2021

De Novo Prediction of Drug–Target Interactions Using Laplacian Regularized Schatten p-Norm Minimization

Gaoyan Wu

Mengyun Yang

Yaohang Li
Old Dominion University

Jianxin Wang

Original Publication Citation

Wu, G., Yang, M., Li, Y., & Wang, J. (2021). De novo prediction of drug–target interactions using Laplacian regularized Schatten p-norm minimization. *Journal of Computational Biology, 28*(7), 660-673. https://doi.org/10.1089/cmb.2020.0538

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 0, 2021

Pp. 1–14
DOI: 10.1089/cmb.2020.0538

ISBRA 2020

De Novo Prediction of Drug–Target Interactions Using Laplacian Regularized Schatten *p*-Norm Minimization

GAOYAN WU,[1] MENGYUN YANG,[2] YAOHANG LI,[3] and JIANXIN WANG[1]

ABSTRACT

In pharmaceutical sciences, a crucial step of the drug discovery is the identification of drug–target interactions (DTIs). However, only a small portion of the DTIs have been experimentally validated. Moreover, it is an extremely laborious, expensive, and time-consuming procedure to capture new interactions between drugs and targets through traditional biochemical experiments. Therefore, designing computational methods for predicting potential interactions to guide the experimental verification is of practical significance, especially for de novo situation. In this article, we propose a new algorithm, namely Laplacian regularized Schatten *p*-norm minimization (LRSpNM), to predict potential target proteins for novel drugs and potential drugs for new targets where there are no known interactions. Specifically, we first take advantage of the drug and target similarity information to dynamically prefill the partial unknown interactions. Then based on the assumption that the interaction matrix is low-rank, we use Schatten *p*-norm minimization model combined with Laplacian regularization terms to improve prediction performance in the new drug/target cases. Finally, we numerically solve the LRSpNM model by an efficient alternating direction method of multipliers algorithm. We evaluate LRSpNM on five data sets and an extensive set of numerical experiments show that LRSpNM achieves better and more robust performance than five state-of-the-art DTIs prediction algorithms. In addition, we conduct two case studies for new drug and new target prediction, which illustrates that LRSpNM can successfully predict most of the experimental validated DTIs.

Keywords: drug–target interactions prediction, Laplacian regularization, matrix completion, Schatten *p*-norm minimization.

1. INTRODUCTION

DRUG DISCOVERY IS A VERY DIFFICULT PROCESS. How to effectively discover new candidate medications has been widely studied by academic researchers. With the development of high-throughput technology, the efficacy of drugs can be simulated in advance. Experts can analyze the chemical molecules of drugs

[1]The Hunan Provincial Key Lab of Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China.
[2]School of Science, Shaoyang University, Shaoyang, China.
[3]Department of Computer Science, Old Dominion University, Norfolk, Virginia, USA.

from the molecular level, and further study whether these molecules can act on the useful target proteins of human, so as to estimate the therapeutic effect of drugs.

Although drug discovery has made great progress in recent years, it is well known to be costly and time consuming. Literally, Eroom's Law (Scannell et al., 2012) indicates that the pharmaceutical sector invests 50 billion dollars annually in research for new medicines, but the number of new drugs approved per billion U.S. dollars spent has halved roughly every 9 years since 1950, falling $\sim$80-fold in inflation-adjusted terms. Furthermore, the number of truly innovative drugs approved by regulatory agencies has declined in recent years, despite the advances in biotechnology. It is reported that the Food and Drug Administration (FDA) of the United States spends twice as much money on a new drug on the market every 9 years, but only about 20 novel drugs are on the market per year with high investment costs (Chen and Zhang, 2013). It is very necessary to develop more efficient methods to reduce the cost of time and expense.

In the drug discovery process, the prediction of drug–target interactions (DTIs) is an important step that aims to identify potential new drugs or new targets for existing drugs. Meanwhile, the study of DTIs has various applications, mainly including screening drug candidates that target-specific disease-associated genes/proteins (Oprea and Mestres, 2012), drug repositioning (Li et al., 2016), drug toxic side effect prediction (Lounkine et al., 2012), and understanding the mechanism of drug operation and disease pathology (Núñez et al., 2012). Knowledge of the associations between drugs and their targets is essential for a wide range of pharmaceutical and bioinformatics studies. According to the statistics, there are >90 million chemical molecules in PubChem (Kim et al., 2016) and DrugBank (Wishart et al., 2008), and >100,000 human target proteins in UniProt (Apweiler et al., 2004). However, only a small partial known DTIs have been verified by biological experiments and much more still remain to be discovered. Therefore, identifying more DTIs is an extremely valuable task, which can bring huge breakthrough in biopharmaceutical and biomedical research.

In recent years, many computational approaches have been developed to infer novel DTIs under the advantage of lower cost and wider coverage. Bleakley and Yamanishi (2009) first proposed a Bipartite Local Model (BLM) to predict target proteins of a given drug, then to predict drugs targeting a given protein. BLM used the chemical structure similarity of drugs and the sequence similarity of targets to improve the prediction accuracy. Analogously, Laplacian regularized least squares (LapRLS) (Xia et al., 2010) is another algorithm based on the BLM. LapRLS used regularized least squares to minimize an objective function that includes an error term as well as a graph regularization term. To perform prediction, Van Laarhoven and Marchiori (2013) utilized a weighted nearest neighbor (WNN) procedure for inferring a profile of a drug by using interaction profiles of the compounds. The experimental results have shown that neighbors' information is indeed beneficial to the prediction results. In addition, Mizutani et al. (2012) made use of protein functions and drugs' side effects to identify novel targets for the known anticancer drugs by sparse canonical correlation analysis, where drugs and targets were represented from different views.

