Bioinformatics, 36(22-23), 2020, 5456–5464
doi: 10.1093/bioinformatics/btaa1024
Advance Access Publication Date: 17 December 2020
Original Paper

Systems biology

Heterogeneous graph inference with matrix completion for computational drug repositioning

Mengyun Yang[1,2], Lan Huang[1], Yunpei Xu[1], Chengqian Lu [1] and Jianxin Wang [1,*]

[1]The Hunan Provincial Key Lab of Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China and [2]School of Science, Shaoyang University, Shaoyang 422000, China

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on August 9, 2020; revised on November 23, 2020; editorial decision on November 25, 2020; accepted on November 26, 2020

Abstract

Motivation: Emerging evidence presents that traditional drug discovery experiment is time-consuming and high costs. Computational drug repositioning plays a critical role in saving time and resources for drug research and discovery. Therefore, developing more accurate and efficient approaches is imperative. Heterogeneous graph inference is a classical method in computational drug repositioning, which not only has high convergence precision, but also has fast convergence speed. However, the method has not fully considered the sparsity of heterogeneous association network. In addition, rough similarity measure can reduce the performance in identifying drug-associated indications.

Results: In this article, we propose a heterogeneous graph inference with matrix completion (HGIMC) method to predict potential indications for approved and novel drugs. First, we use a bounded matrix completion (BMC) model to prefill a part of the missing entries in original drug–disease association matrix. This step can add more positive and formative drug–disease edges between drug network and disease network. Second, Gaussian radial basis function (GRB) is employed to improve the drug and disease similarities since the performance of heterogeneous graph inference more relies on similarity measures. Next, based on the updated drug–disease associations and new similarity measures of drug and disease, we construct a novel heterogeneous drug–disease network. Finally, HGIMC utilizes the heterogeneous network to infer the scores of unknown association pairs, and then recommend the promising indications for drugs. To evaluate the performance of our method, HGIMC is compared with five state-of-the-art approaches of drug repositioning in the 10-fold cross-validation and *de novo* tests. As the numerical results shown, HGIMC not only achieves a better prediction performance but also has an excellent computation efficiency. In addition, cases studies also confirm the effectiveness of our method in practical application.

Availabilityand implementation: The HGIMC software and data are freely available at https://github.com/BioinformaticsCSU/HGIMC, https://hub.docker.com/repository/docker/yangmy84/hgimc and http://doi.org/10.5281/zenodo.4285640.

Contact: jxwang@mail.csu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The traditional drug discovery is time-consuming, complex and expensive (Chong *et al.*, 2007; Pushpakom *et al.*, 2019). The investment in drug discovery has increased dramatically over past few decades, but the number of novel drugs on the market remains low. In fact, about 90% of drug candidates in Phase I clinical trials fail (Li *et al.*, 2016). Drug repositioning is considered as a promising strategy for drug development, whose goal is to find new candidate indications for existing drugs. Since these known drugs has already passed various clinical trials and ensured safety, Phase I clinical trials can be skipped for the repositioned drugs. Therefore, drug repositioning can shorten the period of drug research and reduce costs. There are some recent reviews of drug repositioning in terms of databases (Tanoli *et al.*, 2020), methods (Luo *et al.*, 2020) and trends (Karaman and Sippl, 2019).

So far, many computational drug repositioning methods have been proposed and can be roughly divided into three categories, including machine learning-based methods, network propagation-based methods and recommendation system-based methods. These methods are based on an assumption that similar drugs are associated with similar diseases and vice versa. In machine learning-based

methods, prediction of potential drug–disease associations can be treated as a binary classification problem. The drug–disease associations are considered as samples, while the prior similarities of drug and disease are considered as features. Motivated by the fact that the clinical side-effects provide a human phenotypic profile for drugs, Yang and Agarwal (2011) extended Naive Bayes models to identify potential indications for clinical drugs using the side-effects as features. Integrating multiply similarities of drug and disease, Gottlieb *et al.* (2011) used a logistic regression model to predict promising drug–disease associations. Moreover, support vector machine (SVM) classifier was also used to validate therapeutic classes for known drugs (Napolitano *et al.*, 2013). Deep learning is a new subfield of machine learning, which has been used in computational drug repositioning in recent years. Aliper *et al.* (2016) used a fully connected deep neural network to identify potential new indications for drugs. The model is superior to SVM in classification accuracy. Zeng *et al.* (2019) proposed a network-based deep learning method named deepDR to predict new drug–disease associations, which can learn latent features of drugs from multiple drug-related networks by a multi-modal deep autoencoder.

The network propagation-based methods can infer the missing edges in heterogeneous networks and have a great advantage in computation efficiency. According to the guilt-by-association principle, Wang *et al.* (2013) proposed a heterogeneous graph based inference (HGBI) algorithm for predicting drug-related targets. It updated edge weights among interaction pairs by using all pathways in the heterogeneous graph, which not only has high convergence precision, but also has fast convergence speed. HGBI algorithm was also extended to identify new candidate indications for drugs (Wang *et al.*, 2014 b). Martinez *et al.* (2015) proposed a network-based prioritization method, called DrugNet, to predict novel diseases for existing drugs. Through constructing a heterogeneous drug–target–disease network, DrugNet is able to use propagation flow to implement drug–disease prioritization and disease–drug prioritization. Through exploiting sparse drug–disease associations to enhance the similarity measures of drug and disease, Luo *et al.* (2016) developed a bi-random walk algorithm, namely MBiRW, to perform association prediction.

For recommendation system-based methods, the problem of predicting drug–disease associations is considered as a user–item rating problem. Matrix completion and matrix factorization are the most popular approaches in recommendation system. Luo *et al.* (2018) proposed a drug repositioning recommendation system (DRRS) to predict promising indications for known and novel drugs. DRRS employed a singular value thresholding (SVT) algorithm (Cai *et al.*, 2010) to complete a global adjacency matrix consisting of drug similarity matrix, disease similarity matrix and drug–disease association matrix. Based on the adjacency matrix of DRRS, Yang *et al.* (2019a) developed a low-rank matrix completion method, namely bounded nuclear norm regularization (BNNR). BNNR can handle the noise from drug and disease similarities by integrating a regularization term. Additionally, the bounded constraint can ensure that all the completed values are within a specific interval. In order to incorporate multiple types of drug and disease information, Yang *et al.* (2019b) proposed overlap matrix completion for tri-layer heterogeneous networks (OMC3). Xuan *et al.* (2019) proposed a non-negative matrix factorization method, called DisDrugPred. To fully exploit useful latent information, DisDrugPred took the multi-similarities of drug and disease and the sparsity of drug–disease pairs into account.

