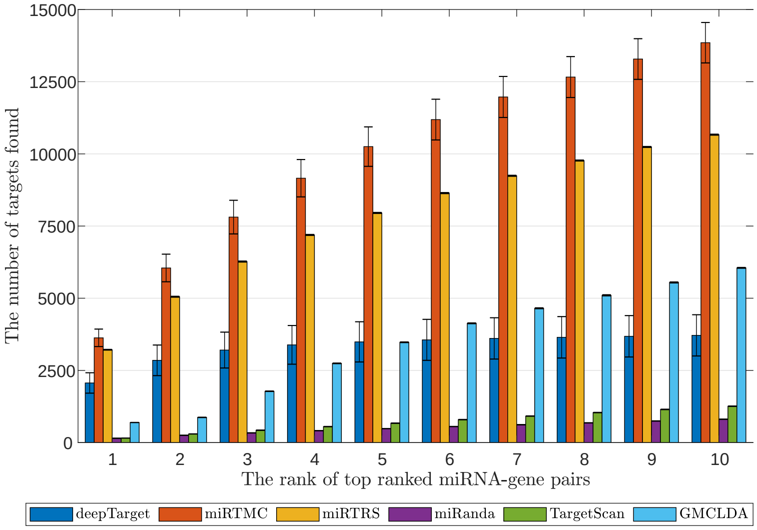

Fig. 7. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by bar chart with the error bar on dataset D2.**

D. Results on independent datasets

In order to evaluate the performance of miRTMC and competing methods systematically, five independent datasets have been prepared, as described in the Section of Materials, for these experiments. IDS1 and IDS2 independent datasets are prepared based on D1 and D2. All known miRNA targets in D1 are considered as the positive samples in the training dataset. For deepTarget methods, 319,172 negative samples (the same number as the positive samples) are generated by the method used by deepTarget. Then, the learning model of deepTarget is trained by these positive and negative samples. In miRTMC, miRTRS and GMCLDA, the models are constructed using the positive samples (D1) only. The newly added miRNA-gene interactions in D2 are predicted by these models as test samples. Datasets IDS3, IDS4, and IDS5 are constructed based on D2 by extracting the miRNA targets which are validated from Luciferase reporter assay and/or Western blot experiments. In these three datasets, table 3 describes the number of training samples (positive) and the number of testing samples, respectively. In miRTMC, miRTRS, and GMCLDA, the models are constructed by the training samples, and the learning model of deepTarget is trained by positive samples and negative samples (the same number as the positive samples) which are generated by the method used by deepTarget. The parameter determination method of miRTMC in the independent dataset experiment is the same as in 10-fold cross validation experiments. We set the parameters $\gamma = 1.618, \lambda = 10, \alpha = 10/100000$ and $tol = 1e - 4$.

The same ranking strategy method and the evaluation method as described in the 10-fold cross validation experiments are used in the independent dataset experiments. Same as in the 10-fold cross validation experiments on D1, in dealing with the scores of deepTarget and miRanda methods, the maximum score between miRNA m_i and the transcripts (mRNA) of the same gene is selected as the score of this miRNA-gene pair. In dealing with the scores of TargetScan, the minimal score between miRNA m_i and the transcripts (mRNA) of the same gene is selected as the score of this miRNA-gene pair. The AUC values of miRTMC and the competing methods are calculated for performance comparison. Table 4 compares the AUC values of miRTMC and the competing methods on these five independent datasets. One can find that miRTMC achieves the best AUC values of 0.717 and 0.717 on both of IDS1 and IDS2, compared to miRTRS, deepTarget, miRanda, TargetScan and GMCLDA. Moreover, miRTMC achieves 0.756, 0.760 and 0.749 on IDS3, IDS4 and IDS5, respectively. Note that because the independent dataset experiments on IDS1 and IDS2 use the same training dataset and IDS1 is a subset of IDS2, the AUC values of the same method on these two independent datasets are very close. We further analyze the prediction results of the two methods and find that miRTMC correctly predicts some important miRNA targets, while miRTRS dose not, although they have close AUC values. For example, in IDS4 experiment, miRTMC predicts the target gene ZEB2 of hsa-miR-335-5p. ZEB2 is related to colorectal cancer (PMID: 24829139) and this miRNA target pair has been verified by three different biological experiments [62]. miRTMC also correctly identifies the target gene HMGA2 of hsa-miR-16-5p. HMGA2 is related to human pituitary tumorigenesis (PMID: 22139073, 21572407). This miRNA target pair is also verified by three biological experiments [63][64].