It is worth noting that matrix factorization and completion methods have exhibited excellent performance for DTI prediction among these methods. Kernelized Bayesian matrix factorization with twin kernels (KBMF2K) (Gönen, 2012) applied a Bayesian probabilistic matrix factorization to perform prediction. KBMF2K defined two kernel matrices based on chemical similarity between drugs and genomic similarity between targets. Besides, KBMF2K used variational approximation to perform nonlinear dimensionality reduction, which can improve the computational efficiency in terms of computation time. collaborative matrix factorization (CMF) (Zheng et al., 2013) employed collaborative filtering for DTI prediction. This approach transforms the input DTI matrix into the inner product between drug features and target features, which are also derived from similarity data. Liu et al. (2016) proposed the neighborhood regularized logistic matrix factorization (NRLMF). NRLMF focused on the probability of DTI using logistic matrix decomposition, in which the features of drug and target are represented by drug-specific and target-specific potential carriers, respectively. The Neighborhood Constraint Matrix Completion (NCMC) method (Fan et al., 2018) applied the similar information of drugs/targets to define the concept of neighborhood. NCMC method combined nuclear norm minimization model with neighborhood constraints to deal with the sparsity of known interactions, which captured the strong correlation between drug and target.

Although these computational methods have been achieved excellent performance for predicting DTIs, it is a challenging task to identify interactions for new drugs or new targets, which is known as de novo prediction. To solve the cold-start problem where drugs or targets have no given interactions in de novo cases, the side information of drugs and targets can be taken advantage to achieve further improvement. To enhance the prediction accuracy in de novo tests, some existing methods have provided insights to

improve the prediction performance for new drugs or targets. However, the results show that there is still room for improvement. It is necessary to develop more effective computational methods to predict potential DTIs.

In this article, we propose a novel matrix completion approach, namely Laplacian regularized Schatten p-norm minimization (LRSpNM) for de novo prediction of DTIs. Based on the assumption that similar drugs are normally interacted with similar targets and similar targets tend to bind with similar drugs, the DTIs matrix can be assumed to be of low rank. Accordingly, matrix completion algorithms, which efficiently construct low-rank matrix approximations consistent with known interactions, can provide tremendous help in discovering the novel DTIs. In our method, we use Schatten p-norm to approximate the matrix rank and combine the Laplacian regularized term to assist prediction. In addition, considering that many of the interactions in the DTIs matrix are unknown cases, we use a prefilling step to enhance prediction. The performances of LRSpNM are empirically evaluated on five benchmark data sets, compared with five state-of-the-art DTI prediction methods. Extensive computational results demonstrate that LRSpNM usually outperforms other competing methods on all data sets under two de novo experimental settings. The code of LRSpNM is freely available at https://github.com/BioinformaticsCSU/LRSpNM

2. MATERIALS

Evaluation experiments are performed using a benchmark data set (Yamanishi et al., 2008) and a larger data set arranged by Wang and Kurgan (2019). Specifically, the former data set, which is generally used in DTIs prediction, consists of four different sub-data sets targeting protein of enzyme, ion channel, G protein-coupled receptor (GPCR), and nuclear receptor. The four sub-data sets are publicly available at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget The latter data set, remarked as WANG, focuses on the human protein targets, which can be downloaded from the website http://biomine.cs.vcu.edu/servers/CONNECTOR Each data set includes three matrices: an interaction matrix $A \in R^{m\times n}$ between m drugs and n targets, a similarity matrix of drugs $S_d \in R^{m\times m}$, and a similarity matrix of targets $S_t \in R^{n\times n}$. The statistical information of DTI matrix in each data set is summarized in Table 1.

The matrix A is the adjacency matrix encoding the DTIs, where A_{ij} is 1 if drug d_i and target t_j are known to interact and 0 otherwise. The drug similarity S_d is computed from the chemical structures of drugs by using SIMCOMP (Hattori et al., 2003), which defines the drug similarity between two drugs d_i and d_j as follows:

$$S_d(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}, \tag{1}$$

where $|d_i \cap d_j|$ is the number of all substructures shared by d_i and d_j, $|d_i \cup d_j|$ is the number of all substructures that either d_i or d_j has. The target similarity S_t is computed according to target sequences by using a normalized Smith–Waterman score (Smith and Waterman, 1981) of target t_i and t_j as follows:

$$S_t(t_i, t_j) = \frac{SW(t_i, t_j)}{\sqrt{SW(t_i, t_i)}\sqrt{SW(t_j, t_j)}}, \tag{2}$$

where $SW(,)$ is the Smith–Waterman score.

Table 1. Statistic of Drug–Target Interactions Data Sets

Data sets	*No. of drugs*	*No. of targets*	*No. of interactions*	*Sparsity*
Enzyme	445	664	2926	0.010
Ion channel	210	204	1476	0.034
GPCR	223	95	635	0.030
Nuclear receptor	54	26	90	0.064
WANG	449	1469	34,456	0.052

GPCR, G protein-coupled receptor.

3. METHODS

To predict DTIs that remain undiscovered, we propose a novel method called LRSpNM, which mainly consists of three steps. First, a preprocessing step is performed to infer partial unknown interaction probability values based on the K nearest neighbor profiles. Second, the Laplacian matrices for drug and target are calculated based on the original similarities matrices. Finally, the framework of LRSpNM is used to infer the potential interactions. The workflow of LRSpNM for predicting potential DTIs is shown in Figure 1.

3.1. Preprocessing step

When an interaction matrix is constructed with drugs as rows, targets as columns, and known DTIs valued 1, unknown valued 0, the DTIs prediction problem can then be modeled as a matrix completion problem by completing the unknown elements with pharmacological space information in the interaction matrix.

In this study, the known DTI matrix A has m drug rows and n target columns. The ith row in A is the interaction profile for drug d_i. Similarly, the jth column in A is the interaction profile for target t_j. A drug or target being known means that it has at least one interaction in its profile, whereas it being new means that it has no interactions in its profile.