Inspired by the low-rank completion of BNNR and the guilt-by-association principle of HGBI, we propose a heterogeneous graph inference with matrix completion (HGIMC) method to predict new drug-associated indications. The main idea of HGIMC is overcoming two deficiencies of HGBI model, including an extremely sparse heterogeneous association network and less accurate similarity matrices. Specifically, first, we calculate five measures of drug similarities and two measures of disease similarities using existing software packages and obtain the average of them. Then, we use bounded matrix completion (BMC) to prefill some drug–disease associations with high confidence for enriching the edges between drug network and disease network. Moreover, we apply Gaussian radial basis function (GRB) to the averaged similarity matrix for obtaining better similarity measures. Finally, integrating these two comprehensive similarity matrices and the updated drug–disease association matrix, the heterogeneous graph inference can quickly and accurately predict the potential drug–disease associations. The overall flowchart of HGIMC model is shown in Figure 1.

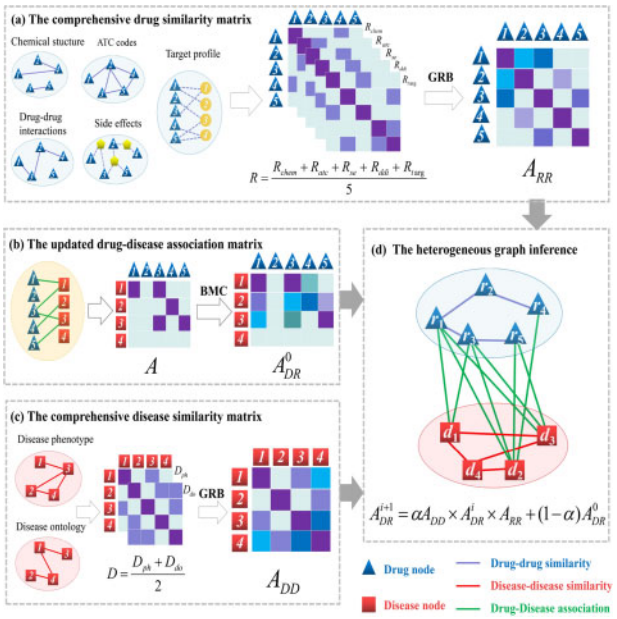

Fig. 1. The overall flowchart of HGIMC model. (**a**) The comprehensive drug similarity matrix based on GRB. (**b**) The updated drug–disease association matrix using BMC. (**c**) The comprehensive disease similarity matrix using GRB. (**d**) The heterogeneous graph inference

2 Materials and methods

The gold standard dataset (Gottlieb *et al.*, 2011) in drug repositioning was used to demonstrate the effectiveness of our proposed method. It contained 1933 validated drug–disease associations involving 593 drugs and 313 diseases. These drugs and diseases were derived from DrugBank (Wishart, 2006) and Online Mendelian Inheritance in Man (OMIM) database (Ada *et al.*, 2002), respectively. The corresponding drug–disease association matrix can be represented by a binary matrix $A \in \{0,1\}^{n\times m}$, where m and n denote the number of drug and disease nodes, respectively. The existing drug–disease pairs are denoted as 1 s, while the unknowns are represented as 0 s.

For drugs, we calculated five measures of drug similarities (Huang *et al.*, 2020), including chemical structures similarity R_{chem}, anatomical therapeutic chemical (ATC) codes similarity R_{atc}, side effects similarity R_{se}, drug–drug interactions similarity R_{ddi} and target profiles similarity R_{targ}. Based on the Canonical SMILES (Weininger, 1988) files of drugs, we used the Chemical Development Kit (CDK) (Steinbeck *et al.*, 2003) tool to compute the hashed fingerprints for all drugs and then obtained R_{chem}. All ATC codes of related drugs were extracted from DrugBank. We applied a semantic similarity algorithm (Resnik, 1995) to calculate the similarity scores among ATC terms and then got R_{atc}. The remaining of similarities, including R_{se}, R_{ddi} and R_{targ}, were measured by Jaccard similarity coefficient (Jaccard, 1908), which can be formulated as follows,

$$R_{se/ddi/targ}(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|},$$

where $|\cdot|$ represented the cardinality of a set. S_i represented the set of side-effect profiles of drug i in R_{se}, drug–drug interaction profiles

of drug i in R_{ddi} and drug–target interaction profiles of drug i in R_{targ}, respectively. Specifically, in R_{ddi}, each drug was represented as an interaction profile consisting of all drugs that were known to interact with it. Additionally, side effects of drugs were extracted from SIDER (Kuhn *et al.*, 2016), and drug–drug interactions and drug–target interactions were extracted from DrugBank.

For disease, we computed two measures of disease similarities, including disease phenotype similarity D_{ph} and disease ontology similarity D_{do}. Specifically, D_{ph} was collected from MimMiner (Van *et al.*, 2006). According to the structure of disease ontology terms, D_{do} was calculated by using the gene ontology-based algorithm (Wang *et al.*, 2007).

In summary, we have collected five drug similarity matrices (i.e. $R_{chem}, R_{atc}, R_{se}, R_{ddi}, R_{targ}$) and two disease similarity matrices (i.e. D_{ph}, D_{do}). In order to simply integrate there multiple similarities, we averaged out the five drug similarity matrices and denoted $R \in \mathrm{IR}^{m\times m}$. Similarly, the averaged disease matrix was obtained and denoted $D \in \mathrm{IR}^{n\times n}$.

3 HGIMC for predicting drug–disease associations

Based on the guilt-by-association principle (Barabási *et al.*, 2011; Chiang and Butte, 2009), it is key to construct a heterogeneous network with rich and reliable edges among association nodes. However, the heterogeneous graph based inference (HGBI) algorithm (Wang *et al.*, 2013) uses original sparse associations as the initial paths, which ignores more potential and important inference edges. Additionally, setting a rough threshold to truncate the similarity weights is not a precise measure of similarity. In order to improve the HGBI algorithm, we propose an HGIMC algorithm, which consists of bounded matrix completion (BMC), Gaussian radial basis (GRB) and HGBI algorithm. Specifically, we introduce BMC to increase promising edges in drug–disease association network, and make use of GRB to obtain more reliable correlations in drug and disease similarity networks. After finishing the above BMC and GRB steps, HGBI algorithm is employed to identify potential drug–disease associations. The details of BMC, GRB and HGBI are described in the next sections.

