Cui et al. BMC Bioinformatics 2019, 20(Suppl 8):287
https://doi.org/10.1186/s12859-019-2768-7

BMC Bioinformatics

RESEARCH Open Access

$L_{2,1}$-GRMF: an improved graph regularized matrix factorization method to predict drug-target interactions

Zhen Cui[1], Ying-Lian Gao[2], Jin-Xing Liu[1,3*], Ling-Yun Dai[1] and Sha-Sha Yuan[1]

From International Conference on Data Science, Medicine and Bioinformatics
Wenzhou, China. 22- 24 June 2018

Abstract

Background: Predicting drug-target interactions is time-consuming and expensive. It is important to present the accuracy of the calculation method. There are many algorithms to predict global interactions, some of which use drug-target networks for prediction (ie, a bipartite graph of bound drug pairs and targets known to interact). Although these algorithms can predict some drug-target interactions to some extent, there is little effect for some new drugs or targets that have no known interaction.

Results: Since the datasets are usually located at or near low-dimensional nonlinear manifolds, we propose an improved GRMF (graph regularized matrix factorization) method to learn these flow patterns in combination with the previous matrix-decomposition method. In addition, we use one of the pre-processing steps previously proposed to improve the accuracy of the prediction.

Conclusions: Cross-validation is used to evaluate our method, and simulation experiments are used to predict new interactions. In most cases, our method is superior to other methods. Finally, some examples of new drugs and new targets are predicted by performing simulation experiments. And the improved GRMF method can better predict the remaining drug-target interactions.

Keywords: Drug-target interaction prediction, Graph regularization, $L_{2,1}$-norm, Matrix factorization, Manifold learning

Background

With advances in drug discovery technologies, the existing methods can identify drug targets to some extent. But drug development is a high-cost, inefficient problem [1]. For drug developers, there has been a great deal of interest in the repositioning of drugs. This repositioning has some potential to reduce risk time and cost [2]. A crucial element for the repositioning of medicines is online biological databases such as KEGG [3], DrugBank [4], STITCH [5] and ChEMBL [6], which store a large number of current drug-target interactions. It is worth noting that there are still many interactions that have not been found [7]. Therefore, the advances of drug-target prediction technology is accelerated, and more and more prediction methods are proposed [8]. These computations, which reasonably predict new and unexplored interactions, have greatly facilitated the drug discovery process, making the process more credible. Recent research shows that there are three popular methods for predicting drug-target interactions, such as ligand-based methods [9], docking-based methods [10], and chemogenomic approaches [11]. Of course, we can also use the opposition-based learning particle swarm optimization to predict interactions, such as SNP-SNP interactions [12]. Moreover, the potential gene-gene interactions network can be identified by LNDriver [13].

Recently, many researchers have used matrix decomposition methods to solve drug-target interaction

* Correspondence: sdcavell@qfnu.edu.cn
[1]School of Information Science and Engineering, Qufu Normal University, Rizhao, China
[3]Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei, China
Full list of author information is available at the end of the article

problems. The main methods are Bayesian matrix factorization, KBMF2K [14] and collaborative matrix factorization method, CMF [15]. A high-dimensional drug-target interaction matrix is decomposed into a plurality of low-dimensional matrices, and these matrices have characteristics of the original matrices, which is the principle of these methods. However, in theories, the above methods of matrix factorization still have some room for improvement. [16].

Using chemogenomic approaches to predict drug-target interactions is an effective method. The reason is that the first two methods have their own drawbacks. If a docking simulation is used, the three-dimensional structure of a target protein must be available. Furthermore, for ligand-based methods, if there are few or no target proteins known, this would be a problem that cannot be ignored. [9]. The advantage of using chemical genomics is that the information from the drugs and targets is used simultaneously for prediction [17]. New interactions are inferred by calculating the similarity of the chemical structures between drugs and the similarity of the genomic sequences between the targets. In this paper, the drug similarity and the target similarity are based on the construction methods in previous studies, which are based on the characteristics of the drug and the characteristics of the target. Its advantage is that we are better able to compare it with other methods, which is universal. However, if the same construction method of the drug similarity and the target similarity is used, this may affect the final results.

Two separate models are used to train drug target pairs, one based on the drug side and the other based on the target side. Thus, the final results are solved by predicting these two aspects. In this paper, to avoid over-fitting and sparing the target, the $L_{2,1}$-norm is added in our method, which can eliminate some unattached target pairs [18]. Ten-fold cross-validation is used to evaluate the performance of our method.

We present the experimental results in Results. In Datasets, we conducted a case study. And we summarize this paper in Cross-validation experiments. In Interaction prediction under CVd, we clearly introduced the methods, including specific iteration formulas and algorithms.

Results

Datasets

Four datasets are used to experiment: the nuclear receptor (NR), the G protein-coupled receptor (GPCR), the ion channel (IC) and the enzyme (E). The size of these four datasets is different. Nuclear receptors are one of the most abundant transcriptional regulators in metazoans. NR includes some steroid hormones, vitamin D and quinone. In recent years, nuclear receptors have received widespread attention. For example, they are closely related to the development of diseases such as diabetes and fatty liver. Among them, PPAR-g agonist thiazolidinedione rosiglitazone can effectively improve insulin sensitivity in diabetic patients. GPCRs are one of the target enzymes that are important proteins in cell signaling and have so far been found as therapeutic drugs. The total number of targets is about 500, and GPCR targets account for the vast majority of receptors therein. In recent years, indications for targeting GPCR drugs are expanding from traditional areas such as allergies, hypertension, anesthesia and schizophrenia to new areas such as obesity. An ion channel is a pore-forming protein that traverses the channel by allowing an ion of a particular type to rely on an electrochemical gradient. ICs are small pores in the cell membrane that allow ions to enter and exit the cell. Therefore, most of them have become the targets of some mainstream drugs. Enzymes are macromolecular biocatalysts. Some common drugs use enzymes as targets, and some effects on enzymes such as inhibition, induction, activation or reactivation are exerted. In addition, drugs like this are mostly enzyme inhibitors. According to statistics, half of the top 20 drugs in the world are enzyme inhibitors. It is worth noting that some drugs are enzymes themselves, such as pepsin and trypsin.

Each dataset contains three matrices, $\mathbf{Y}$, $\mathbf{S_d}$ and $\mathbf{S_t}$. Matrix $\mathbf{Y}$ represents the drug-target interactions. It is worth noting that this matrix is an adjacency matrix. If it is known that the drug d_i is related to the target t_j, Y_{ij} is 1, otherwise Y_{ij} is 0. The matrix $\mathbf{S_d}$ represents the chemical pairing structural similarity [19] and the matrix $\mathbf{S_t}$ represents the genome sequence similarity of the target pair [20]. Table 1 lists the specific information for the four datasets. More information about the datasets are published in https://github.com/cuizhensdws/L21-GRMF.

Cross-validation experiments

We compare the existing matrix decomposition methods CMF (Collaborative matrix factorization), GRMF (Graph regularized matrix factorization), WGRMF (Weighted graph regularized matrix factorization) and our proposed method and compare WKNKN preprocessing on these methods. We use cross-validation experiments on these methods. In this paper, we use a ten-fold cross-validation (CV). The original dataset $\mathbf{Y}$ is divided into ten subsets, each of which is tested once and the rest as a training set. The cross-validation is repeated five times, one subset is selected each time as a test set, and the average

Table 1 Drugs, Targets, and Interactions in Each Dataset

Datasets	NR	GPCR	IC	E
Drugs	54	223	210	445
Targets	26	95	204	664
Interactions	90	635	1476	2923

cross-validation recognition accuracy rate of five times is taken as a result.