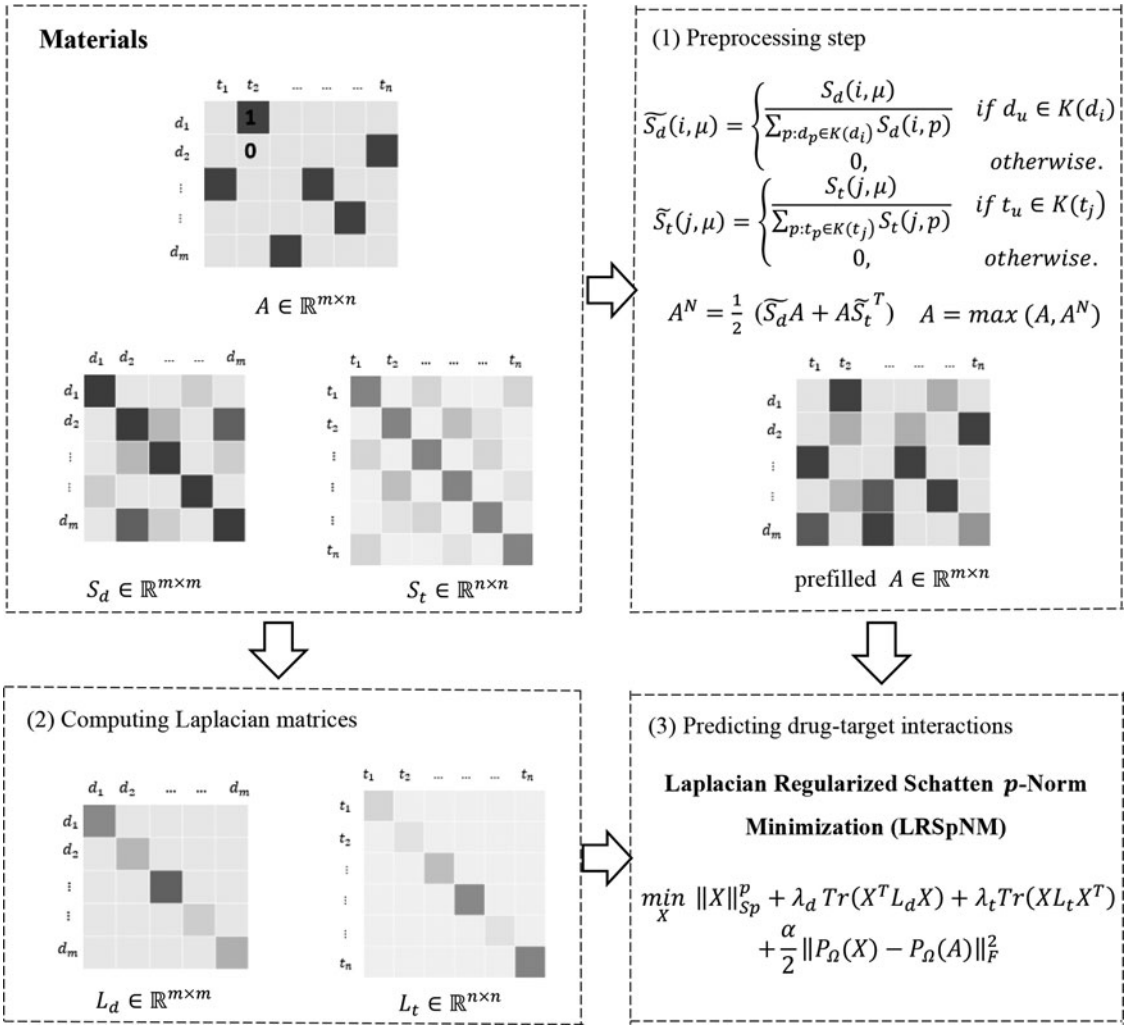

FIG. 1. Overall workflow of LRSpNM for discovering potential drug–target interactions. (1) Preprocessing step. (2) Computing Laplacian matrices. (3) Predicting drug–target interactions. LRSpNM, Laplacian regularized Schatten p-norm minimization.

Many of the noninteractions in A are unknown cases that may potentially be positive interactions. Previous studies (Keiser et al., 2007; Jacob and Vert, 2008) show that the interaction probability between drug d_i and target t_j should be close to the interaction probabilities between d_i's neighbors and t_j's neighbors. Hence, we use a preprocessing step that utilizes the similarity information between drugs and targets to estimate the interaction likelihoods for unknown interactions in A.

First, for drug d_i, we select the K most similar drugs as its neighbors based on drug similarity and use $K(d_i)$ to denote the set of them. We use an adjacency matrix $\tilde{S}_d$ to represent the drug neighborhood information, which is defined as follows:

$$\tilde{S}_d(d_i, d_\mu) = \begin{cases} \frac{S_d(d_i, d_\mu)}{\sum_{p:d_p \in K(d_i)} S_d(d_i, d_p)} & if\ d_\mu \in K(d_i) \\ 0, & otherwise, \end{cases} \tag{3}$$

where $S_d(d_i, d_\mu)$ is the original similarity score between d_i and d_μ.

Similarly, we use $K(t_j)$ to represent the set of t_j's neighbors, and calculate the adjacency matrix $\tilde{S}_t$ in the same way, which is defined as

$$\tilde{S}_t(t_j, t_\mu) = \begin{cases} \frac{S_t(t_j, t_\mu)}{\sum_{p:t_p \in K(t_j)} S_t(t_j, t_p)} & if\ t_\mu \in K(t_j) \\ 0, & otherwise, \end{cases} \tag{4}$$

Based on the K nearest known neighbors' information from drugs and targets, we can obtain the DTIs likelihoods for partial unknown pairs, which is marked as A^N and calculated in the following equation:

$$A^N = \frac{\tilde{S}_d A + A \tilde{S}_t}{2}. \tag{5}$$

Finally, we combine the prefilled interaction probabilities with known interactions as the input matrix to be completed, expressed as

$$A = \max(A, A^N). \tag{6}$$

Note that to infer the interaction likelihood for drug–target pairs, the processing step uses the K nearest known neighbors, which is reasonable since known neighbors carry more additional interaction information.

3.2. Schatten p-norm minimization

Given partial and possibly noisy observations on some entries, our aim is to find a low-rank matrix that recovers all unknown entries, which is called matrix completion in the literature (Candes and Recht, 2009). The general approach is to find a matrix with the minimum rank under certain conditions from the observations. Specifically, the general rank minimization problem to fill out the missing entries is formulated as the following formula:

$$\begin{aligned} &\min_X rank(X) \\ &s.t.\ \ P_\Omega(X) = P_\Omega(A), \end{aligned} \tag{7}$$

where A is the prefilled interaction matrix, $X \in R^{m \times n}$ is the variable matrix, $rank(X)$ denotes the rank function of X, Ω is a set containing index pairs of all known entries in A and P_Ω is the projection operator onto Ω, which is defined as

$$(P_\Omega(X))_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega \\ 0, & (i,j) \notin \Omega. \end{cases} \tag{8}$$

Based on the assumption that similar drugs share the similar molecular pathways to interact with similar targets, the matrix A is inherently low rank. Thus, the DTIs prediction problem can be modeled as a matrix completion problem, which predicts the unknown DTIs by completing the elements in the interaction matrix. Hence, we can use matrix completion algorithms to predict unknown DTIs.