3.1 The updated drug–disease association matrix via BMC

Under the low-rank assumption, matrix completion methods (Ramlatchan *et al.*, 2018; Yang *et al.*, 2020) can recover the missing entries of a low-rank matrix. Specifically, the drug–disease association matrix is often incomplete since only a few of associations have been verified by clinic trials and the remaining are missing. Based on the assumption that similar drugs are associated with similar diseases, the latent factors determining drug–disease associations are highly correlated. So the association matrix is low-rank. In this study, we don't choose inductive matrix completion (Chen *et al.*, 2018a) since it requires high quality prior information in the optimization. We utilize a bounded matrix completion (BMC) method (Yang *et al.*, 2019a) without using any similarity information, our previous work, to fill out the missing elements in drug–disease association matrix, which can be formulated as follows,

$$\begin{aligned} \min_X \quad & \|X\|_* + \frac{\alpha}{2}\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(A)\|_F^2 \\ s.t. \quad & 0 \le X \le 1. \end{aligned} \tag{1}$$

where $\|X\|_*$ represents the nuclear norm of X, which can lead to a low-rank approximation for X. α is a harmonic parameter balancing the nuclear norm and the error term. A is the original drug–disease association matrix, Ω is a set of index pairs containing all non-zero elements and $\mathcal{P}_\Omega$ is the projection operator projecting X onto Ω, which is defined as

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

Moreover, the constraint $0 \le X \le 1$ ensures that the values of each element in X are in the interval [0,1]. To solve this model, an alternating direction method of multipliers (ADMM) is employed by introducing a new variable matrix W. The model (1) can be formulated as the following equivalent form,

$$\begin{aligned} \min_X \quad & \|X\|_* + \frac{\alpha}{2}\|\mathcal{P}_\Omega(W) - \mathcal{P}_\Omega(A)\|_F^2 \\ s.t. \quad & X = W, \\ & 0 \le W \le 1. \end{aligned} \tag{2}$$

The augmented Lagrangian function of model (2) is

$$\begin{aligned} \ell(W, X, Y, \alpha, \beta) = \quad & \|X\|_* + \frac{\alpha}{2}\|\mathcal{P}_\Omega(W) - \mathcal{P}_\Omega(A)\|_F^2 \\ & + Tr(Y^T(X - W)) + \frac{\beta}{2}\|X - W\|_F^2, \end{aligned} \tag{3}$$

where Y is a Lagrange multiplier and $\beta > 0$ is a parameter. We can give the following optimization using ADMM:

$$W_{k+1} = \underset{0 \le W \le 1}{\operatorname{argmin}}\, \ell(W, X_k, Y_k, \alpha, \beta), \tag{4}$$

$$X_{k+1} = \underset{X}{\operatorname{argmin}}\, \ell(W_{k+1}, X, Y_k, \alpha, \beta), \tag{5}$$

$$Y_{k+1} = Y_k + \beta(X_{k+1} - W_{k+1}). \tag{6}$$

We use the inverse operator (Yang and Yuan, 2012) to solve Equation (4) and acquire a closed-form solution W^* as follows,

$$W^* = (\mathcal{I} - \frac{\alpha}{\alpha+\beta}\mathcal{P}_\Omega)\left(\frac{1}{\beta}Y_k + \frac{\alpha}{\beta}\mathcal{P}_\Omega(M) + X_k\right),$$

where $\mathcal{I}$ denotes the identity operator. Moreover, to limit the element values of W_{k+1} in the range of [0,1], we employ the following projection operator

$$W_{k+1} = \mathcal{Q}_{[0,1]}(W^*), \tag{7}$$

where $\mathcal{Q}_{[0,1]}$ is defined as

$$(\mathcal{Q}_{[0,1]}(W^*))_{ij} = \begin{cases} 1, & W^*_{ij} > 1 \\ W^*_{ij}, & 0 \le W^*_{ij} \le 1 \\ 0. & W^*_{ij} < 0 \end{cases}$$

Equation (5) can be rearranged as

$$\begin{aligned} X_{k+1} \quad & = \operatorname{argmin}_X \|X\|_* + \frac{\beta}{2}\|X - (W_{k+1} - \frac{1}{\beta}Y_k)\|_F^2 \\ & = \mathcal{D}_{\frac{1}{\beta}}(W_{k+1} - \frac{1}{\beta}Y_k), \end{aligned} \tag{8}$$

where $\mathcal{D}_\tau(X)$ is the singular value shrinkage (SVT) operator (Cai *et al.*, 2010; Ma *et al.*, 2011). Specifically, SVT operator is described as

$$\mathcal{D}_\tau(X) = \sum_{i=1}^{\sigma_i \ge \tau}(\sigma_i - \tau)u_i v_i^T,$$

where σ_i is the ith singular value of X larger than threshold τ, while u_i and v_i are the left and right singular vectors corresponding to σ_i, respectively.

Additionally, we terminate the BMC model when the following stopping criteria are satisfied:

$$f_k \leq tol_1, \frac{|f_{k+1} - f_k|}{max\{1, |f_k|\}} \leq tol_2, \tag{9}$$

where $f_k = \frac{\|X_{k+1} - X_k\|_F}{\|X_k\|_F}$, tol_1 and tol_2 are the given tolerances, which are set as 2×10^{-3} and 10^{-5}, respectively.

After BMC optimization, we remove all entries in the completed association matrix with scores less than a threshold (set to 0.1 empirically). Because the main aim of BMC is to add the strong edges to enhance the connectivity of heterogeneous drug–disease network, not the weak edges. Moreover, if the matrix A contains the whole zero rows/columns, BMC cannot effectively deal with them and the corresponding predicted values are meaningless, which are close to 0, such as 10^{-5}. These completed values in this case are simultaneously removed. Finally, we denotes the updated association matrix as A^0_{DR}.