To verify the effect of the prediction, we use the evaluation index which has been widely used before, the AUPR (Area under the Precision-Recall curve) [21]. There is also an evaluation scale called AUC (Area under the receiver operating characteristic curve). We can use this method when forecasting. In our experiments, ten AUPR values are calculated for each ten-fold cross-validation, an average is obtained and we repeat five times, so we take the average of the five AUPRs as the final result [22]. In general, the AUPR value is less than the AUC value. The AUPR value is above 0.3, so the experimental results are reasonable.

We test two aspects [23], one is CVd which is based on the drug-interaction profiles and the other is CVt, which is based on the target-interaction profiles. CVd is used to test the ability to predict new drugs, CVt is used to test the ability to predict new targets. In addition, we perform a convergence analysis of each method using the NR and GPCR datasets as examples, and each method is subjected to 100 iterations. When the number of iterations is about 20, our method achieves convergence. It is worth noting that we have different tolerances for errors, considering the size and type of the datasets. Generally speaking, as long as the error is within a reasonable range, this is acceptable. Figures 1 and 2 show the convergence of different methods on the NR and GPCR datasets, respectively.

Interaction prediction under CVd

Table 2 lists the experimental results at CVd. And Standard deviations are given in parentheses. Under the NR dataset, the $L_{2,1}$-GRMF ($L_{2,1}$-norm Graph regularized matrix factorization) method is superior to the GRMF method and is almost the same as the GRMF method after adding the WKNKN. Importantly, our improved method $L_{2,1}$-GRMF, with the addition of WKNKN, has seen significant improvements. Moreover, after adding the weight matrix to $L_{2,1}$-GRMF and using WKNKN, the accuracy of prediction is also improved. Figure 3 shows the PR curves on the CVd side of each method on the NR dataset.

However, on the GPCR dataset, we run our method and find that it is not outperform the previous method, and initially estimate that there is a problem with the dataset itself. Figure 4 shows the PR curves on the CVd side of each method on the GPCR dataset. We observe that using the weight matrix when performing CVd experiments is higher than the AUPR value obtained without using the weight matrix. In addition, the $L_{2,1}$-WGRMF (Weighted $L_{2,1}$-norm graph regularized matrix factorization) method using WKNKN is superior to any other method in the IC dataset, slightly better than the WGRMF method using WKNKN. Figure 5 shows the PR curves on the CVd side of each method on the IC dataset. In the E dataset, the best method is $L_{2,1}$-WGRMF but the AUPR score drops instead after applying WKNKN. In other words, in the E dataset, the preprocessing step will actually have a negative effect on the forecast result. Figure 6 shows the PR curves on the CVd side of each method on the E dataset. In general, not all methods use WKNKN to improve AUPR scores, which have a positive effect on most datasets and negative effects on some datasets. In practice, the negative impact of the WKNKN method is unavoidable on some datasets. One important reason is that the WKNKN

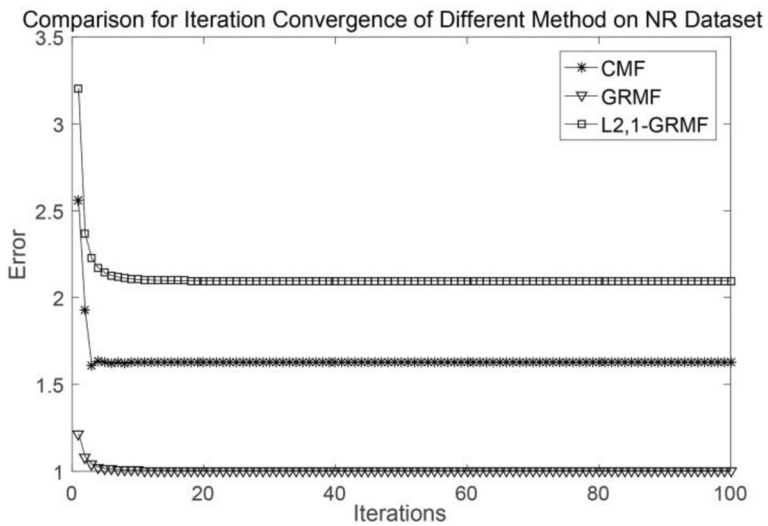

Fig. 1 Comparison of convergence about three methods on the NR dataset

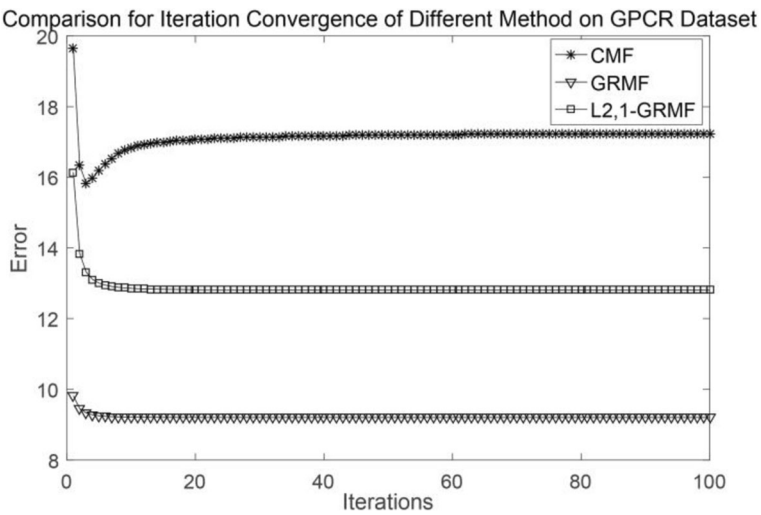

Fig. 2 Comparison of convergence about three methods on the GPCR dataset

method assigns an inaccurate value to the 0 element of the matrix $\mathbf{Y}$ on the E dataset. When we add the $L_{2,1}$-GRMF method to make more accurate predictions, these inaccurate values will reduce the prediction accuracy.

Interaction prediction under CVt

We can see in Table 3 that under most datasets, the AUPR value of CVt is generally higher than the AUPR value of CVd. This shows that hiding the interactions of the target can still get a better prediction result. But hiding the drug interactions and the prediction result will be greatly reduced. And standard deviations are given in parentheses. It is worth noting that in most datasets, the CMF method has lower AUPR values than any other method, and its AUPR value is far less than our method, especially in the NR dataset.

Discussion

Among the NR, GPCR and IC datasets, the superior methods are the $L_{2,1}$-GRMF method using the preprocessing steps, and our improved method has some improvement on all three datasets. Figures 7, 8, 9 and 10 show the PR curves on the CVt side of each method on the NR, GPCR, IC and E datasets, respectively. On the E dataset, it is still the best GRMF method. We can also see that some instances are ignored after using the weight matrix, whereas the GRMF method does not use the weight matrix $\mathbf{W}$. Therefore, based on the previous conclusions, the information of the target is more important than the information of the drug. Therefore, using the GRMF method, the AUPR value is higher than the AUPR value using WGRMF.

On most datasets, the $L_{2,1}$-norm does play a key role in predicting the results. The $L_{2,1}$-norm can provide a sparse solution for the final result. Compared with the CMF method, the $L_{2,1}$-norm also promotes the final convergence. Therefore, the overall performance of the $L_{2,1}$-GRMF method and $L_{2,1}$-WGRMF is superior to other methods.