TABLE IV
THE AUC VALUES OF THE MIRTMC AND COMPETING METHODS WITH DIFFERENT INDEPENDENT DATASETS.

	IDS1	IDS2	IDS3	IDS4	IDS5
miRTMC	**0.717**	**0.717**	**0.756**	**0.760**	**0.749**
miRTRS	0.70	0.699	0.749	0.759	0.740
deepTarget	0.690	0.684	0.701	0.712	0.693
miRanda	0.690	0.690	0.694	0.689	0.683
TargetScan	0.666	0.667	0.702	0.709	0.695
GMCLDA	0.612	0.611	0.690	0.700	0.688

E. Results on de novo experiment

miRTMC, miRTRS and GMCLDA are proposed based on the recommendation algorithms, which are often puzzled by the cold start problem. Both methods attempt to address the cold start issue for new genes without previously known miRNA targets. Compared with dataset D1, 243 new genes are newly added in IDS2 and these genes are not associated with any of miRNAs in D1. These genes have 2,670 interactions in total in IDS2. In order to compare the performance of miRTMC and the competing methods in predicting miRNA targets for the new genes, a *de novo* experiment is designed. For miRTMC, miRTRS and GMCLDA, all known miRNA targets in D1 are used to construct models for predicting the 2,670 miRNA targets newly added in IDS2. For deepTarget, all known miRNA targets in dataset D1 are considered as positive samples, meanwhile, the same number of negative samples as positive samples are generated. Then these samples are used to train deepTarget for predicting the 2,670 miRNA-target pairs of the newly added 243 gene. miRTMC is the best among these methods in terms of AUC. miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA achieve the AUC values of 0.726, 0.700, 0.690, 0.690, 0.666 and 0.620, respectively. We count the numbers of the correct miRNA targets found by each method at top 1 to top 10, and the results are shown in Figure 8. In particular, miRTMC correctly predicts 11 miRNA target pairs in top 5, while deepTarget, miRTRS, miRanda, TargetScan and GMCLDA only correctly predict 3 pairs, 1 pair, 5 pairs, 7 pairs and 8 pairs, respectively.

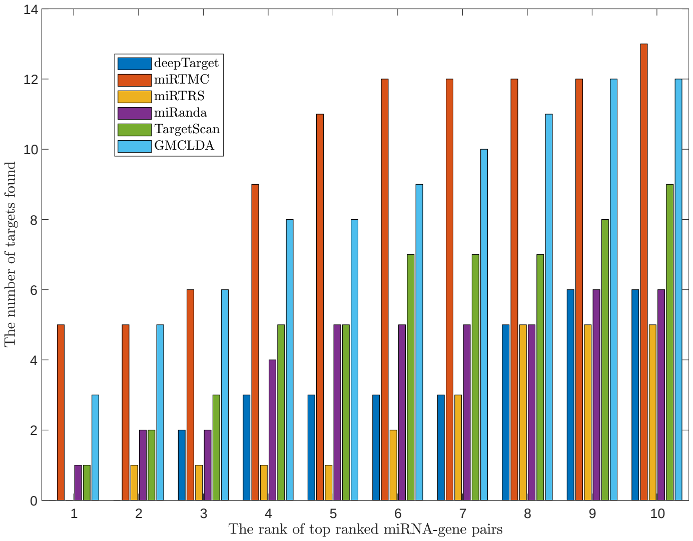

Fig. 8. **Comparison among miRTMC, miRTRS, deepTarget, miRanda, TargetScan and GMCLDA by bar chart at top 1 to top 10 ranked candidates, respectively, in *de novo* experiment.**

V. CONCLUSIONS

Computational approaches for miRNA target prediction are valuable complements to the biological experimental studies on miRNA targets identification. Existing computational methods for predicting miRNA targets suffer from the relatively high false positive rates. In order to reduce the false positive rate of miRNA target prediction, inspired by the collaborative filtering recommendation algorithm, we propose a miRNA target prediction method formulated as a matrix completion problem. The miRNA-gene interaction that is not experimentally validated is predicted by filling out the unknown elements in the miRNA-gene interaction matrix. The biologically experimentally validated miRNA targets are used to construct a heterogeneous network, composed of the miRNA sequence similarity network, and the gene sequence similarity network, and the miRNA-gene association network.

Matrix completion is carried out on the adjacency matrix of the heterogeneous network. ADMM is adopted to solve matrix completion problem with non-negative constraints. The SVD operation on a large miRNA-gene matrix is accelerated by R^{4}SVD method. Compared to the existing machine learning methods for miRNA target prediction, miRTMC does not need negative samples, which are often difficult to obtain in practice. 10-fold cross validations and independent dataset experiments are carried out to evaluate miRTMC and the competing methods. Our results show that the accuracy of miRTMC is superior to those of competing methods.