Unfortunately, the rank minimization problem [Eq. (7)] is known to be NP-hard. One of the solutions is that turns the rank function to a more tractable solution by minimizing the nuclear norm, which has been proven to be the convex relaxation of matrix rank (Fazel, 2002). Although the nuclear norm minimization

model is a convex problem with a global solution, the relaxation may deviate from the solution of the original problem. Meanwhile, most completion methods minimize the squared prediction errors on the observed entries, which is sensitive to outliers. It is desired to solve a better approximation of the rank minimization problem without introducing much computational cost. To solve the problem, Nie et al. (2012, 2015) proposed nonconvex optimization models where the Schatten p-norm of a matrix X is used to replace the rank function of Equation (7), which is defined as

$$\|X\|_{S_p}^p = \sum_{i=1}^{\min\{n,m\}} \sigma_i^p = Tr\left(\left(X^T X\right)^{\frac{p}{2}}\right), \tag{9}$$

where σ_i is the ith singular value of X and when $p=1$, the Schatten 1-norm is the well-known nuclear norm. That is to say, the nuclear norm is the special case of Schatten p-norm. As a result, the baseline of Schatten p-norm minimization is formulated as

$$\min_X \|X\|_{S_p}^p + \frac{\alpha}{2}\|P_\Omega(X) - P_\Omega(A)\|_F^2, \tag{10}$$

where α is the harmonic parameter that balances the Schatten p-norm and the error term, we optimize the effectiveness of matrix completion by fine-tuning the value of p.

3.3. Laplacian regularized Schatten p-norm minimization

In this section, we first introduce Schatten p-norm to approximate the rank of the interaction matrix, and then present a new objective function through incorporation of the drug–drug similarity and target–target similarity into the matrix completion framework for DTI prediction. We use a Laplacian regularized term to constrain that drugs with similar chemical structure are more likely to have connections with similar targets. Similarly, targets with similar genomic sequence similarity are more likely to have interactions with similar drugs. Specially, a LRSpNM model is proposed for DTI prediction. The optimization problem of LRSpNM can be formulated as follows:

$$\min_X \|X\|_{S_p}^p + \lambda_d Tr\left(X^T L_d X\right) + \lambda_t Tr\left(X L_t X^T\right) + \frac{\alpha}{2}\|P_\Omega(X) - P_\Omega(A)\|_F^2, \tag{11}$$

where $L_d \in R^{m\times m}$ is the drug Laplacian matrix with $L_d = D_d - S_d$, D_d is the diagonal matrix with $D_d(i,i) = \sum S_d(i,i)$, $L_t \in R^{n\times n}$ is the target Laplacian matrix with $L_t = D_t - S_t$, D_t is the diagonal matrix with $D_t(i,i) = \sum S_t(i,i)$ and λ_d, λ_t are parameters balancing the reconstruction terms of LRSpNM model.

To solve the optimization problem in Equation (11), we use the alternating direction method of multipliers (ADMM) (Chen et al., 2012) framework and introduce two auxiliary variables W and Z to make the objective function separable:

$$\begin{gathered}\min_X \|W\|_{S_p}^p + \lambda_d Tr(X^T L_d X) + \lambda_t Tr(X L_t X^T) + \tfrac{\alpha}{2}\|P_\Omega(Z) - P_\Omega(A)\|_F^2 \\ s.t.\ \ X = W,\ \ X = Z.\end{gathered} \tag{12}$$

The corresponding augmented Lagrange function of Equation (12) is:

$$\begin{aligned}\mathcal{L}(W,Z,X,U,V) &= \|W\|_{S_p}^p + \frac{\alpha}{2}\|P_\Omega(Z) - P_\Omega(A)\|_F^2 + \lambda_d Tr\left(X^T L_d X\right) + \lambda_t Tr\left(X L_t X^T\right) \\ &+ Tr\left(U^T(X-W)\right) + \frac{\mu_1}{2}\|X-W\|_F^2 + Tr\left(V^T(X-Z)\right) + \frac{\mu_2}{2}\|X-Z\|_F^2,\end{aligned} \tag{13}$$

where U and V are the Lagrange multipliers, $\mu_1 > 0$ and $\mu_2 > 0$ control the penalties for violating the linear constraints. Then the variables of LRSpNM can be approximated alternatively through the following steps:

Compute $\boldsymbol{W}_{k+1}$: The variable W can be calculated by the following equation with other variables fixed:

$$\begin{aligned}W_{k+1} &= arg\min_W \mathcal{L}(W, Z_k, X_k, U_k, V_k) \\ &= arg\min_W \|W\|_{S_p}^p + Tr\left(U_k^T(X_k - W)\right) + \frac{\mu_1}{2}\|X_k - W\|_F^2 \\ &= arg\min_W \|W\|_{S_p}^p + \frac{1}{\mu_1}\left\|W - X_k - \frac{1}{\mu_1}U\right\|_{kF}^2,\end{aligned} \tag{14}$$

where W_{k+1} can be obtained by the algorithm provided in Nie et al. (2015), which guaranteed convergence when $0 < p < 2$.

Compute $\boldsymbol{Z}_{k+1}$: When other variables are fixed, Z can be obtained by minimizing following function:

$$\begin{aligned} Z_{k+1} &= arg\min_{Z} \mathcal{L}(W_{k+1}, Z, X_k, U_k, V_k) \\ &= arg\min_{Z} \frac{\alpha}{2}\|P_\Omega(Z) - P_\Omega(A)\|_F^2 + Tr\left(V_k^T(X_k - Z)\right) + \frac{\mu_2}{2}\|X_k - Z\|_F^2, \end{aligned} \tag{15}$$

which is a convex optimization problem and can be solved by setting the derivative of Equation (15) to zero. Referred to the solution of Yang et al. (2019), which provides detailed derivation process, then we directly obtain

$$Z_{k+1} = \frac{1}{\mu_2}V_k + \frac{\alpha}{\mu_2}P_\Omega(A) + X_k - \frac{\alpha}{\alpha+\mu_2}P_\Omega\left(\frac{1}{\mu_2}V_k + \frac{\alpha}{\mu_2}P_\Omega(A) + X_k\right). \tag{16}$$