Table 2 AUPR Results for Interaction Prediction Under CVd

Methods	NR	GPCR	IC	E
CMF	0.482(0.034)	0.406(0.008)	0.350(0.008)	0.375(0.007)
GRMF	0.517(0.025)	0.369(0.011)	0.341(0.016)	0.349(0.012)
WGRMF	0.520(0.025)	0.408(0.010)	0.364(0.018)	0.404(0.014)
$L_{2,1}$-GRMF	0.543(0.034)	0.373(0.011)	0.345(0.012)	0.346(0.013)
$L_{2,1}$-WGRMF	0.542(0.024)	0.400(0.010)	0.370(0.016)	0.408(0.013)
WKNKN+CMF	0.515(0.032)	0.409(0.010)	0.350(0.014)	0.385(0.004)
WKNKN+GRMF	0.542(0.028)	0.404(0.011)	0.356(0.014)	0.390(0.010)
WKNKN+WGRMF	0.528(0.033)	0.410(0.012)	0.369(0.017)	0.401(0.013)
WKNKN+$L_{2,1}$-GRMF	0.573(0.011)	0.394(0.007)	0.356(0.012)	0.386(0.013)
WKNKN+$L_{2,1}$-WGRMF	0.544(0.026)	0.394(0.012)	0.374(0.016)	0.385(0.007)

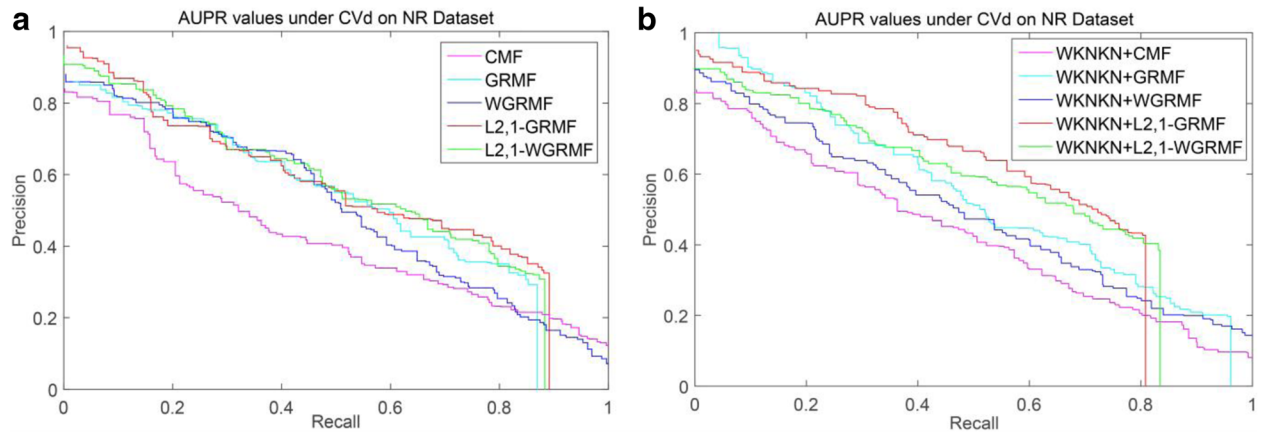

Fig. 3 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVd side of each method on the NR dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

Case study

In this section, we conduct a simulation experiment. First, we erase some of the known drug targets in the original dataset. That is, those elements that are originally 1 in the original matrix become 0. This process is performed randomly by the computer. In the second step, we perform the experiment. We examine the results of the experiment and see if the erased condition is successfully predicted.

The experimental procedure we implement is that in the NR dataset, ten drugs with the interaction of the target estrogen receptor alpha (KEGG ID: hsa2099) are removed. This target is the main cause of breast cancer. After the experiment is done, we count the experimental results. We predict five of the hidden interactions. At the same time, we also predict a portion of new drugs and take the most reliable top five new drugs stated in Table 4. Among them, the sixth drug Testosterone is the drug with the highest correlation with this target.

In IC dataset, for the drug Diazoxide (KEGG ID: D00294), a blood pressure lowering drug. We also use a similar approach. Before using the $L_{2,1}$-GRMF method, we eliminate twenty of them in the matrix $\mathbf{Y}$. Because the GPCR dataset is larger than the NR dataset and there are many targets associate with this drug, we have removed twenty interactions here. After conducting simulation experiments, we successfully predicted twelve known targets and eight new targets. We then list the top twenty targets in Table 5. The first 12 are known targets and the remaining part is our prediction of a new target.

For these two cases, the similarity of the estrogen receptor alpha to its nearest neighbor target is less than 0.02 in the matrix $\mathbf{S^t}$. In the matrix $\mathbf{S^d}$, the similarity of Diazoxide to its nearest neighbor is 0.3, which is also

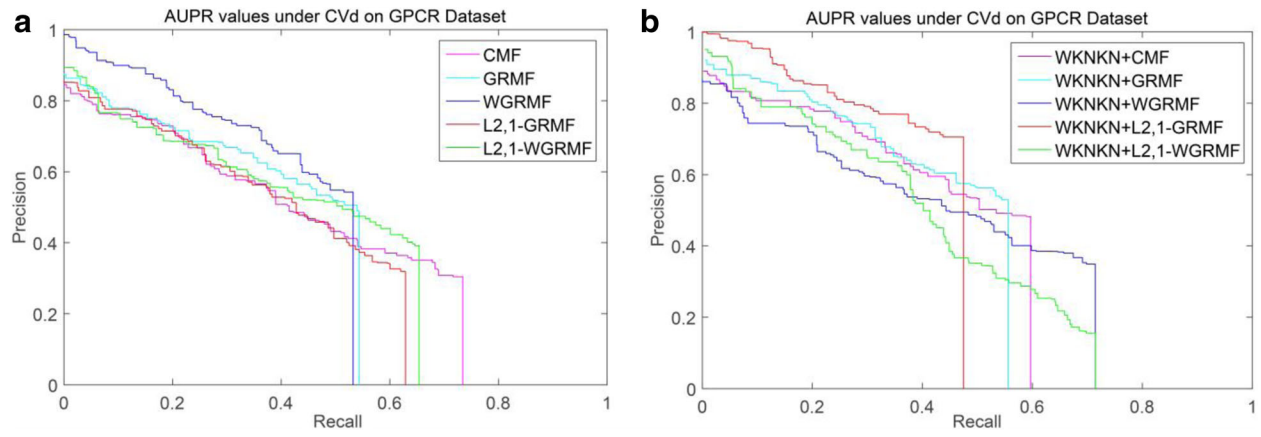

Fig. 4 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVd side of each method on the GPCR dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

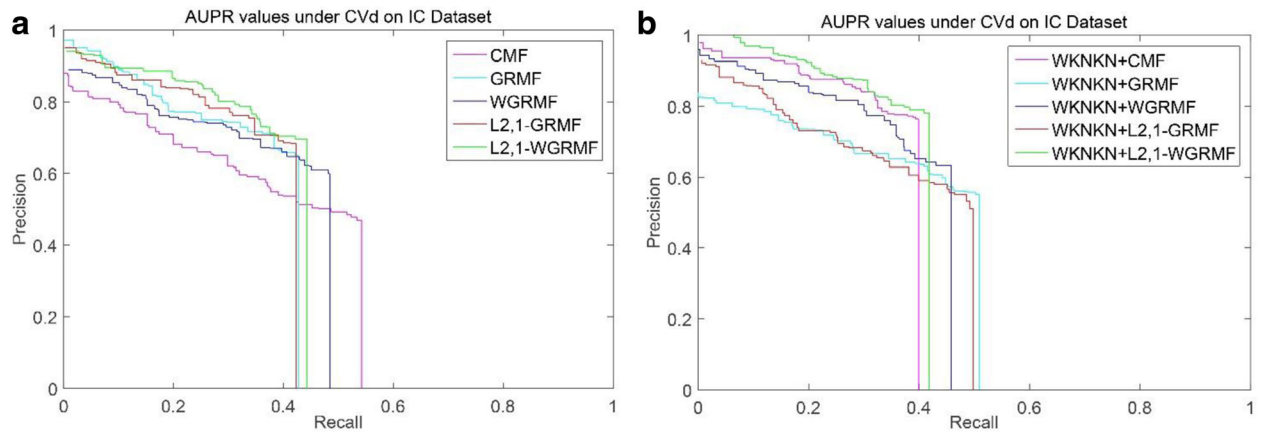

Fig. 5 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVd side of each method on the IC dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

quite low. Therefore, we are more difficult to make predictions. Thus, this shows that our proposed $L_{2,1}$-GRMF method is excellent and reliable results can be obtained when predicting some challenging drugs and targets. Of course, there are still some limitations to the two methods proposed. If we add a weight matrix, the time required for the experiment will multiply. Compared with other methods, our time complexity is relatively high. In addition, the method does not predict new drugs and new targets without any interaction.