3.2 The comprehensive similarity matrix via GRB

To improve the similarity measure, we recalculate a comprehensive similarity with the row vectors of the similarity matrix. Specifically, the ith row vector of the drug similarity matrix R represents the similarity values among the ith drug node and all drug nodes. In fact, it can also be considered as the feature representation of ith drug node. We introduce Gaussian radial basis function (GRB) to compute the distance among these feature representations. The distance scores of drug nodes are treated as the comprehensive similarity values, which can better measure the similarity from a global perspective. The Gaussian radial basis similarity of drug can be defined as,

$$A_{RR}(i,j) = \exp\left(\frac{\|\boldsymbol{r_i} - \boldsymbol{r_j}\|^2}{-2\sigma^2}\right), \tag{10}$$

where $\boldsymbol{r_i}$, $\boldsymbol{r_j}$ are ith and jth row vectors of matrix R, respectively. Parameter σ is used to control the function bandwidth. In this study, we uniformly set σ to 0.5. A_{RR} is the comprehensive similarity matrix after GRB processing. In the same manner, based on the disease matrix D, we can also compute the Gaussian radial basis similarity of disease. The corresponding similarity matrix denotes A_{DD}.

3.3 The heterogeneous graph inference algorithm

After completing the above BMC and GRB steps, we can construct a completely new drug–disease heterogeneous graph, using the updated drug–disease associations, Gaussian radial basis similarities of drug and disease. The iteration of heterogeneous graph based inference (HGBI) (Wang *et al.*, 2013) can be formulated as matrix multiplications:

$$A^{i+1}_{DR} = \gamma A_{DD} \times A^i_{DR} \times A_{RR} + (1-\gamma)A^0_{DR}. \tag{11}$$

In each iteration, the all drug–disease associations can be updated by drug and disease similarities inference with a probability γ and the prefilled associations with a probability $1-\gamma$. The parameter γ is fixed to 0.1 in this study. Moreover, when similarity matrices A_{RR} and A_{DD} are normalized as

$$\begin{aligned} A_{RR}(i,j) &= \frac{A_{RR}(i,j)}{\sqrt{\sum_{k=1}^{m} A_{RR}(i,k) \sum_{k=1}^{m} A_{RR}(k,j)}}, \\ A_{DD}(i,j) &= \frac{A_{DD}(i,j)}{\sqrt{\sum_{k=1}^{n} A_{DD}(i,k) \sum_{k=1}^{n} A_{DD}(k,j)}}, \end{aligned} \tag{12}$$

Equation (11) will converge (Wang *et al.*, 2013). The stopping criterion of heterogeneous graph inference algorithm is

$$max(A^{i+1}_{DR} - A^i_{DR}) \leq tol_3, \tag{13}$$

where $max(A)$ denotes the maximum value in matrix A and tol_3 is set to 10^{-10}.

In summary, the novelty of HGIMC method is improving the HGBI algorithm from two perspectives. One is enriching the sparse heterogeneous association network by BMC. The BMC can add some strong drug–disease associations without any prior similarity, which can provide more reliable paths for the graph inference. But the BMC does not predict any valid associations for new nodes. The other is enhancing similarity measures by GRB. The GRB can compute a comprehensive measure based on the row/column vectors of similarity matrix for the graph inference. It plays an important role in identifying potential associations for new nodes.

4 Results and discussion

4.1 Experimental settings

We conducted a 10-fold cross-validation and a *de novo* test to evaluate the performance of HGIMC. In the 10-fold cross-validation, all known drug–disease pairs were randomly divided into ten parts. Each part was in turn treated as the test samples, while the rest of associations were considered as the training samples. We had repeated the cross-validation ten times and averaged the corresponding results. In the *de novo* test, the existing drugs with only one known association had been tested. In the gold standard dataset, the number of these drugs was 171. Specifically, for each of these drugs, its association was in turn removed as the test sample and the remaining associations were treated as the training samples. It is worth mentioning that the *de novo* test can show the capability of a method in predicting potential indications for new drugs. For the sake of rigor, we removed the side effects similarity and drug-drug interactions similarity in the *de novo* test, since new drugs may have no side effect information and drug-drug interaction data. In addition, we selected three evaluation metrics to assess the prediction performance, including the area under the receiver operating characteristic (ROC) curve (AUC), the area under the Precision-Recall curve (AUPR) and Precision.

In HGIMC algorithm, there are two parameters needed to determine, including α and β. We pick the α and β values from {0.1, 1, 10, 100} based on cross validation by grid search. Supplementary Table S1 gives the AUC values of HGIMC under various α and β in the 10-fold cross-validation. As the Supplementary Table S1 shown, the AUC value is the highest when $\alpha = 10$ and $\beta = 10$. For other datasets, we also set both α and β to 10 in this study. In addition, there are three hyper-parameters set empirically, including σ (0.5), the threshold (0.1) and γ (0.1). To give a sensitivity analysis for them, we vary one hyper-parameter while the other two hyper-parameters are fixed. These hyper-parameters are selected from {0.1, 0.3, 0.5, 0.7, 0.9} by grid search. Supplementary Figure S1 shows the performance trend of HGIMC with respect to different settings for these hyper-parameters in the 10-fold cross-validation.

4.2 Comparisons with the state-of-the-arts

HGIMC was compared with five latest computational methods: BNNR (Yang *et al.*, 2019a), DRRS (Luo *et al.*, 2018), MBiRW (Luo *et al.*, 2016), DrugNet (Martinez *et al.*, 2015) and HGBI (Wang *et al.*, 2013). The parameters involved in these compared methods are set to the recommended values by the authors or grid search. Specifically, in BNNR algorithm, two parameters α and β are selected from {0.1, 1, 10, 100} by grid search based on the cross validation. In DRRS algorithm, there are two adaptive parameters, i.e. τ and δ. In MBiRW algorithm, we use the recommended values of three parameters, i.e. $\alpha = 0.3, l = r = 2$. Because these default parameters were also chosen by grid search on the gold standard dataset in the MBiRW literature. In DrugNet algorithm, the parameter α is picked from {0.1, 0.2,...,0.9} by grid search. In HGBI algorithm, we find that a parameter ($\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$) has little influence on the performance of HGBI algorithm, so we choose to use the author's recommended value ($\alpha = 0.4$). In addition, prior similarity data in there methods are those referred to the original literatures, i.e. the chemical structures similarity of drug and the phenotype similarity of disease.