The performance of miRTMC is sensitive to the computation of miRNA similarity or gene similarity. Since the matrix completion method is based on known relationships, the performance of the miRTMC method may also be affected by the sparsity of the adjacency matrix of miRNA-gene heterogeneous network. The accurate identification of a miRNA targets relies on extracting effective features characterizing miRNAs and genes. Deep learning graph convolution has demonstrated promising results in matrix completion models, such as Mgcnn [65] and GC-MC [66], for filling out user-item associations, compared to traditional recommendation system methods. Our future research will be investigating novel graph convolution models appropriate for miRNA and gene graphs to extract effective features to further improve miRNA-target association predictions.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grants (No.61972423, No. U1909208); the 111 Project (No.B18059); and the Hunan Provincial Science and Technology Program (2018WK4001); Research Project Foundation of the education department of Hunan Province (No. 17C1377).

References

[1] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, “Microrna targets in drosophila,” *Genome biology*, vol. 5, no. 1, p. R1, 2003.

[2] P. Brodersen and O. Voinnet, “Revisiting the principles of microrna target recognition and mode of action,” *Nature reviews Molecular cell biology*, vol. 10, no. 2, pp. 141–148, 2009.

[3] A. Esquela-Kerscher and F. J. Slack, “Oncomirs—micrornas with a role in cancer,” *Nature Reviews Cancer*, vol. 6, no. 4, pp. 259–269, 2006.

[4] P. Jin, D. C. Zarnescu, S. Ceman, M. Nakamoto, J. Mowrey, T. A. Jongens, D. L. Nelson, K. Moses, and S. T. Warren, “Biochemical and genetic interaction between the fragile x mental retardation protein and the microrna pathway,” *Nature neuroscience*, vol. 7, no. 2, pp. 113–117, 2004.

[5] C. Xu, Y. Lu, Z. Pan, W. Chu, X. Luo, H. Lin, J. Xiao, H. Shan, Z. Wang, and B. Yang, “The muscle-specific micrornas mir-1 and mir-133 produce opposing effects on apoptosis by targeting hsp60, hsp70 and caspase-9 in cardiomyocytes,” *Journal of cell science*, vol. 120, no. 17, pp. 3045–3052, 2007.

[6] Y. Huang, X. J. Shen, Q. Zou, S. P. Wang, S. M. Tang, and G. Z. Zhang, “Biological functions of micrornas: a review,” *Journal of physiology and biochemistry*, vol. 67, no. 1, pp. 129–139, 2011.

[7] D. W. Thomson, C. P. Bracken, and G. J. Goodall, “Experimental strategies for microrna target identification,” *Nucleic acids research*, vol. 39, no. 16, pp. 6845–6853, 2011.

[8] X. Fan and L. Kurgan, “Comprehensive overview and assessment of computational prediction of microrna targets in animals,” *Briefings in bioinformatics*, vol. 16, no. 5, pp. 780–794, 2015.

[9] L. Wei, Y. Huang, Y. Qu, Y. Jiang, and Q. Zou, “Computational analysis of mirna target identification,” *Current Bioinformatics*, vol. 7, no. 4, pp. 512–525, 2012.

[10] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, “Prediction of mammalian microrna targets,” *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[11] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets,” *cell*, vol. 120, no. 1, pp. 15–20, 2005.

[12] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, “Fast and effective prediction of microrna/target duplexes,” *Rna*, vol. 10, no. 10, pp. 1507–1517, 2004.

[13] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, “The role of site accessibility in microrna target recognition,” *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[14] C. Burgler and P. M. Macdonald, “Prediction and verification of microrna targets by movingtargets, a highly adaptable prediction method,” *BMC genomics*, vol. 6, no. 1, p. 88, 2005.

[15] X. Wang, “Improving microrna target prediction by modeling with unambiguously identified microrna-target pairs from clip-ligation studies,” *Bioinformatics*, vol. 32, no. 9, pp. 1316–1322, 2016.

[16] H. Liu, D. Yue, Y. Chen, S.-J. Gao, and Y. Huang, “Improving performance of mammalian microrna target prediction,” *BMC bioinformatics*, vol. 11, no. 1, p. 476, 2010.

[17] S. Yu, J. Kim, H. Min, and S. Yoon, “Ensemble learning can significantly improve human microrna target prediction,” *Methods*, vol. 69, no. 3, pp. 220–229, 2014.