Conclusions

In this paper, we propose two improved matrix decomposition methods, $L_{2,1}$-GRMF and $L_{2,1}$-WGRMF. Both methods are used to predict drug-target interactions. We use cross-validation to calculate AUPR values and predict on the drug side (CVd) and the target side (CVt), respectively. We compare them with the most advanced matrix factorization methods currently available. In most cases, our improved methods can provide the best results, which means that the predictive performance is improved with the use of the $L_{2,1}$-norm.

WKNKN preprocessing steps are used to help the experimental results. In addition, it can also be used as an independent method to predict the interactions of drug-target. Considering that the dimensions of the data are relatively small, so the drug-target interactions contained in each dataset are also limited. And our approach applies to these datasets.

In the future, we expect more and more known interaction of drug targets will be found, providing more valuable datasets for our prediction. We will explore more effective prediction methods to solve drug-target interaction problems. For example, we can use matrix factorization of hyper-graph method to improve the reliability of predictive interactions.

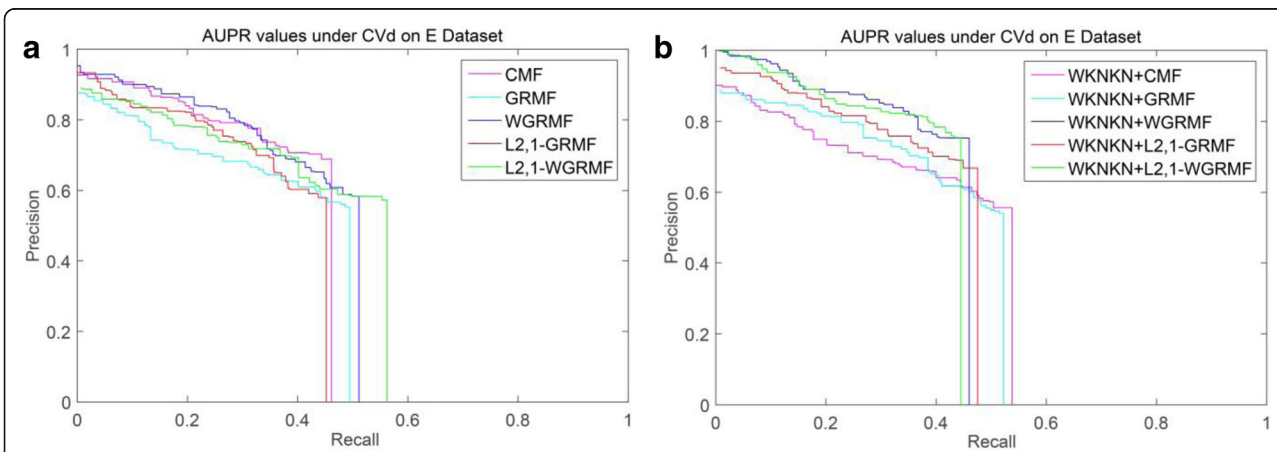

Fig. 6 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVd side of each method on the E dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

Table 3 AUPR Results for Interaction Prediction Under CVt

Methods	NR	GPCR	IC	E
CMF	0.379(0.020)	0.540(0.028)	0.751(0.014)	0.740(0.014)
GRMF	0.423(0.032)	0.567(0.027)	0.745(0.008)	0.763(0.020)
WGRMF	0.423(0.017)	0.574(0.027)	0.801(0.008)	0.801(0.018)
L_{21}-GRMF	0.465(0.056)	0.607(0.020)	0.823(0.012)	0.804(0.021)
$L_{2,1}$-WGRMF	0.425(0.023)	0.603(0.026)	0.801(0.007)	0.802(0.016)
WKNKN+CMF	0.434(0.029)	0.557(0.021)	0.742(0.015)	0.772(0.014)
WKNKN+GRMF	0.500(0.028)	0.615(0.023)	0.815(0.010)	0.807(0.016)
WKNKN+WGRMF	0.446(0.015)	0.585(0.027)	0.799(0.007)	0.798(0.018)
WKNKN+$L_{2,1}$-GRMF	0.519(0.038)	0.617(0.024)	0.826(0.008)	0.799(0.016)
WKNKN+$L_{2,1}$-WGRMF	0.457(0.032)	0.548(0.021)	0.799(0.012)	0.791(0.014)

Methods

CMF

Co-matrix factorization is an effective method to predict the interactions of drug-target [15]. The objective function of CMF method is

$$\min_{\mathbf{A},\mathbf{B}} = \left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_F^2 + \lambda_l\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2\right) + \lambda_d\left\|\mathbf{S}^{\mathbf{d}}-\mathbf{A}\mathbf{A}^{\mathrm{T}}\right\|_F^2 + \lambda_t\left\|\mathbf{S}^{\mathbf{t}}-\mathbf{B}\mathbf{B}^{\mathrm{T}}\right\|_F^2, \quad (1)$$

where $\mathbf{W}$ represents a weight matrix, $W_{ij} = 1$ when Y_{ij} is known, $W_{ij} = 0$ otherwise. Obviously, the last two items of the objective function are regularization terms. We use L to represent the objection function in Eq. (1), a_i represents the i-th vector of $\mathbf{A}$, and b_j represents the j-th vector of $\mathbf{B}$. Two update rules are used to solve $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$. Finally, the two update rules are executed using least square until convergence:

$$\mathbf{A} = \left(\mathbf{Y}\mathbf{B} + \lambda_d\mathbf{S}^{\mathbf{d}}\mathbf{A}\right)\left(\mathbf{B}^{\mathrm{T}}\mathbf{B} + \lambda_l\mathbf{I}_{\mathrm{k}} + \lambda_d\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1}, \quad (2)$$

$$\mathbf{B} = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{A} + \lambda_t\mathbf{S}^{\mathbf{t}}\mathbf{B}\right)\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda_l\mathbf{I}_{\mathrm{k}} + \lambda_t\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}. \quad (3)$$

In summary, after the potential feature matrices $\mathbf{A}$ and $\mathbf{B}$ are updated, the predicted score matrix can be obtained by multiplying $\mathbf{A}$ and $\mathbf{B}$. This predicted score matrix can be used to predict new drug-target interactions by comparing with the original drug-target interactions matrix $\mathbf{Y}$.