Table 1. AUC, AUPR and precision values of all compared methods in the 10-fold cross-validation and *de novo* tests on the gold standard dataset

Tests	Metrics	HGIMC	BNNR	DRRS	MBiRW	DrugNet	HGBI
10-fold CV	AUC	**0.941**	0.932	0.930	0.917	0.868	0.829
	AUPR	0.394	**0.402**	0.341	0.264	0.155	0.102
	Precision	0.438	**0.440**	0.375	0.304	0.192	0.130
de novo	AUC	**0.877**	0.830	0.824	0.818	0.782	0.746
	AUPR	**0.356**	0.199	0.197	0.189	0.102	0.075
	Precision	**0.433**	0.251	0.269	0.234	0.135	0.099

Note: The best results are **bold** and the second best results are underlined.

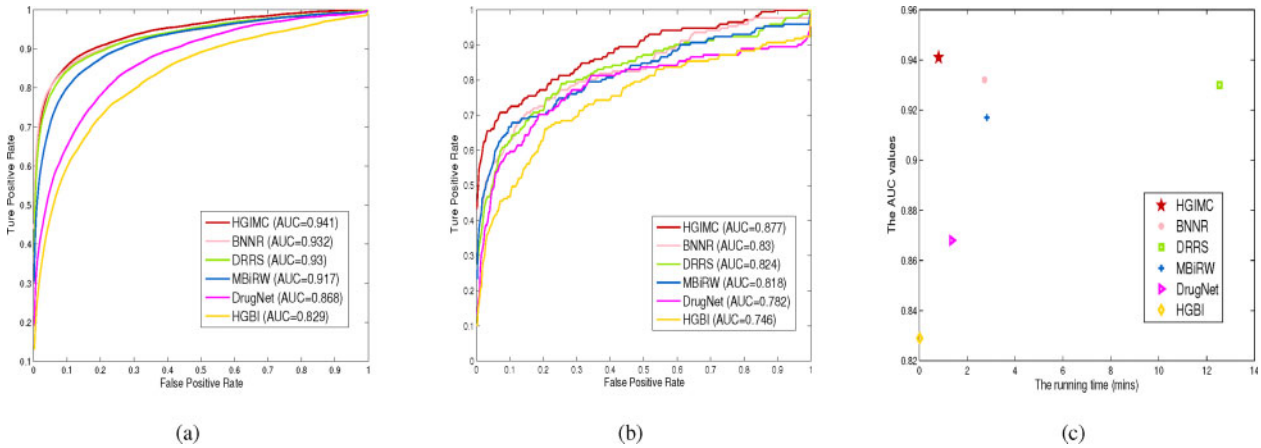

Fig. 2. The prediction results of all methods on gold standard dataset. (**a**) The ROC curves in the 10-fold cross-validation. (**b**) The ROC curves in *de novo*. (b) The running time and AUC values in the 10-fold cross-validation

Table 1 shows the AUC, AUPR and Precision values obtained by six drug repositioning methods on the gold standard dataset. As the results shown in Table 1, HGIMC achieves the best AUCs of 0.941 in the 10-fold cross-validation and 0.877 in *de novo*, which is 13.510% and 17.560% higher than the corresponding AUCs of HGBI, respectively. In the 10-fold cross-validation, HGIMC obtains 0.428 and 0.463 on AUPR and Precision, respectively, which is slightly lower than BNNR method. However, in *de novo*, HGIMC obtains the best AUPR of 0.356 and the best Precision of 0.433, which is 78.894% and 72.510% better than BNNR on AUPR and Precision, respectively. The computational results demonstrate the superior prediction accuracy of our method in identifying potential indications for existing and novel drugs. The ROC curves of all compared approaches in the 10-fold cross-validation and *de novo* tests are shown in Figure 2(a) and (b).

Furthermore, in order to illustrate the computational efficiency of all compared approaches, we have recorded the running time of a 10-fold cross-validation on a Linux server with CPU 2.60 GHz and 1 TB memory. As shown in Figure 2(c), the running time and the AUC values of all methods are reported. Specifically, only HGIMC and HGBI can complete the cross-validation test in less than 1 min, but HGBI yields the lowest AUC value. Additionally, only HGIMC, BNNR and DRRS can obtain the AUC value greater than 0.92, but the running time of DRRS is more than 12 min. In summary, HGIMC is a promising prediction method from both prediction effectiveness and computation efficiency.

4.3 Comparisons with the multi-similarities fusion methods

In this section, we present the comparisons among HGIMC and the multi-similarities fusion methods. First, we select a popular and effective fusion method, named similarity network fusion (SNF) (Wang *et al.*, 2014a), to integrate the multi-similarities of drug and disease. SNF is a nonlinear method for data integration. It uses K nearest neighbors and cross-diffusion process to iteratively update each similarity network capturing the common and complementary information from the other similarity networks. The two key parameters in SNF algorithm, K nearest neighbors (K) and iteration steps (t), are both set to 10 empirically. Second, these fused similarity matrices are input into BNNR, DRRS, MBiRW, DrugNet and HGBI models in turn. We denoted these new approaches as BNNR-SNF, DRRS-SNF, MBiRW-SNF, DrugNet-SNF and HGBI-SNF, respectively. Table 2 presents the prediction results of these methods on the gold standard dataset. As shown in Table 2, HGIMC outperforms all SNF methods in the 10-fold cross-validation. In *de novo*, the AUC value of our method (0.877) is slightly lower than that of DrugNet-SNF (0.881), but the AUPR and Precision values of HGIMC are the best, which are higher 26.690% and 27.729% than that of DrugNet-SNF. Comparing with Tables 1 and 2, we find that the SNF technology can improve DrugNet and HGBI methods comprehensively. However, the results of MBiRW-SNF are worse than MBiRW. MBiRW is a Bi-Random Walk (BiRW) method with improved similarity measurement. Specifically, before the BiRW algorithm is implemented, a logistic function was employed to enhance the similarity measure of drug and disease. In MBiRW-SNF algorithm, the logistic function has changed the structure of fused similarity matrix by SNF method. That is the main reason for the inferior of MBiRW-SNF. Furthermore, for BNNR-SNF and BNNR, the performance of BNNR is better than BNNR-SNF in 10-fold cross-validation, while BNNR-SNF is better than BNNR in terms of AUC, AUPR and Precision in *de novo* test. It illustrates a method with multi-similarities is not necessarily better than single similarity approach, which depends on the actual data and the method characteristics.

