frontiers in Genetics

ORIGINAL RESEARCH
published: 03 May 2021
doi: 10.3389/fgene.2021.666575

Predicting Drug-Disease Association Based on Ensemble Strategy

Jianlin Wang[1], *Wenxiu Wang*[1], *Chaokun Yan*[1]*, *Junwei Luo*[2]* *and Ge Zhang*[1]

[1] School of Computer and Information Engineering, Henan University, Kaifeng, China, [2] College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

Drug repositioning is used to find new uses for existing drugs, effectively shortening the drug research and development cycle and reducing costs and risks. A new model of drug repositioning based on ensemble learning is proposed. This work develops a novel computational drug repositioning approach called CMAF to discover potential drug-disease associations. First, for new drugs and diseases or unknown drug-disease pairs, based on their known neighbor information, an association probability can be obtained by implementing the weighted K nearest known neighbors (WKNKN) method and improving the drug-disease association information. Then, a new drug similarity network and new disease similarity network can be constructed. Three prediction models are applied and ensembled to enable the final association of drug-disease pairs based on improved drug-disease association information and the constructed similarity network. The experimental results demonstrate that the developed approach outperforms recent state-of-the-art prediction models. Case studies further confirm the predictive ability of the proposed method. Our proposed method can effectively improve the prediction results.

Keywords: drug repositioning, ensemble strategy, similarity measure, matrix completion, drug-disease association

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Bolin Chen,
Northwestern Polytechnical University, China
Wei Peng,
Kunming University of Science and Technology, China

***Correspondence:**
Chaokun Yan
ckyan@henu.edu.cn
Junwei Luo
luojunwei@hpu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *10 February 2021*
Accepted: *23 March 2021*
Published: *03 May 2021*

Citation:
Wang J, Wang W, Yan C, Luo J and Zhang G (2021) Predicting Drug-Disease Association Based on Ensemble Strategy. Front. Genet. 12:666575. doi: 10.3389/fgene.2021.666575

1. INTRODUCTION

Traditional drug discovery is a high-risk, high-investment, and long-term process (Li et al., 2015). It is well-known that it usually takes more than 10 years and more than $800 million to bring a new drug to market (Adams and Brantner, 2006). Additionally, the probability of drug approval success is below 10% (Ashburn and Thor, 2004). Considering the challenges of traditional drug discovery, the drug repositioning method is rising in popularity (Cano et al., 2017) and has attracted increasing interest from the research community and pharmaceutical industry (Shameer et al., 2015). Some successful repositioning drugs, such as duloxetine, sildenafil, and thalidomide, have generated high revenues in the history of their patent holders or companies (Ashburn and Thor, 2004).

The purpose of drug repositioning is to discover new indications for old drugs. Recently, many computational drug repositioning techniques, such as machine learning-based models, have been used to identify potential drug-disease interactions (Li et al., 2015). For example, Napolitano et al. (2013) melded drug-related features into a single information layer, which was used to train a multi-class support vector machine classifier whose output was a therapeutic class for a given drug. Chen and Li (2017) proposed the flexible and robust multiple-source learning (FRMSL) method to integrate multiple heterogeneous data sources to obtain drug-drug similarity and disease-disease similarity, and used the Kronecker regularized least squares (KronRLS) approach to solve the prediction problem. Liang et al. (2017) used Laplacian regularized sparse subspace learning to find

novel drug indications, integrating multiple pieces of information. Most machine learning-based models using negative samples are generated randomly from unknown associations, among which some false negatives may be included, resulting in a biased decision boundary (Liu et al., 2016a).