GRMF

In the GRMF method, the benefits of regularization items is that it can avoid over-fitting [20]. The objective function of GRMF is as follows:

$$\min_{\mathbf{A},\mathbf{B}} = \left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_F^2 + \lambda_l\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2\right) + \lambda_d\mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{d}}\mathbf{A}\right) + \lambda_t\mathrm{Tr}\left(\mathbf{B}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{t}}\mathbf{B}\right), \quad (4)$$

Then, matrix $\mathbf{A}$ and $\mathbf{B}$ are initialized. The SVD (singular value decomposition) method is used to decompose matrix $\mathbf{Y} \in R^{n\times m}$ into $\mathbf{U} \in R^{n\times k}$, $\mathbf{S}_{\mathbf{k}} \in R^{k\times k}$, and $\mathbf{V} \in R^{n\times k}$. In matrix $\mathbf{Y}$, the largest possible

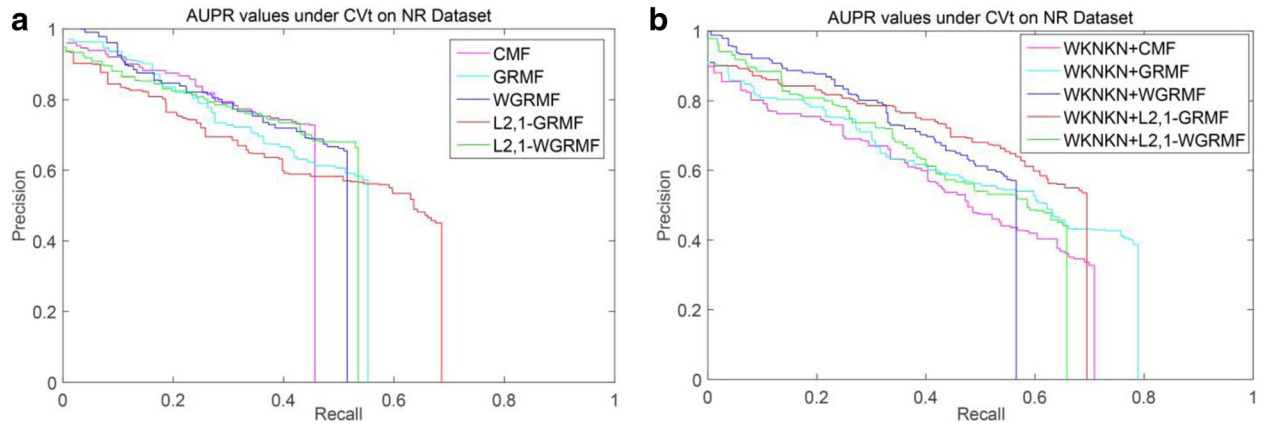

Fig. 7 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVt side of each method on the NR dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

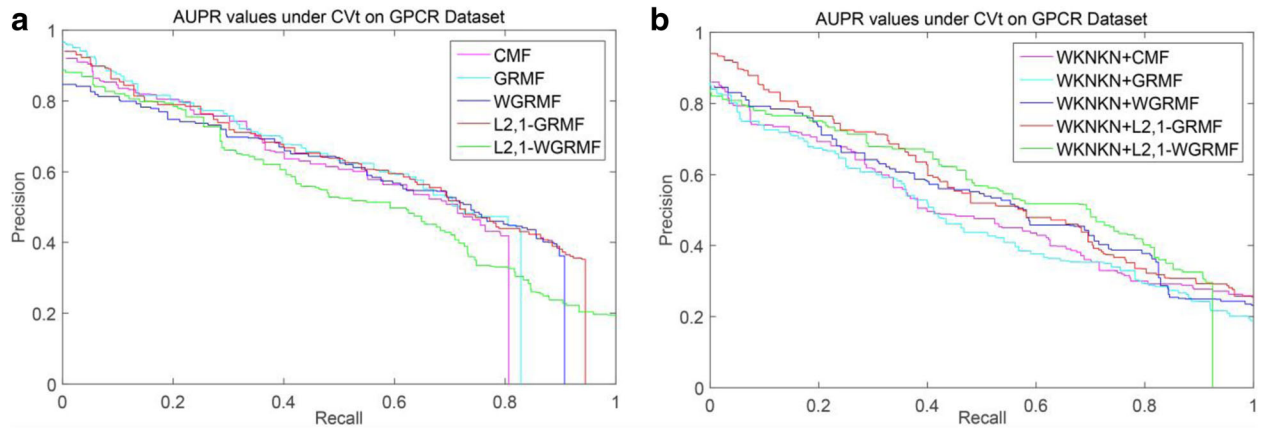

Fig. 8 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVt side of each method on the GPCR dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

number of singular values is $\min(n, m)$, so $k\max = \min(n, m)$. Finally, the square root of $\mathbf{S_k}$ can be obtained, where $\mathbf{A} = \mathbf{US_k^{1/2}}$, $\mathbf{B} = \mathbf{VS_k^{1/2}}$.

Next, the least square method is used to update **A** and **B**. This objective function in Eq. (4) can be replaced by L. These two update rules are used to solve $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$. Finally, the two update rules are executed by using least square until convergence.

WGRMF

Like CMF, the weight matrix **W** in WGRMF is the same as **W** in CMF. Behind the weight matrix, either to prevent unknown interactions, the purpose is to help find the latent feature matrix **A** and **B**. The objective function of WGRMF method is as follows

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}} = &\left\|\mathbf{W}\odot\left(\mathbf{Y}-\mathbf{AB}^{\mathrm{T}}\right)\right\|_F^2 \\ &+\lambda_l\left(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2\right) \\ &+\lambda_d\mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{d}}\mathbf{A}\right)+\lambda_t\mathrm{Tr}\left(\mathbf{B}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{t}}\mathbf{B}\right). \end{aligned} \quad (5)$$

This objective function in Eq. (5) can be replaced by L, where a_i represents the i-th vector of **A**, and b_j represents the j-th vector of **B**. These two update rules are used to solve $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$. Finally, the two update rules are executed by using least square until convergence. However, it is worth noting that the update rules here are not the same as the update rules in GRMF. In GRMF, the rules are matrix updates, but in WGRMF the rules are row updates.

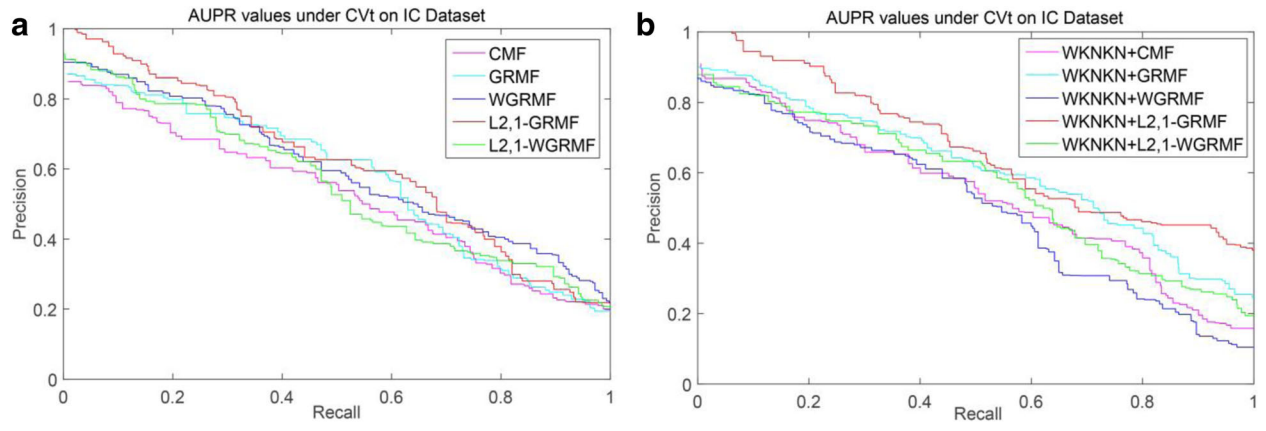

Fig. 9 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVt side of each method on the IC dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