Table 2. AUC, AUPR and precision values of all compared methods with multi-similarities fusion on the gold standard dataset

Tests	Metrics	HGIMC	BNNR-SNF	DRRS-SNF	MBiRW-SNF	DrugNet-SNF	HGBI-SNF
10-fold CV	AUC	**0.941**	0.919	0.916	0.881	0.908	0.923
	AUPR	**0.394**	0.385	0.361	0.240	0.166	0.305
	Precision	**0.438**	0.425	0.400	0.290	0.200	0.348
de novo	AUC	0.877	0.831	0.813	0.795	**0.881**	0.873
	AUPR	**0.356**	0.291	0.318	0.149	0.281	0.336
	Precision	**0.433**	0.345	0.386	0.170	0.339	0.398

Note: The best results are **bold** and the second best results are underlined.

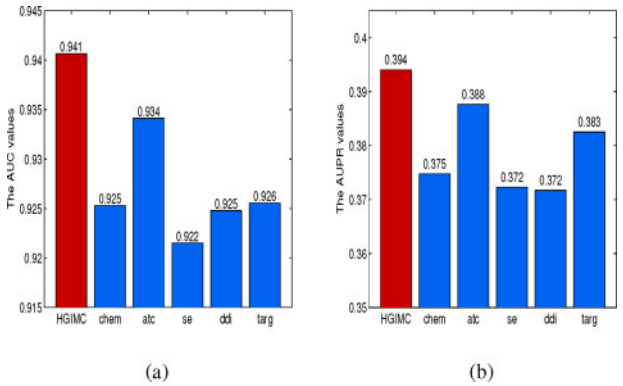

Fig. 3. The performance of individual similarity and the averaged multi-similarities in 10-fold cross-validation. (**a**) The AUC comparison. (**b**) The AUPR comparison

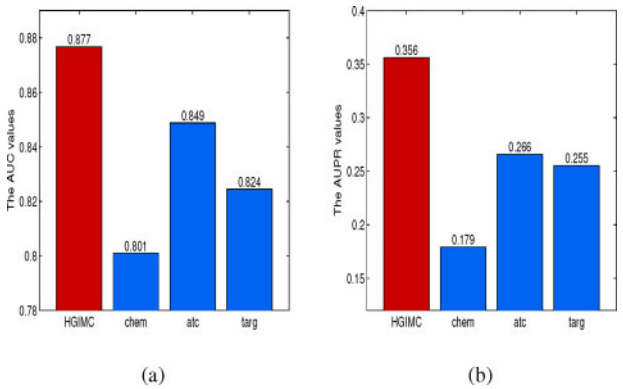

Fig. 4. The performance of individual similarity and the averaged multi-similarities in *de novo*. (**a**) The AUC comparison. (**b**) The AUPR comparison

4.4 Analysis of individual similarity and multi-similarities in HGIMC

To compare the performance of individual similarity and the averaged multi-similarities in HGIMC algorithm, we take drug similarities as examples while fixing the averaged disease similarity. In HGIMC algorithm, we use single drug similarity matrix ($R_{chem}, R_{atc}, R_{se}, R_{ddi}, R_{targ}$) to replace the averaged matrix of drug multi-similarities (R), respectively. The comparison results of 10-fold cross-validation and *de novo* tests are shown in Figures 3 and 4. As shown in Figures 3 and 4, HGIMC with the averaged drug similarity obtains the best AUC and AUPR values. It illustrates HGIMC can effectively intergrate multiple similarities. Additionally, R_{atc} seems more effective compared with the other individual similarity.

4.5 Case studies

In this section, we perform case studies to demonstrate capability of HGIMC in practical applications. Specifically, all known drug–disease associations on gold standard dataset are treated as training set,

Table 3. The compared results of HGBI, BMC-HGBI, GRB-HGBI and HGIMC in the 10-fold cross-validation and *de novo* tests under the single similarity case

Tests	Metrics	HGBI	BMC-HGBI	GRB-HGBI	HGIMC
10-fold CV	AUC	0.829	0.908	0.872	**0.920**
	AUPR	0.102	0.359	0.184	**0.377**
	Precision	0.130	0.404	0.229	**0.424**
de novo	AUC	0.746	0.741	**0.818**	0.806
	AUPR	0.075	0.068	0.171	**0.209**
	Precision	0.099	0.088	0.222	**0.281**

Note: The best results are **bold** and the second best results are underlined.

HGIMC is implemented to predict the scores of unknown association pairs. Then, we rank the candidate diseases of each drug based on the predicted scores in a descending order. Finally, we choose four representative drugs, including Mitomycin, Vincristine, Fluorouracil and Leucovorin, to check their top candidates in comparative toxicogenomics database (CTD) (Davis *et al.*, 2013). Supplementary Table S2 lists the top 10 candidate indications for the four drugs and the bold font represents the corresponding association has been confirmed in CTD. We observe that 4–7 indications for each drug have been verified in the top 10. For example, vincristine (DB00541) is an antitumor vinca alkaloid isolated from Vinca Rosea, which is indicated for the treatment of small cell cancer of the lung, glioma susceptibility 1, lung cancer, kaposi sarcoma, osteogenic sarcoma, breast cancer and colorectal cancer. Additionally, mismatch repair cancer syndrome and chronic lymphocytic leukemia, ranked the top-1 and top-2, have not been confirmed. The two indications are treated by vincristine with a high probability and worthy of further study. In order to give more reliable references for medical researchers, we list the 11 drug–disease association pairs with predicted scores higher than 0.9 in Supplementary Table S3, which have not been validated by CTD.

4.6 The advantages of BMC and GRB for heterogeneous graph inference

In order to show the advantages of BMC and GRB for heterogeneous graph inference, we introduce BMC and GRB techniques separately into HGBI model, denoted BMC-HGBI and GRB-HGBI. For fair and distinct comparison, the same prior similarity information is used to HGBI, BMC-HGBI, GRB-HGBI and HGIMC. We employ two cases to present these comparisons, one is using single similarity information, such as chemical structures similarity of drug and phenotype similarity of disease, while the other is using the averaged multi-similarities of drug and disease.