Compute $\boldsymbol{X}_{k+1}$: When other variables are fixed, X can be solved by minimizing the following objective function:

$$\begin{aligned} X_{k+1} &= arg\min_{X} \mathcal{L}(W_{k+1}, Z_{k+1}, X, U_k, V_k) \\ &= arg\min_{X} \lambda_d Tr\left(X^T L_d X\right) + \lambda_t Tr\left(X L_t X^T\right) + Tr\left(U_k^T(X - W_{k+1})\right) \\ &\quad + \frac{\mu_1}{2}\|X - W_{k+1}\|_F^2 + Tr\left(V_k^T(X - Z_{k+1})\right) + \frac{\mu_2}{2}\|X - Z_{k+1}\|_F^2. \end{aligned} \tag{17}$$

By setting the derivative of Equation (17) with respect to X to zero, we have

$$(2\lambda_d L_d + \mu_1 I)X + X(2\lambda_t L_t + \mu_2 I) = \mu_1 W_{k+1} + \mu_2 Z_{k+1} - U_k - V_k. \tag{18}$$

Equation (18) is a Sylvester equation (Bartels and Stewart, 1972), which provides the solution $X = Sylvester(A, B, C)$ of the matrix equation $AX + XB = C$. Thus, X_{k+1} can be solved by the following equation:

$$X_{k+1} = Sylvester(2\lambda_d L_d + \mu_1 I,\ 2\lambda_t L_t + \mu_2 I,\ \mu_1 W_{k+1} + \mu_2 Z_{k+1} - U_k - V_k). \tag{19}$$

Compute U_{k+1}, $\boldsymbol{V}_{k+1}$: We update the multipliers by

$$\begin{aligned} U_{k+1} &= U_k + \mu_1(X_{k+1} - W_{k+1}) \\ V_{k+1} &= V_k + \mu_2(X_{k+1} - Z_{k+1}). \end{aligned} \tag{20}$$

The variables W, Z, and X are iteratively updated until convergence. Finally, we obtain the predicted DTIs based on the completed entities in matrix X. LRSpNM repeats the ADMM iterations until convergence is reached.

Table 2. Area Under the Precision-Recall Curve Results for Drug–Target Interaction Prediction Under CV_drug

AUPR	*Enzyme*	*Ion channel*	*GPCR*	*Nuclear receptor*	*WANG*
LapRLS	0.111 (0.002)	0.172 (0.005)	0.219 (0.004)	0.370 (0.020)	0.417 (0.018)
WNN	*0.393 (0.013)*	0.334 (0.010)	0.367 (0.007)	*0.540 (0.020)*	*0.623 (0.002)*
KBMF2K	0.254 (0.010)	0.317 (0.009)	0.390 (0.014)	0.483 (0.030)	0.432 (0.011)
CMF	0.386 (0.008)	0.353 (0.014)	*0.406 (0.010)*	0.523 (0.030)	0.601 (0.005)
NRLMF	0.335 (0.031)	*0.355 (0.039)*	0.353 (0.028)	0.539 (0.059)	0.597 (0.007)
LRSpNM	**0.399 (0.009)**	**0.357 (0.014)**	**0.408 (0.012)**	**0.546 (0.021)**	**0.630 (0.012)**

Best and second-best AUPR results are bold and italics, respectively. Standard deviations are given in parentheses.

AUPR, area under the precision-recall curve; CMF, collaborative matrix factorization; CV, cross-validation; KBMF2K, kernelized Bayesian matrix factorization with twin kernels; LapRLS, Laplacian regularized least squares; LRSpNM, Laplacian regularized Schatten p-norm minimization; NRLMF, neighborhood regularized logistic matrix factorization; WNN, weighted nearest neighbor.

Table 3. Area Under the Precision-Recall Curve Results for Drug–Target Interaction Prediction Under CV_target

AUPR	*Enzyme*	*Ion channel*	*GPCR*	*Nuclear receptor*	*WANG*
LapRLS	0.638 (0.005)	0.702 (0.004)	0.310 (0.011)	0.369 (0.023)	0.298 (0.005)
WNN	0.778 (0.018)	0.763 (0.007)	0.574 (0.021)	0.492 (0.033)	*0.370 (0.007)*
KBMF2K	0.672 (0.024)	0.727 (0.013)	0.528 (0.018)	0.406 (0.021)	0.309 (0.024)
CMF	0.781 (0.013)	0.779 (0.011)	*0.599 (0.032)*	0.475 (0.016)	0.332 (0.003)
NRLMF	**0.810 (0.017)**	*0.795 (0.026)*	0.539 (0.039)	*0.523 (0.082)*	0.348 (0.031)
LRSpNM	*0.803 (0.017)*	**0.812 (0.011)**	**0.605 (0.022)**	**0.554 (0.047)**	**0.372 (0.008)**

Best and second-best AUPR results are bold and italics, respectively. Standard deviations are given in parentheses.

4. RESULTS

4.1. Experimental settings

In this experiment, we conduct 5 trials of 10-fold cross-validation (CV) to evaluate the de novo performance of LRSpNM. To evaluate the different aspect performance of the prediction methods, we consider two following types of de novo tests from new drugs and new targets aspects, respectively. The first one is called CV_drug where all drugs are randomly divided into 10 subsets. Another is CV_target where all targets are randomly divided into 10 subsets. That is to say, for a given DTI prediction method, CV_drug tests its ability to predict interactions for new drugs and CV_target tests its ability to predict interactions for new targets. Each subset is treated as the testing set in turn, whereas the remaining nine subsets are used as the training set. Both two types of de novo tests are repeated five times and the average accuracy values are showed as the final results. We use area under the precision-recall curve (AUPR) (Davis and Goadrich, 2006) as the evaluation metric. AUPR is a more sensitive metric to assess the prediction result of sparse data and more applicable in this experiment compared with another metric area under the ROC curve.

We perform the 10-fold CV on the training set for setting four parameters of LRSpNM, α, p, λ_d, and λ_t. The best parameter combination is selected by grid search from the range of values where $\alpha \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, $p \in \{0.25, 0.50, 0.75, 1, 1.25, 1.50, 1.75, 2\}$, λ_d and $\lambda_t \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. As for the preprocessing step, the parameter $K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ is also set by grid search.