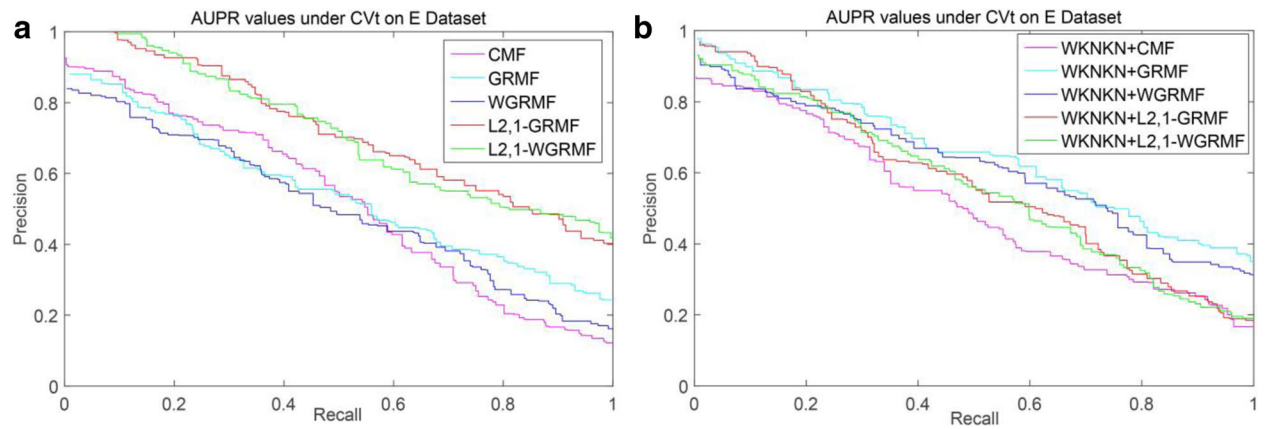

Fig. 10 PR curves for different methods are plotted together, providing a visual comparison between their prediction performances. The PR curves on the CVt side of each method on the E dataset. **a** WKNKN is not used, the PR curves for each method. **b** WKNKN is used, the PR curves for each method

Our proposed methods

Here, our improved approach is used to solve the prediction of drug-target interactions problem. WKNKN (weighted K nearest known neighbors) [20] as a preprocessing step is used to solve unknown missing value problems. Two methods are proposed, Graph Regularization Matrix factorization based on $L_{2,1}$-norm, and a variant called $L_{2,1}$-WGRMF, both of which are used to predict drug-target interactions. Figure 11 shows a flow chart of the proposed method.

$L_{2,1}$-GRMF

Sparsification of the drug similarity matrix and target similarity matrix

Graph regularization terms are used to fully consider the internal structure of the similarity matrix $\mathbf{S^d}$ and $\mathbf{S^t}$. In addition, the graph regularization terms can keep the internal structure of the matrices unchanged. We derive a p-nearest neighbor graph from each drug and target similarity matrix [24] $\mathbf{S^d}$ and $\mathbf{S^t}$ in this work. Therefore, given a drug similarity matrix $\mathbf{S^d}$, a p-nearest neighbor graph [25] N can be generated as

$$\forall_{i,j},\ N_{ij} = \begin{cases} 1, & j \in N_p(i) \quad i \in N_p(j) \\ 0, & j \notin N_p(i) \quad i \notin N_p(j) \\ 0.5, & otherwise, \end{cases} \tag{6}$$

where N is used to sparsify the matrix $\mathbf{S^d}$, which can be written as

$$\forall_{i,j},\ \hat{S} = N_{ij} S^d_{ij}. \tag{7}$$

This result is for a sparse drug similarity matrix. Similarly, the target similarity matrix $\mathbf{S^t}$ can be obtained in the same way. We use the Euclidean distance to calculate the nearest neighbor. In general, Euclidean distance will give better results because it represents the true distance.

Graph regularization helps to facilitate the study the manifold from learning drugs and target spaces. In the original space, there are points that are close to each other, and when the manifold learning is performed, the points are also close to each other in learning.

Low-rank approximation

The idea of low rank approximation (LRA) is applied to GRMF [26]. It decomposes the target matrix $\mathbf{Y}$ into two low-rank latent feature matrices $\mathbf{A}$ and $\mathbf{B}$, i.e., $\mathbf{Y} \approx \mathbf{A}\mathbf{B}^{\mathrm{T}}$ [27]. And the objective function of GRMF can be written as the following optimization problem:

$$\min_{\mathbf{A},\mathbf{B}} = \left\| \mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} \right\|_F^2, \tag{8}$$

where $\|\cdot\|_F$ is Frobenius norm. In addition, the number of potential features of $\mathbf{A}$ and $\mathbf{B}$ is represented by k.

Table 4 Predicted Drugs for estrogen receptor alpha, NR Dataset

Rank	Drug	Drug ID
1	Progesterone	D00066
2	Estrone	D00067
3	Ethinylestradiol	D00554
4	Etodolac	D00315
5	Ethynodiol diacetate	D01294
6	Testosterone	D00075
7	Budesonide	D00246
8	Isotretinoin	D00348
9	Mometasone furoate	D00690
10	Paricalcitol	D00930

Table 5 Predicted Targets for Diazoxide, IC Dataset

Rank	Target	Target ID
1	potassium voltage-gated channel subfamily J member 16	hsa3773
2	potassium voltage-gated channel subfamily A member regulatory beta subunit 1	hsa7881
3	potassium voltage-gated channel subfamily J member 15	hsa3772
4	potassium voltage-gated channel modifier subfamily S member 2	hsa3788
5	potassium voltage-gated channel subfamily H member 5	hsa27133
6	potassium voltage-gated channel subfamily D member 1	hsa3750
7	glutamate ionotropic receptor AMPA type subunit 1	hsa2890
8	potassium voltage-gated channel subfamily D member 3	hsa3752
9	potassium calcium-activated channel subfamily N member 4	hsa3783
10	potassium voltage-gated channel subfamily H member 1	hsa3756
11	potassium calcium-activated channel subfamily N member 3	hsa3782
12	potassium voltage-gated channel subfamily D member 2	hsa3751
13	chloride voltage-gated channel 2	hsa1181
14	calcium voltage-gated channel auxiliary subunit beta 4	hsa785
15	sodium channel epithelial 1 gamma subunit	hsa6340
16	ryanodine receptor 3	hsa6263
17	cholinergic receptor nicotinic delta subunit	hsa1144
18	solute carrier family 6 member 4	hsa6532
19	sodium voltage-gated channel alpha subunit 3	hsa6328
20	sodium voltage-gated channel alpha subunit 9	hsa6335

Regularization

In general, the Tikhonov and graph regularization terms can be used to avoid over-fitting and enhance generalization capability. Here is the objective function of $L_{2,1}$-GRMF:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}} = &\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_F^2+\lambda_l\left(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2\right)\\&+\lambda_l\|\mathbf{B}\|_{2,1}+\lambda_d\sum_{i,r=1}^{n}\widehat{S_{ir}^d}\|a_i-a_r\|^2\\&+\lambda_t\sum_{j,q=1}^{m}\widehat{S_{jq}^t}\left\|b_j-b_q\right\|^2,\end{aligned}\tag{9}$$

where λ_l, λ_d and λ_t are positive parameters, a_i is the i-th rows of $\mathbf{A}$, and b_j is the j-th rows of $\mathbf{B}$, n is the number of drugs, and m is the number of targets. The first term is an approximate model of the matrix $\mathbf{Y}$. The second term is the Tikhonov regularization. Its main purpose is to minimize the norms of $\mathbf{A}$, $\mathbf{B}$. The third term is the $L_{2,1}$-norm applied on $\mathbf{B}$ to increase the target matrix sparsity and discard unwanted target pairs. Considering that we are more concerned with certain drugs, we use the $L_{2,1}$-norm to sparse the potential feature matrix of the target, so that we can better predict new drugs. However, while the $L_{2,1}$-norm is added to $\mathbf{A}$, some of the more important drugs may be lost. The last two terms are graph regularization of drugs and targets, respectively. Moreover, the drug-target model can be re-written as:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}} = &\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_F^2+\lambda_l\left(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2\right)\\&+\lambda_l\|\mathbf{B}\|_{2,1}+\lambda_d\mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{L_d}\mathbf{A}\right)\\&+\lambda_t\mathrm{Tr}\left(\mathbf{B}^{\mathrm{T}}\mathbf{L_t}\mathbf{B}\right),\end{aligned}\tag{10}$$