Table 3 shows the results of various approaches under the single similarity case. As shown in Table 3, the AUC, AUPR and Precision values of BMC-HGBI achieve 0.908, 0.359 and 0.404 in the 10-fold cross-validation, which are 10.012%, 251.961%, 213.077% higher than that of HGBI, respectively. It illustrates that the new associations generated by BMC can exactly improve the prediction performance. However, in *de novo* experiment, BMC-HGBI has almost no positive effect on AUC, AUPR and Precision. This is because BMC cannot predict any meaningful values for zero row/column

vectors in the target matrix. In other word, BMC cannot identify valid associations for novel disease/drug nodes only using existing drug–disease associations. In addition, improving similarity measures is an effective way to handle *de novo*. For GRB-HGBI, the AUC, AUPR and Precision values are 0.818, 0.171 and 0.222 in *de novo*. Compared with the results of HGBI, GRB-HGBI has a great improvement in *de novo*. The result demonstrates GRB does enhance the similarity measures for HGBI. Finally, integrating BMC and GRB together, HGIMC can achieve better results in both the 10-fold cross-validation and *de novo* tests.

Table 4 lists the results of these compared methods under the averaged multi-similarities case. As shown in Table 4, compared with HGBI, the BMC-HGBI and GRB-HGBI algorithm have a great improvement in the 10-fold cross-validation and *de novo*. It illustrates that in the case of multiple similarities, both BMC and GRB can improve the prediction performance of HGBI. Moreover, the HGIMC algorithm outperforms BMC-HGBI and GRB-HGBI on AUPR and Precision for the 10-fold cross-validation and *de novo* tests.

4.7 Experiments on the other datasets

To demonstrate the flexibility and computation efficiency of HGIMC algorithm on different drug–disease datasets, we perform HGIMC algorithm on the other two datasets, including Cdataset (Luo *et al.*, 2016) and a new dataset we collected, namely Ydataset. Cdataset contained 2352 known drug–disease associations, where 663 drugs obtained in DrugBank and 409 diseases picked from OMIM database. Ydataset was obtained by integrating Cdataset and the latest CTD database. It included 1478 drugs, 655 diseases and 8448 validated drug–disease associations. The Supplementary Figure S2 shows a Venn diagram of association pairs among the gold standard dataset, Cdataset and Ydataset. We used the same way described in Section 2 to calculate the same five types of drug similarities and two kinds of disease similarities for Cdataset and Ydataset. Moreover, in Cdataset and Ydataset, the same value of the parameters (i.e. $\alpha = 10$ and $\beta = 10$) are used in the 10-fold cross-validation and *de novo* tests. In *de novo*, there are 177 and 479 drugs which have only one known associated disease in Cdataset and Ydataset, respectively.

Table 5 shows the AUC, AUPR and Precision values obtained by various approaches on Cdataset. As shown in Table 5, in the 10-fold cross-validation, HGIMC obtains the best AUC value of 0.953 and also has competitive results on AUPR and Precision. In *de novo*, HGIMC is 8.251%, 7.326%, 9.328%, 11.975% and 20.082% higher than BNNR, DRRS, MBiWR, DrugNet and HGBI on AUC, respectively. Meanwhile, our method is 52.332%, 40.157% better than the second best method (BNNR) on AUPR and Precision. The ROC curves of all compared approaches on Cdataset are displayed in Supplementary Figure S3.

Table 4. The compared results of HGBI, BMC-HGBI, GRB-HGBI and HGIMC in the 10-fold cross-validation and *de novo* tests under the averaged multi-similarities case

Tests	Metrics	HGBI	BMC-HGBI	GRB-HGBI	HGIMC
10-fold CV	AUC	0.878	0.911	0.923	**0.941**
	AUPR	0.265	0.384	0.286	**0.394**
	Precision	0.310	0.419	0.335	**0.438**
de novo	AUC	0.813	0.809	**0.884**	0.877
	AUPR	0.268	0.274	0.341	**0.356**
	Precision	0.310	0.333	0.409	**0.433**

Note: The best results are **bold** and the second best results are underlined.

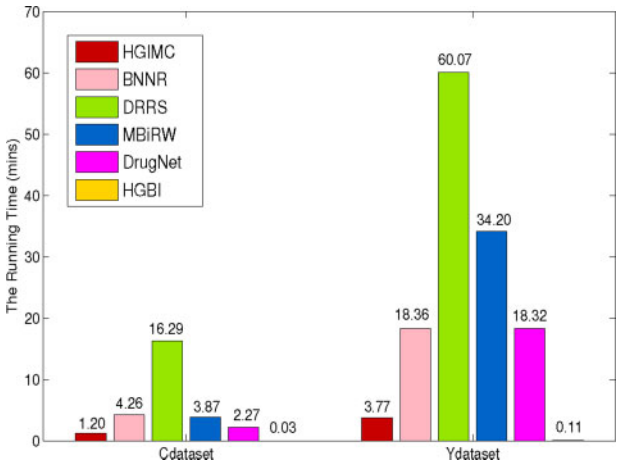

Fig. 5. The running time of a 10-fold cross-validation on Cdataset and Ydataset

Table 5. AUC, AUPR and precision values of all compared methods in the 10-fold cross-validation and *de novo* tests on Cdataset

Tests	Metrics	HGIMC	BNNR	DRRS	MBiRW	DrugNet	HGBI
10-fold CV	AUC	**0.953**	0.948	0.947	0.933	0.903	0.858
	AUPR	0.428	**0.441**	0.378	0.310	0.201	0.129
	Precision	0.463	**0.471**	0.403	0.351	0.239	0.168
de novo	AUC	**0.879**	0.812	0.819	0.804	0.785	0.732
	AUPR	**0.294**	0.193	0.181	0.174	0.106	0.075
	Precision	**0.356**	0.254	0.243	0.232	0.147	0.107

Note: The best results are **bold** and the second best results are underlined.

Table 6. AUC, AUPR and precision values of all compared methods in the 10-fold cross-validation and *de novo* tests on Ydataset

Tests	Metrics	HGIMC	BNNR	DRRS	MBiRW	DrugNet	HGBI
10-fold CV	AUC	0.956	**0.957**	0.956	0.912	0.910	0.864
	AUPR	**0.353**	0.229	0.241	0.145	0.122	0.069
	Precision	**0.386**	0.246	0.262	0.184	0.153	0.088
de novo	AUC	**0.913**	0.897	0.887	0.878	0.845	0.733
	AUPR	**0.174**	0.107	0.103	0.105	0.072	0.047
	Precision	**0.217**	0.129	0.121	0.146	0.088	0.069

Note: The best results are **bold** and the second best results are underlined.

The comparison results of all methods on Ydataset are given in Table 6. As shown in Table 6, in the 10-fold cross-validation, the AUC values of HGIMC, BNNR and DRRS greater than or equal to 0.956, but the AUPR and Precision of HGIMC is at least 46.473%, 47.328% higher than that of the two methods. In *de novo*, HGIMC outperforms other methods in terms of AUC, AUPR and Precision. The ROC curves of various methods on Ydataset are shown in Supplementary Figure S4.