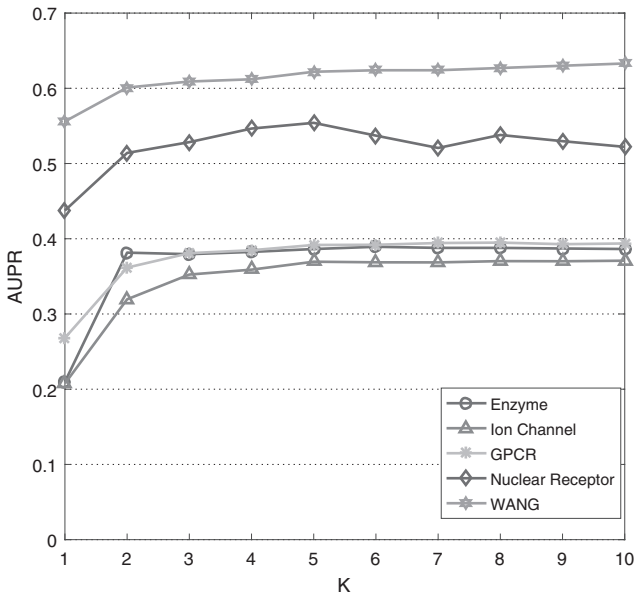

FIG. 2. AUPR with different settings of parameter K under CV_drug. AUPR, area under the precision-recall curve; CV, cross-validation.

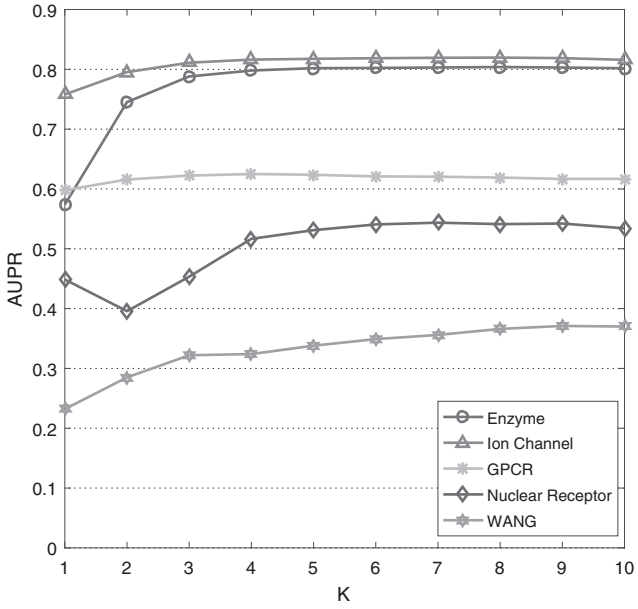

FIG. 3. AUPR with different settings of parameter K under CV_target.

4.2. *Performance results*

To comprehensively measure the prediction performance, five existing state-of-the-art DTI prediction methods are selected to compare with our LRSpNM model, including LapRLS (Xia et al., 2010), WNN (Van Laarhoven and Marchiori, 2013), KBMF2K (Gönen, 2012), CMF (Zheng et al., 2013), and NRLMF (Liu et al., 2016). For these competing methods, all parameters are set to their best values according to the authors' recommendation.

Table 2 shows the result of AUPR under the setting CV_drug. As shown in Table 2, LRSpNM outperforms all five competing methods on five data sets for new drug predictions. It means that LRSpNM outperforms than other methods and provides more accurate prediction under the setting CV_drug.

The results obtained under setting CV_target is presented in Table 3. For new target prediction, LRSpNM outperforms the competing methods except for the enzyme data set, where LRSpNM performs

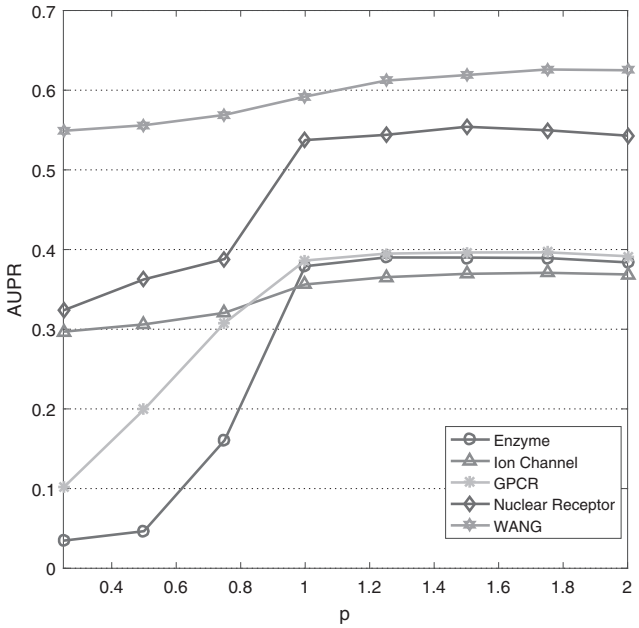

FIG. 4. AUPR with different settings of parameter p under CV_drug.

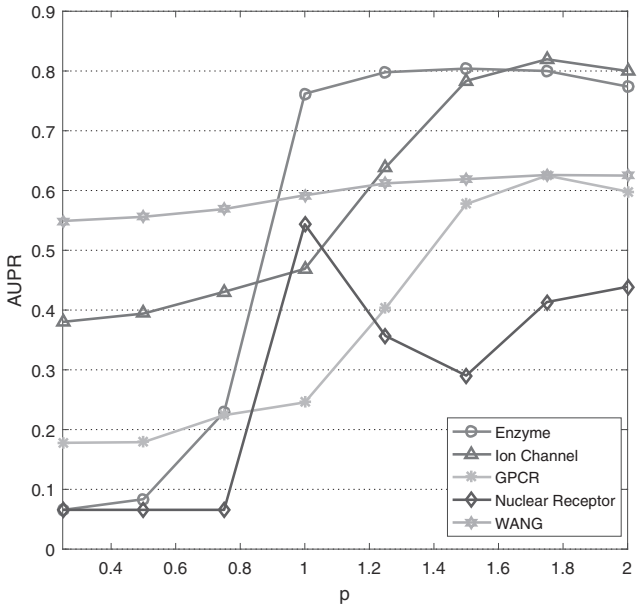

FIG. 5. AUPR with different settings of parameter p under CV_target.

slightly worse than NRLMF algorithm. LRSpNM reports AUPR values that are 2.094%, 0.992%, 5.596%, and 0.538% higher than the methods with second-best performance in other four data sets, respectively.