where $\mathrm{Tr}(\cdot)$ is the trace of the matrix, $\mathbf{L_d}=\mathbf{D^d}-\widehat{\mathbf{S}^{\mathbf{d}}}$ is the graph Laplacian for $\widehat{\mathbf{S}^{\mathbf{d}}}$, $\mathbf{L_t}=\mathbf{D^t}-\widehat{\mathbf{S}^{\mathbf{t}}}$ is the graph Laplacian for $\widehat{\mathbf{S}^{\mathbf{t}}}$. Please refer to [28] for more details on rewriting graph regularization. We know that the known normalized Laplacian is better than unknown, so we replace $\mathbf{L_d}$ and $\mathbf{L_t}$ with $\tilde{\mathbf{L}_{\mathbf{d}}}=(\mathbf{D^d})^{-1/2}\mathbf{L_d}(\mathbf{D^d})^{-1/2}$ and $\tilde{\mathbf{L}_{\mathbf{t}}}=(\mathbf{D^t})^{-1/2}\mathbf{L_t}(\mathbf{D^t})^{-1/2}$. The function can be written as:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}} = &\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}\right\|_F^2+\lambda_l\left(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2\right)\\&+\lambda_l\|\mathbf{B}\|_{2,1}+\lambda_d\mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{d}}\mathbf{A}\right)\\&+\lambda_t\mathrm{Tr}\left(\mathbf{B}^{\mathrm{T}}\tilde{\mathbf{L}}_{\mathbf{t}}\mathbf{B}\right).\end{aligned}\tag{11}$$

We use the minimization of the objective function to predict the outcome of the interactions, but this could lead to unsatisfactory results. Because there are many zeros that have not been found. Therefore, we use WKNKN pre-processing method to solve this problem.

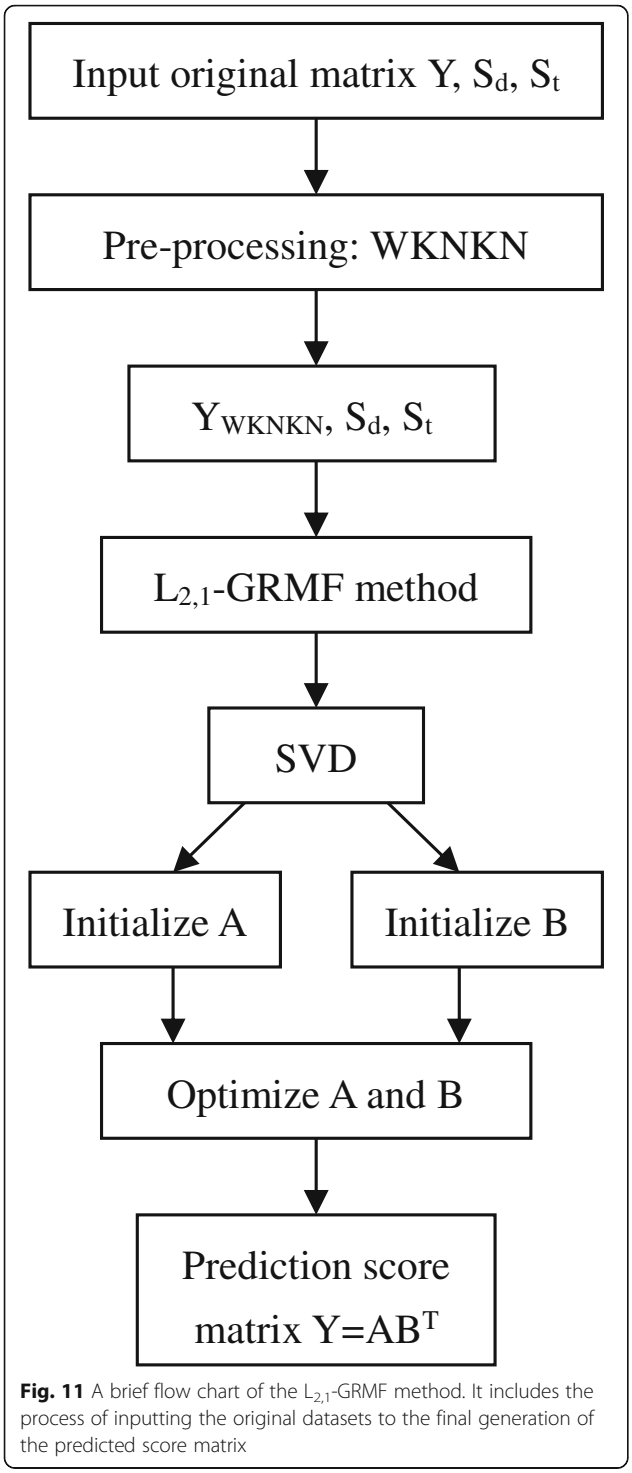

Fig. 11 A brief flow chart of the $L_{2,1}$-GRMF method. It includes the process of inputting the original datasets to the final generation of the predicted score matrix

Initialization of A and B

For the input matrix **Y**, SVD (Singular Value Decomposition) method is used to obtain the initial value of matrix **A** and matrix **B**:

$$[\mathbf{U},\mathbf{S},\mathbf{V}] = SVD(\mathbf{Y},k), \mathbf{A} = \mathbf{U}\mathbf{S}_\mathbf{k}^{\mathbf{1/2}}, \mathbf{B} = \mathbf{V}\mathbf{S}_\mathbf{k}^{\mathbf{1/2}}. \quad (12)$$

Among them, $\mathbf{S_k}$ is a diagonal matrix and contains the k largest singular values. In matrix **Y**, the number of singular values is $k_{\max} = \min(n, m)$. According to the SVD method, $k_{\max}$ is the maximum possible number.

Optimization algorithm

In this paper, we can update **A** and **B** by using the least square method. Let the partial derivative of **A** be equal to 0, the partial derivative of **B** be equal to 0, the objective function in Eq. (11) can be replaced by L, that is, $\partial L/\partial \mathbf{A} = 0$ and $\partial L/\partial \mathbf{B} = 0$. The two update rules are executed by using least square until convergence. When we perform the $L_{2,1}$-GRMF method, λ_l, λ_d and λ_t are determined by the cross-validation on the training set to the optimal parameter values. We use grid search, $\lambda_l \in \{2^{-2}, 2^{-1}, 2^0, 2^1\}$. Then we choose the optimal parameters from this set. Derivation process is as follows:

$$\mathbf{A} = \left(\mathbf{Y}\mathbf{B} - \lambda_d \tilde{\mathbf{L}}_\mathbf{d}\mathbf{A}\right)\left(\mathbf{B}^\mathrm{T}\mathbf{B} + \lambda_l \mathbf{I}_\mathbf{k}\right)^{-1}, \quad (13)$$

$$\mathbf{B} = \left(\mathbf{Y}^\mathrm{T}\mathbf{A} - \lambda_t \tilde{\mathbf{L}}_\mathbf{t}\mathbf{B}\right)\left(\mathbf{A}^\mathrm{T}\mathbf{A} + \lambda_l \mathbf{I}_\mathbf{k} + \lambda_l \mathbf{D}\mathbf{I}_\mathbf{k}\right)^{-1}, \quad (14)$$

where **D** is a diagonal matrix with the i-th diagonal element as $d_{ii} = 1/2\|(\mathbf{B})^i\|_2$. The specific algorithm of $L_{2,1}$-GRMF is as follows:

Algorithm 1: $L_{2,1}$-GRMF

Data Input: $\mathbf{Y} \in R^{n \times m}$

Output: $\hat{\mathbf{Y}}$

Parameters: $K, \eta, k, \lambda_l, \lambda_d, \lambda_t$

Pre-processing: $\mathbf{Y} = WKNKN(\mathbf{Y}, \mathbf{S^d}, \mathbf{S_g^t}, K, \eta), [\mathbf{U}, \mathbf{S}, \mathbf{V}] = SVD(\mathbf{Y}, k)$

Initialization: $\mathbf{A} = \mathbf{U}\mathbf{S}_\mathbf{k}^{\mathbf{1/2}}, \mathbf{B} = \mathbf{V}\mathbf{S}_\mathbf{k}^{\mathbf{1/2}}$

Repeat

Update **A** using Eq.(13).