[18] S. Cheng, M. Guo, C. Wang, X. Liu, Y. Liu, and X. Wu, “Mirtdl: A deep learning approach for mirna target prediction,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 6, pp. 1161–1169, 2016.

[19] B. Lee, J. Baek, S. Park, and S. Yoon, “deeptarget: end-to-end learning framework for microrna target prediction using deep recurrent neural networks,” in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2016, pp. 434–442.

[20] M. Wen, P. Cong, Z. Zhang, H. Lu, and T. Li, “Deepmirtar: a deep-learning approach for predicting human mirna targets,” *Bioinformatics*, vol. 34, no. 22, pp. 3781–3787, 2018.

[21] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, “mirecords: an integrated resource for microrna–target interactions,” *Nucleic acids research*, vol. 37, no. suppl_1, pp. D105–D110, 2008.

[22] I. S. Vlachos, M. D. Paraskevopoulou, D. Karagkouni, G. Georgakilas, T. Vergoulis, I. Kanellos, I.-L. Anastasopoulos, S. Maniou, K. Karathanou, D. Kalfakakou *et al.*, “Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions,” *Nucleic acids research*, vol. 43, no. D1, pp. D153–D159, 2015.

[23] R. A. Fisher and F. Yates, *Statistical tables for biological, agricultural and medical research*. Oliver and Boyd Ltd, London, 1943.

[24] M. Menor, T. Ching, X. Zhu, D. Garmire, and L. X. Garmire, “mirmark: a site-level and utr-level classifier for mirna target prediction,” *Genome biology*, vol. 15, no. 10, p. 500, 2014.

[25] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, “Mapping the human mirna interactome by clash reveals frequent noncanonical binding,” *Cell*, vol. 153, no. 3, pp. 654–665, 2013.

[26] H. Jiang, J. Wang, M. Li, W. Lan, F. Wu, and Y. Pan, “mirtrs: A recommendation algorithm for predicting mirna targets,” *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[27] C. Lu, M. Yang, M. Li, Y. Li, F. Wu, and J. Wang, “Predicting human lncrna-disease associations based on geometric matrix completion,” *IEEE Journal of Biomedical and Health Informatics*, 2019.

[28] 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.

[29] M. Yang, H. Luo, Y. Li, F.-X. Wu, and J. Wang, “Overlap matrix completion for predicting drug-associated indications.” *PLoS Computational Biology*, vol. 15, no. 12, 2019.

[30] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, “Computational drug repositioning using low-rank matrix approximation and randomized algorithms,” *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, 2018.

[31] H. Luo, J. Wang, C. Yan, M. Li, F. Wu, and Y. Pan, “A novel drug repositioning approach based on collaborative metric learning,” *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[32] M. Yang, H. Luo, Y. Li, and J. Wang, “Drug repositioning based on bounded nuclear norm regularization,” *Bioinformatics*, vol. 35, no. 14, pp. i455–i463, 2019.

[33] G. Duan, C. Yan, F. Wu, Y. Pan, and J. Wang, "Mchmda: Predicting microbe-disease associations based on similarities and low-rank matrix completion," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[34] C. Tang, H. Zhou, X. Zheng, Y. Zhang, and X. Sha, "Dual laplacian regularized matrix completion for microrna-disease associations prediction," *RNA biology*, vol. 16, no. 5, pp. 601–611, 2019.

[35] X. Chen, L. Wang, J. Qu, N.-N. Guan, and J.-Q. Li, "Predicting mirna–disease association based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 24, pp. 4256–4265, 2018.

[36] S.-P. Yu, C. Liang, Q. Xiao, G.-H. Li, P.-J. Ding, and J.-W. Luo, "Mclpmda: A novel method for mi rna-disease association prediction based on matrix completion and label propagation," *Journal of cellular and molecular medicine*, vol. 23, no. 2, pp. 1427–1438, 2019.

[37] X. Chen, L.-G. Sun, and Y. Zhao, "Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion," *Briefings in Bioinformatics*, 2020.

[38] X. Chen, J. Yin, J. Qu, and L. Huang, "Mdhgi: Matrix decomposition and heterogeneous graph inference for mirna-disease association prediction," *PLoS computational biology*, vol. 14, no. 8, p. e1006418, 2018.

[39] J. Ha, C. Park, and S. Park, "Pmamca: prediction of microrna-disease association utilizing a matrix completion approach," *BMC systems biology*, vol. 13, no. 1, p. 33, 2019.