These results adequately demonstrate that LRSpNM has a higher accuracy on top-ranked drug–target pairs for novel prediction, which is more meaningful in drug discovery process.

4.3. Parameters analysis

In this section, we conduct cross-validation to investigate the effectiveness of parameters K and p of preprocessing step and Schatten p-norm, respectively. First, we analyze the parameter K in the prefilling step in five data sets. We can see that sensitivity analyses are provided for K in Figures 2 and 3. The result displays that the most K nearest neighbors' information of drugs and targets will assist the DTI prediction. It can be noticed the higher value of K, the better the prediction performance at first. After reaching stability, the change of K has little effect on prediction performance.

In addition, we analyze the parameter p to explore the prediction accuracy of Schatten p-norm. From the results of Figures 4 and 5, the AUPR values increase as the increase of the values of p and then become stable after certain value of p is reached under CV_drug setting in all data sets. Under CV_target setting, the nuclear receptor data set appears a special situation, where the AUPR values fluctuate with the increase of p. The reason may be that the number of known DTI in the nuclear receptor data set is excessively smaller than the other data sets, by which random division in CV might be less stable. From the results, we can find that the Schatten p-norm-based objective can approximate the rank minimization problem much better than the nuclear norm minimization (when $p=1$) to achieve better matrix completion results.

4.4. The effects of different Laplacian regularized terms of LRSpNM on performance

To evaluate the effectiveness of different Laplacian regularized terms, we compare LRSpNM with three circumstances in 10-fold cross-validation. We set $\lambda_d=0$ for the first case, $\lambda_t=0$ for the second case and

Table 4. The Performance of Different Laplacian Regularized Terms Under CV_drug

AUPR	*Enzyme*	*Ion channel*	*GPCR*	*Nuclear receptor*	*WANG*
LRSpNM ($\lambda_d=0$)	0.378	0.352	0.377	0.521	0.622
LRSpNM ($\lambda_t=0$)	**0.399**	0.361	0.383	**0.546**	0.612
SpNM	0.369	0.343	0.376	0.519	0.612
LRSpNM	**0.399**	**0.362**	**0.408**	**0.546**	**0.626**

Best AUPR results in different circumstances is bold.

Table 5. The Performance of Different Laplacian Regularized Terms Under CV_target

AUPR	*Enzyme*	*Ion channel*	*GPCR*	*Nuclear receptor*	*WANG*
LRSpNM ($\lambda_d = 0$)	**0.803**	0.810	0.603	0.543	**0.372**
LRSpNM ($\lambda_t = 0$)	0.797	0.790	0.598	0.532	0.371
SpNM	0.786	0.781	0.589	0.456	0.284
LRSpNM	**0.803**	**0.812**	**0.605**	**0.554**	**0.372**

Best AUPR results in different circumstances is bold.

$\lambda_d = \lambda_t = 0$ at the same time for the third case. In fact, the third situation is the form of Equation (10) without any regularization constraint, which can be marked as SpNM. It should be mentioned that the optimal remaining parameters are selected again.

Tables 4 and 5 show the different prediction results under CV_drug setting and CV_target setting, respectively. It can be found that as for λ_d, setting it to 0 negatively impacts results under CV_drug, but not so much under CV_target. Vice versa for λ_t, setting it to 0 negatively impacts results under CV_target, but not so much under CV_drug. Whereas setting $\lambda_d = \lambda_t = 0$ at the same time, the circumstance of SpNM negatively impacts the performance under both CV_drug and CV_target. This means that λ_d is important under CV_drug, whereas λ_t is important under CV_target. We can find that incorporating the Laplacian regularized term leads to more robust prediction results compared with simply minimizing the Schatten p-norm when compared SpNM with LRSpNM under two cross-validation settings.

4.5. Case study

In this section, we simulate some real-world situations to illustrate the prediction ability of LRSpNM. Specifically, we use LRSpNM method to predict a new drug or new target to see if its interactions will be predicted successfully. The largest data set—WANG is used as the training set to obtain the optimal parameters under CV_drug and CV_target, respectively. For new drug prediction, we recalculate the drug matrix S_d between the new drug and original drugs in WANG data set. The original targets are regarded as a candidate set. Finally, the predicted interactions are verified through the database DrugBank (Wishart et al., 2008). The same procedure is done for the new target prediction.

We select the new drug—Fludiazepam (PubChem ID: 3369), for which there are 13 validated targets of all 1469 targets in WANG data set. After LRSpNM is run on the modified data set, all targets are sorted in descending order of how likely they would interact with Fludiazepam. The top 15 predicted targets for Fludiazepam are given in Table 6. From the result, we can find that the 8 of 13 targets are predicted successfully in the top 15. As for the new target prediction, CALM1 (Uniprot ID: P0DP23) is chosen to

Table 6. Predicted Targets for New Drug Fludiazepam

Rank	*Target Uniport ID*	*Target name*
1	A8K177	GABRA1
2	A0A024R9X6	GABRA2
3	**P31644**	**GABRA5**
4	**P34903**	**GABRA3**
5	Q16445	GABRA6
6	P48169	GABRA4
7	B2RCW8	GABRB3
8	**P18505**	**GABRB1**
9	**P47870**	**GABRB2**
10	P08684	CYP3A4
11	**Q8N1C3**	**GABRG1**
12	**Q99928**	**GABRG3**
13	**P78334**	**GABRE**
14	**P18507**	**GABRG2**
15	Q9UN88	GABRQ

Successfully predicted interactions are in bold.

Table 7. Predicted Targets for New Target CALM1

Rank	*Drug PubChem ID*	*Drug name*
1	**5566**	**Trifluoperazine**
2	**3333**	**Felodipine**
3	**2726**	**Chlorpromazine**
4	**4485**	**Nifedipine**
5	**3372**	**Fluphenazine**
6	**4927**	**Promethazine**
7	**4748**	**Perphenazine**
8	4189	Miconazole
9	**16362**	**Pimozide**
10	**4768**	**Phenoxybenzamine**
11	**3763**	**Isoflurane**
12	5591	Troglitazone
13	1986	Acetazolamide
14	2972	Deferiprone
15	2973	Deferoxamine

Successfully predicted interactions are in bold.

conduct experiment, for which there are 10 validated drugs of all 449 drugs. The top 15 predicted drugs that interact with CALM1 are given in Table 7. The all 10 interactions of CALM1 are predicted successfully in the top 15.