Update **B** using Eq.(14).

Until convergence

$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{B}^\mathrm{T}$

Return **Y**

$L_{2,1}$-WGRMF

A variant of $L_{2,1}$-GRMF, called $L_{2,1}$-WGRMF, is obtained here by adding a weight matrix **W** to the $L_{2,1}$-GRMF. The advantage is that it helps to determine the latent feature matrices **A** and **B** of the drug-target matrix **Y**. So, we write the objective function that contains **W** as follows:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} = & \left\|\mathbf{W} \odot \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\mathrm{T}\right)\right\|_F^2 \\ & + \lambda_l\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2\right) + \lambda_l\|\mathbf{B}\|_{2,1} \\ & + \lambda_d \mathrm{Tr}\left(\mathbf{A}^\mathrm{T}\mathbf{L}_\mathbf{d}\mathbf{A}\right) + \lambda_t \mathrm{Tr}\left(\mathbf{B}^\mathrm{T}\mathbf{L}_\mathbf{t}\mathbf{B}\right). \end{aligned} \quad (15)$$

Let objective function be set to F such that $\partial F/\partial a_i = 0$ and $\partial F/\partial b_j = 0$. The update rules are used to obtain **A** and **B** until convergence

$$\forall_i = 1 \ldots n,$$
$$a_i = \left(\sum_{j=1}^{m} W_{ij} Y_{ij} b_j - \lambda_d \left(\tilde{\mathbf{L}}_{\mathbf{d}} \right)_{i*} \mathbf{A} \right) \left(\sum_{j=1}^{m} W_{ij} b_j^{\mathrm{T}} b_j + \lambda_l \mathbf{I}_{\mathbf{k}} \right)^{-1}, \tag{16}$$

$$\forall_j = 1 \ldots m,$$
$$b_j = \left(\sum_{i=1}^{n} W_{ij} Y_{ij} a_i - \lambda_t \left(\tilde{\mathbf{L}}_t \right)_{j*} \mathbf{B} \right) \left(\sum_{i=1}^{n} W_{ij} a_i^{\mathrm{T}} a_i + \lambda_l \mathbf{I}_{\mathbf{k}} + \lambda_l \mathbf{D} \mathbf{I}_{\mathbf{k}} \right)^{-1}. \tag{17}$$

Abbreviations

AUPR: Area under the precision-recall curve; CMF: Collaborative matrix factorization method; CV: Cross-validation; GRMF: Graph regularized matrix factorization; $L_{2,1}$-GRMF: $L_{2,1}$-norm Graph regularized matrix factorization; $L_{2,1}$-WGRMF: Weighted $L_{2,1}$-norm graph regularized matrix factorization; LRA: Low rank approximation; SVD: Singular value decomposition; WGRMF: Weighted graph regularized matrix factorization; WKNKN: Weighted K nearest known neighbors

Acknowledgements

Not applicable.

Funding

Publication consts are founded by the National Natural Science Foundation of China under grant Nos. 61872220, 61572284, and 61701279.

Availability of data and materials

The datasets that support the findings of this study are available in https://github.com/cuizhensdws/L21-GRMF.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 8, 2019: Decipher computational analytics in digital health and precision medicine*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-8.

Authors' contributions

ZC and YLG jointly contributed to the design of the study. ZC designed and implemented the $L_{2,1}$-GRMF and $L_{2,1}$-WGRMF method, performed the experiments, and drafted the manuscript. JXL gave statistical and computational advice to the project, and participated in designing evaluation criteria. LYD and SSY contributed to the data analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Information Science and Engineering, Qufu Normal University, Rizhao, China. [2]Library of Qufu Normal University, Qufu Normal University, Rizhao, China. [3]Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei, China.

Published: 10 June 2019

References

1. Novac N. Challenges and opportunities of drug repositioning. Trends Pharmacol Sci. 2013;34(5):267–72.
2. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther. 2013; 93(4):335–41.
3. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(Database issue):D353–61.
4. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011;39(Database issue:D1035.
5. Kuhn M, Szklarczyk D, Pletscherfrankild S, Blicher TH, Mering CV, Jensen LJ, Bork P. STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res. 2014;42(Database issue):401–7.
6. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, Mcglinchey S, Michalovich D, Allazikani B. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(Database issue):1100–7.
7. Yonan AL, Palmer AA, Smith KC, Feldman I, Lee HK, Yonan JM, Fischer SG, Pavlidis P, Gilliam TC. Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. Genes Brain Behav. 2003;2(5):303–20.
8. Klipp E, Wade RC, Kummer U. Biochemical network-based drug-target prediction. Curr Opin Biotechnol. 2010;21(4):511–6.
9. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007; 25(2):197–206.
10. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol. 2007;25(1):71–5.
11. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–40.
12. Shang J, Sun Y, Li S, Liu JX, Zheng CH, Zhang J. An improved opposition-based learning particle swarm optimization for the detection of SNP-SNP interactions. Biomed Res Int. 2015;2015:524821.
13. Wei PJ, Zhang D, Xia J, Zheng CH. LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network. Bmc Bioinformatics. 2016;17(Suppl 17):467.
14. Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics. 2012;28(18):2304–10.
15. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2013. p. 1025–33.
16. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J. Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45.
17. Ge SG, Xia J, Sha W, Zheng CH. Cancer subtype discovery based on integrative model of multigenomic data. IEEE/ACM Transactions on Computational Biology & Bioinformatics. 2017;14(5):1115–21.
18. Wang DQ, Zheng CH, Gao YL, Liu JX, Wu SS, Shang JL. L21-iPaD: an efficient method for drug-pathway association pairs inference. In: IEEE international conference on bioinformatics and biomedicine; 2017. p. 664–9.
19. Takahashi Y, Fujishima S, Kato H. Chemical data mining based on structural similarity. Journal of Computer Chemistry Japan. 2003;2(4):119–26.
20. Ezzat A, Zhao P, Wu M, Li X-L, Kwoh C-K. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2017;14(3):646-56.
21. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: ICML '06 : proceedings of the international conference on machine learning, New York, Ny, Usa; 2006. p. 233–40.
22. Li J, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. J R Stat Soc. 2010;59(4):673.
23. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T. Toward more realistic drug–target interaction predictions. Brief Bioinform. 2015;16(2):325–37.
24. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci. 2003;43(2):391.

25. Wang B, Pan F, Hu KM, Paul JC. Manifold-ranking based retrieval using k -regular nearest neighbor graph. Pattern Recogn. 2012;45(4):1569–77.
26. Liberty E, Woolfe F, Martinsson PG, Rokhlin V, Tygert M. Randomized algorithms for the low-rank approximation of matrices. Proc Natl Acad Sci U S A. 2007;104(51):20167–72.
27. Wang J, Liu J-X, Zheng C-H, Wang Y-X, Kong X-Z, Wen C-G. A Mixed-Norm Laplacian Regularized Low-Rank Representation Method for Tumor Samples Clustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2019;16(1):172-82.
28. Gu Q, Zhou J, Ding CHQ. Collaborative filtering: weighted nonnegative matrix factorization incorporating user and item graphs. SDM:199-210. In: Siam international conference on data mining, SDM 2010, April 29–may 1, 2010, Columbus, Ohio, Usa; 2010. p. 199–210.