In addition, Figure 5 illustrates the running time of various methods on Cdataset and Ydataset obtained under a 10-fold cross-validation. As shown in Figure 5, HGIMC has the least running time on the two datasets, except for HGBI. Although the HGBI algorithm faster than the other compared methods, its prediction performance is the worst. Actually, Ydataset is larger in scale than Cdataset. Compared with the running time on Cdataset, HGIMC, BNNR, DRRS, MBiRW and DrugNet algorithms increase the time on Ydataset by approximately 3, 15, 45, 30 and 16 min, respectively. It indicates that HGIMC is more suitable for large-scale prediction tasks.

5 Conclusions

In this study, we have proposed a new computational method named HGIMC for predicting potential drug-associated indications. In HGIMC, BMC can enrich the edges between drug network and disease network, and GRB can further enhance the similarity measures of drug and disease. With the cooperation of BMC and GRB, the performance of heterogeneous graph inference has been greatly improved in the 10-fold cross-validation and *de novo* tests. Moreover, HGIMC has an excellent computation efficiency. Additionally, case studies have confirmed the reliability of HGIMC on identifying new indications for known drugs.

The HGIMC algorithm can also be applied to other association prediction of biological entities, such as miRNA-disease associations (Chen *et al.*, 2018b), lncRNA-miRNA interactions (Fan *et al.*, 2020) and drug combination prediction (Chen *et al.*, 2016; Kim *et al.*, 2020; Zagidullin *et al.*, 2019). Specifically, drug combination can be regarded as a new way of drug repositioning. DrugComb (Kim *et al.*, 2020) provides an efficient utilization of drug combination resources, which is useful for predicting, testing and understanding drug combinations. For synergistic drug combination prediction, these are an assumption that principal drugs which obtain synergistic effect with similar adjuvant drugs are often similar and vice versa. It means drug combination matrix is also low-rank, which is the key condition for HGIMC algorithm to be effective. Using BMC and GRB in the principal and adjuvant drug networks, we believe HGIMC can perform well in drug combination prediction.

Funding

This work was supported by the National Natural Science Foundation of China [61972423], the Graduate Research Innovation Project of Hunan [CX20190125], Hunan Provincial Science and technology Program [2018wk4001] and 111Project [B18059].

Conflict of Interest: none declared.

References

Ada,H. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Aliper,A. *et al.* (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharmaceutics*, **13**, 2524–2530.

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Cai,J. *et al.* (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, **20**, 1956–1982.

Chen,X. *et al.* (2016) NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.*, **12**, e1004975.

Chen,X. *et al.* (2018a) Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*, **34**, 4256–4265.

Chen,X. *et al.* (2018b) MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.*, **14**, e1006418.

Chiang,A.P. and Butte,A.J. (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Therap.*, **86**, 507–510.

Chong,C. *et al.* (2007) New uses for old drugs. *Nature*, **448**, 645–646.

Davis,A. *et al.* (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.

Fan,Y. *et al.* (2020) Heterogeneous graph inference based on similarity network fusion for predicting lncRNA-miRNA interaction. *RSC Advances*, **10**, 11634–11642.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Huang,L. *et al.* (2020) Drug–drug similarity measure and its applications. *Brief. Bioinform.*, doi: 10.1093/bib/bbaa265.

Jaccard,P. (1908) Nouvelles recheres sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, **44**, 223–270.

Karaman,B. and Sippl,W. (2019) Computational drug repurposing: current trends. *Curr. Med. Chem.*, **26**, 5389–5409.

Kim,Y. *et al.* (2020) Anticancer drug synergy prediction in understudied tissues using transfer learning. *J. Am. Med. Inf. Assoc.*, doi: 10.1093/jamia/ocaa212.

Kuhn,M. *et al.* (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res.*, **44**, D1075–D1079.

Li,J. *et al.* (2016) A survey of current trends in computational drug repositioning. *Brief. Bioinform.*, **17**, 2–12.

Luo,H. *et al.* (2016) Drug repositioning based on comprehensive similarity measures and Bi-random walk algorithm. *Bioinformatics*, **32**, 2664–2671.

Luo,H. *et al.* (2018) Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, **34**, 1904–1912.

Luo,H. *et al.* (2020) Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief. Bioinform.*, doi: 10.1093/bib/bbz176.

Ma,S. *et al.* (2011) Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Program.*, **128**, 321–353.

Martinez,V. *et al.* (2015) DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.*, **63**, 41–49.

Napolitano,F. *et al.* (2013) Drug repositioning: a machine-learning approach through data integration. *J. Cheminf.*, **5**, 30.

Pushpakom,S. *et al.* (2019) Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, **18**, 41–58.

Ramlatchan,A. *et al.* (2018) A survey of matrix completion methods for recommendation systems. *Big Data Mining Anal.*, **1**, 308–323.

Resnik,P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada*, 448–453.

Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): An open-source java library for chemo-and bioinformatics. *Cheminformatics*, **34**, 493–500.

Tanoli,Z. *et al.* (2020) Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief. Bioinform.*, doi: 10.1093/bib/bbaa003.

Van,D. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wang,J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Wang,W. *et al.* (2013) Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput.*, **18**, 53–64.

Wang,B. *et al.* (2014a) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,W. *et al.* (2014b) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–2930.

Weininger,D. (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

Wishart,D. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–672.

Xuan,P. *et al.* (2019) Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*, **35**, 4108–4119.

Yang,J. and Yuan,X. (2012) Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.*, **82**, 301–329.

Yang,L. and Agarwal,P. (2011) Systematic drug repositioning based on clinical side-effects. *PLoS One*, **6**, e28025.

Yang,M. *et al.* (2019a) Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics (ISMB/ECCB 2019)*, **35**, i455–i463.

Yang,M. *et al.* (2019b) Overlap matrix completion for predicting drug-associated indications. *PLoS Comput. Biol.*, **15**, e1007541.

Yang,M. *et al.* (2020) Feature and nuclear norm minimization for matrix completion. *IEEE Trans. Knowl. Data Eng.*, doi: 10.1109/tkde.2020.3005978.

Zagidullin,B. *et al.* (2019) DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.*, **47**, W43–W51.

Zeng,X. *et al.* (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, **35**, 5191–5198.