We can notice that the aforementioned two cases (i.e., Fludiazepam and CALM1) are considered challenges where they are totally new to other drugs and targets. According to the case studies, it is shown that LRSpNM performs reasonably well. In summary, LRSpNM is generally able to predict targets for new drugs and drugs for new target.

5. CONCLUSIONS

This article presents a novel matrix completion method, named LRSpNM for de novo prediction of DTIs. In detail, we transform the task of DTIs prediction into a matrix completion problem, in which the potential interactions between drugs and targets can be discovered based on the prediction scores after the matrix completion procedure. The novelty of LRSpNM comes from first integrating Schatten p-norm minimization with Laplacian regularization to predict the interaction probability of an unknown drug–target pair. Specifically, when p becomes a tunable parameter in completing the DTI matrix, it can be adjusted to achieve better performance than nuclear norm where $p = 1$. Moreover, we use a preprocessing step that transforms the 0's in the given drug–target matrix into interaction likelihood values. This step fully utilizes the information of drug and target to fill partial unknown drug–target pairs.

A couple of experiments have been conducted to compare our method with five state-of-the-art methods under two different types of cross-validations for de novo prediction. In most of the cases, LRSpNM achieves the highest accuracy and presents the reliability of LRSpNM. Meanwhile, the two real case studies demonstrate that the proposed owns the capacity to predict potential novel interactions in de novo situation.

Of course, experimental results also illustrate that there is still much room for improvement. In this article, we only consider one type of representation for drugs or targets. Practically, each drug and target can have multiple representations from different aspects. For example, a drug also can be represented by its Anatomical Therapeutic Chemical (ATC) code or drug side effects. A target can be described by its gene expression values in cell level. As for future study, we will aim to integrate these multiview representations for DTIs prediction to further improve the prediction performance.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

This study is supported by the National Natural Science Foundation of China (Grant No. 61972423), Hunan Provincial Science and Technology Program (No. 2018wk4001), and 111Project (No. B18059).

REFERENCES

Apweiler, R., Bairach, A., Wu, C.H., et al. 2004. Uniport: the universal protein knowledgebase. *Nucleic Acids Res*. 32, D115–119.

Bartels, R.H., and Stewart, G.W. 1972. Solution of the matrix equation AX+XB = C [F4]. *Commun. ACM* 15, 820–826.

Bleakley, K., and Yamanishi, Y. 2009. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403.

Candes, E.J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Found Comput. Math*. 9, 717–772.

Chen, C., He, B., and Yuan, X. 2012. Matrix completion via an alternating direction method. *IMA J. Numer. Anal*. 32, 227–245.

Chen, H., and Zhang, Z. 2013. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS One* 8, e62975.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves, 233–240. Proceedings of the 23rd International Conference on Machine Learning. June 25–29, 2006. Pittsburgh, PA, USA.

Fan, X., Hong, Y., Liu, X., et al. 2018. Neighborhood constraint matrix completion for drug-target interaction prediction, 348–360. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer. June 3–6, 2018. Melbourne, VIC, Australia.

Fazel, M. 2002. Matrix rank minimization with applications [PhD thesis]. Stanford University. Palo Alto, CA, USA.

Gönen, M. 2012. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 28, 2304–2310.

Hattori, M., Okuno, Y., Goto, S., et al. 2003. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* 125, 11853–11865.

Jacob, L., and Vert, J.P. 2008. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 24, 2149–2156.

Keiser, M.J., Roth, B.L., Armbruster, B.N., et al. 2007. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206.

Kim, S., Thiessen, P.A., Bolton, E.E., et al. 2016. PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213.

Li, J., Zheng, S., Chen, B., et al. 2016. A survey of current trends in computational drug repositioning. *Brief Bioinform.* 17, 2–12.

Liu, Y., Wu, M., Miao, C., et al. 2016. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* 12, e1004760.

Lounkine, E., Keiser, M.J., Whitebread, S., et al. 2012. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486, 361–367.

Mizutani, S., Pauwels, E., Stoven, V., et al. 2012. Relating drug-protein interaction network with drug side effects. *Bioinformatics* 28, i522–i528.

Nie, F., Huang, H., and Ding, C. 2012. Low-rank matrix recovery via efficient schatten p-norm minimization. Proceedings of the 26th AAAI Conference on Artificial Intelligence. July 22–26, 2012. Toronto, Ontario, Canada.

Nie, F., Wang, H., Huang, H., et al. 2015. Joint Schatten p-norm and ℓp-norm robust matrix completion for missing value recovery. *Knowl. Inf. Syst*. 42, 525–544.

Núñez, S., Venhorst, J., and Kruse, C.G. 2012. Target-drug interactions: first principles and their application to drug discovery. *Drug Discov. Today* 17, 10–22.

Oprea, T.I., and Mestres, J. 2012. Drug repurposing: far beyond new targets for old drugs. *AAPS J.* 14, 759–763.

Scannell, J.W., Blanckley, A., Boldon, H., et al. 2012. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11, 191–200.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

van Laarhoven, T., and Marchiori, E. 2013. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 8, e66952.

Wang, C., and Kurgan, L. 2019. Review and comparative assessment of similarity-based methods for prediction of drug-protein interactions in the druggable human proteome. *Brief Bioinform.* 20, 2066–2087.

Wishart, D.S., Knox, C., Guo, A.C., et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906.

Xia, Z., Wu, L.Y., Zhou, X., et al. 2010. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* 4(Suppl 2), S6.

Yamanishi, Y., Araki, M., Gutteridge, A., et al. 2008. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240.

Yang, M., Luo, H., Li, Y., et al. 2019. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455–i463.

Zheng, X., Ding, H., Mamitsuka, H., et al. 2013. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, 1025–1033. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 11–14, 2013. Chicago, IL, USA.

Address correspondence to:
Dr. Mengyun Yang
School of Science
Shaoyang University
Xueyuan Road, Daxiang District
Shaoyang 422000, Hunan
P.R. China

E-mail: mengyunyang@csu.edu.cn