[40] Z. Shen, Y.-H. Zhang, K. Han, A. K. Nandi, B. Honig, and D.-S. Huang, "mirna-disease association prediction with collaborative matrix factorization," *Complexity*, vol. 2017, 2017.

[41] L. Peng, M. Peng, B. Liao, G. Huang, W. Liang, and K. Li, "Improved low-rank matrix recovery method for predicting mirna-disease association," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[42] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[43] Y. Li and W. Yu, "A fast implementation of singular value thresholding algorithm using recycling rank revealing randomized singular value decomposition," *arXiv preprint arXiv:1704.05528*, 2017.

[44] C.-H. Chou, S. Shrestha, C.-D. Yang, N.-W. Chang, Y.-L. Lin, K.-W. Liao, W.-C. Huang, T.-H. Sun, S.-J. Tu, W.-H. Lee *et al.*, "mirtarbase update 2018: a resource for experimentally validated microrna-target interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D296–D302, 2017.

[45] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "mirwalk–database: prediction of possible mirna binding sites by "walking" the genes of three genomes," *Journal of biomedical informatics*, vol. 44, no. 5, pp. 839–847, 2011.

[46] R.-M. Tsoupidi, I. Kanellos, T. Vergoulis, I. S. Vlachos, A. G. Hatzigeorgiou, and T. Dalamagas, "Tarminer: automatic extraction of mirna targets from literature," in *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*. ACM, 2015, p. 12.

[47] D. Karagkouni, M. D. Paraskevopoulou, S. Chatzopoulos, I. S. Vlachos, S. Tastsoglou, I. Kanellos, D. Papadimitriou, I. Kavakiotis, S. Maniou, G. Skoufos *et al.*, "Diana-tarbase v8: a decade-long collection of experimentally supported mirna–gene interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D239–D245, 2017.

[48] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[49] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[50] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[51] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, no. 2, pp. 227–234, 1995.

[52] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing," *SIAM Journal on Imaging sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[53] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[54] H. Ji, W. Yu, and Y. Li, "A rank revealing randomized singular value decomposition (r3svd) algorithm for low-rank matrix approximations," *arXiv preprint arXiv:1605.08134*, 2016.

[55] A. Kozomara and S. Griffiths-Jones, "mirbase: annotating high confidence micrornas using deep sequencing data," *Nucleic acids research*, vol. 42, no. D1, pp. D68–D73, 2014.

[56] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu *et al.*, "mirtarbase 2016: updates to the experimentally validated mirna-target interactions database," *Nucleic acids research*, vol. 44, no. D1, pp. D239–D247, 2016.

[57] Y.-Y. Ou *et al.*, "Prediction of fad binding sites in electron transport proteins according to efficient radial basis function networks and significant amino acid pairs," *BMC bioinformatics*, vol. 17, no. 1, p. 298, 2016.

[58] Q.-T. Ho, Y.-Y. Ou *et al.*, "Classifying the molecular functions of rab gtpases in membrane trafficking using deep convolutional neural networks," *Analytical biochemistry*, vol. 555, pp. 33–41, 2018.

[59] N. Le and B. Nguyen, "Prediction of fmn binding sites in electron transport chains based on 2-d cnn and pssm profiles." *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[60] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microrna target sites in mammalian mrnas," *elife*, vol. 4, p. e05005, 2015.

[61] R. Glowinski and J. T. Oden, "Numerical methods for nonlinear variational problems," *Journal of Applied Mechanics*, vol. 52, p. 739, 1985.

[62] Z. Sun, Z. Zhang, Z. Liu, B. Qiu, K. Liu, and G. Dong, "Microrna-335 inhibits invasion and metastasis of colorectal cancer by targeting zeb2," *Medical oncology*, vol. 31, no. 6, p. 982, 2014.

[63] D. Palmieri, D. D'angelo, T. Valentino, I. De Martino, A. Ferraro, A. Wierinckx, M. Fedele, J. Trouillas, and A. Fusco, "Downregulation of hmga-targeting micrornas has a critical role in human pituitary tumorigenesis," *Oncogene*, vol. 31, no. 34, pp. 3857–3865, 2012.

[64] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, "A quantitative analysis of clip methods for identifying binding sites of rna-binding proteins," *Nature methods*, vol. 8, no. 7, p. 559, 2011.

[65] F. Monti, M. Bronstein, and X. Bresson, "Geometric matrix completion with recurrent multi-graph neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 3697–3707.

[66] R. v. d. Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," *arXiv preprint arXiv:1706.02263*, 2017.