In recent years, with the rapid advance of high-throughput biology, huge amounts of multi-omic data have been yielded and several databases have been developed to store these valuable data (Chen et al., 2019; Luo et al., 2020). With the development of publicly available drug-related or disease-related databases, the network-based method is widely used in drug repositioning. The network-based method discovered potential drug–disease associations by propagating information in a heterogeneous biological network containing some information about diseases, drugs, or targets (Luo et al., 2018). For example, Yu et al. (2015) used drugs, protein complexes, and diseases to construct a tripartite network, which inferred the association probabilities of drug-disease pairs. Martìnez et al. (2015) developed DrugNet, a model for drug-disease and disease-drug prioritization; a network of interconnected drugs, proteins, and diseases was built, and DrugNet was used for drug repositioning. Luo et al. (2016) utilized drug- and disease-related properties to compute comprehensive similarity measures and the utility bi-random walk (BiRW) algorithm to find new uses for existing drugs. In recent years, the matrix factorization-based method has been successfully applied to biological association prediction, such as lncRNA-disease (Fu et al., 2017; Lan et al., 2020), drug-target (Liu et al., 2016b; Shi et al., 2018), and drug-disease (Zhang et al., 2018). The method can integrate prior information flexibly and integrate much information and many features into the framework to improve the accuracy of prediction. Zhang et al. (2018) developed a similarity-constrained matrix factorization approach (SCMFDD), which utilizes known drug-disease interactions, drug features, and disease features to predict potential drug-disease associations. Gönen and Kaski (2014) developed a new probabilistic method KBMF2MKL, which extended kernelized matrix factorization by incorporating multiple kernel learning. However, association prediction with matrix factorization has some limitations on the accuracy and prediction performance, especially for new diseases or drugs, which are called cold start problems. So, given different prediction approaches, an ensemble method is a promising way to combine their capacity in predicting the associations between drugs and diseases.

In this work, we develop a new drug repositioning model, CMAF, which integrates three methods (matrix factorization-based, label propagation-based, and network consistency projection-based methods) to obtain the final prediction result. To assess the performance of the developed approach, 10-fold cross-validation was implemented, and from the experimental results, we can see that ensemble models can combine different information to achieve high-accuracy performance. The experimental results demonstrate that CMAF obtained better results than the other four recent models in predicting potential drug-disease associations.

2. MATERIALS AND METHODS

In this section, we first introduce the gold standard dataset used in this study. Then, a proposed drug repositioning method named CMAF is presented to discover new uses for existing drugs. The overall flowchart of CMAF is shown in **Figure 1**, which contains the following three steps. First, the WKNKN algorithm is used as a preconditioning step to compute the temporary association score for new drugs and diseases or unknown drug-disease pairs. Second, a new drug-drug similarity network and a new disease-disease similarity network can be established. Third, three classical models are used to predict potential drug-disease associations separately, and their prediction results are ensembled to obtain the final association possibility of drug-disease pairs.

2.1. Dataset

The dataset used in this paper is curated manually from multiple biological datasets (Gottlieb et al., 2011). The dataset has 593 drugs and 313 diseases involving 1,933 validated drug-disease pairs. The drugs are collected from DrugBank (Wishart et al., 2006), and the diseases are extracted from Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2002).

The drug similarity is computed by the Chemical Development Kit (CDK) (Steinbeck et al., 2006) in terms of SMILES (Weininger, 1988) chemical structures, and the similarity between drug pairs is denoted as the Tanimoto score (Tanimoto, 1958) of their 2D chemical fingerprints. The disease similarity is computed using MimMiner (van Driel et al., 2006), which measures the similarity of two diseases by calculating the similarity between the MeSH terms (Lipscomb, 2000) present in the medical description information from the OMIM database.

2.2. Improved Drug-disease Association

A known drug-disease association Y can be modeled as a two-dimensional matrix, which has m drug rows and n disease columns, where each entry is denoted by Y_{ij}. The i-th row vector of the adjacency matrix Y, $Y(r_i) = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$, is the interaction profile for drug r_i. Similarly, the j-th column vector of the adjacency matrix Y, $Y(d_j) = (Y_{1j}, Y_{2j}, \ldots, Ymj)$, is the interaction profile for disease d_j.

It should be noted that the interaction profiles of new drugs or new diseases are all zero values. Additionally, many of the non-associations in Y are unobserved situations that could have potential interactions (i.e., false negatives). Therefore, we used WKNKN (Ezzat et al., 2017) to obtain the interaction likelihood value for non-associated drug-disease pairs in terms of their K nearest known neighbors [the K nearest known neighbors can be obtained by the K nearest neighbors (KNN) function according to their drug or disease similarity]. Here, we set $K = 5$. For every drug r_i, the similarity of its chemical structure with the K known drugs nearest to it and their corresponding values in the interaction profiles are utilized to obtain the interaction

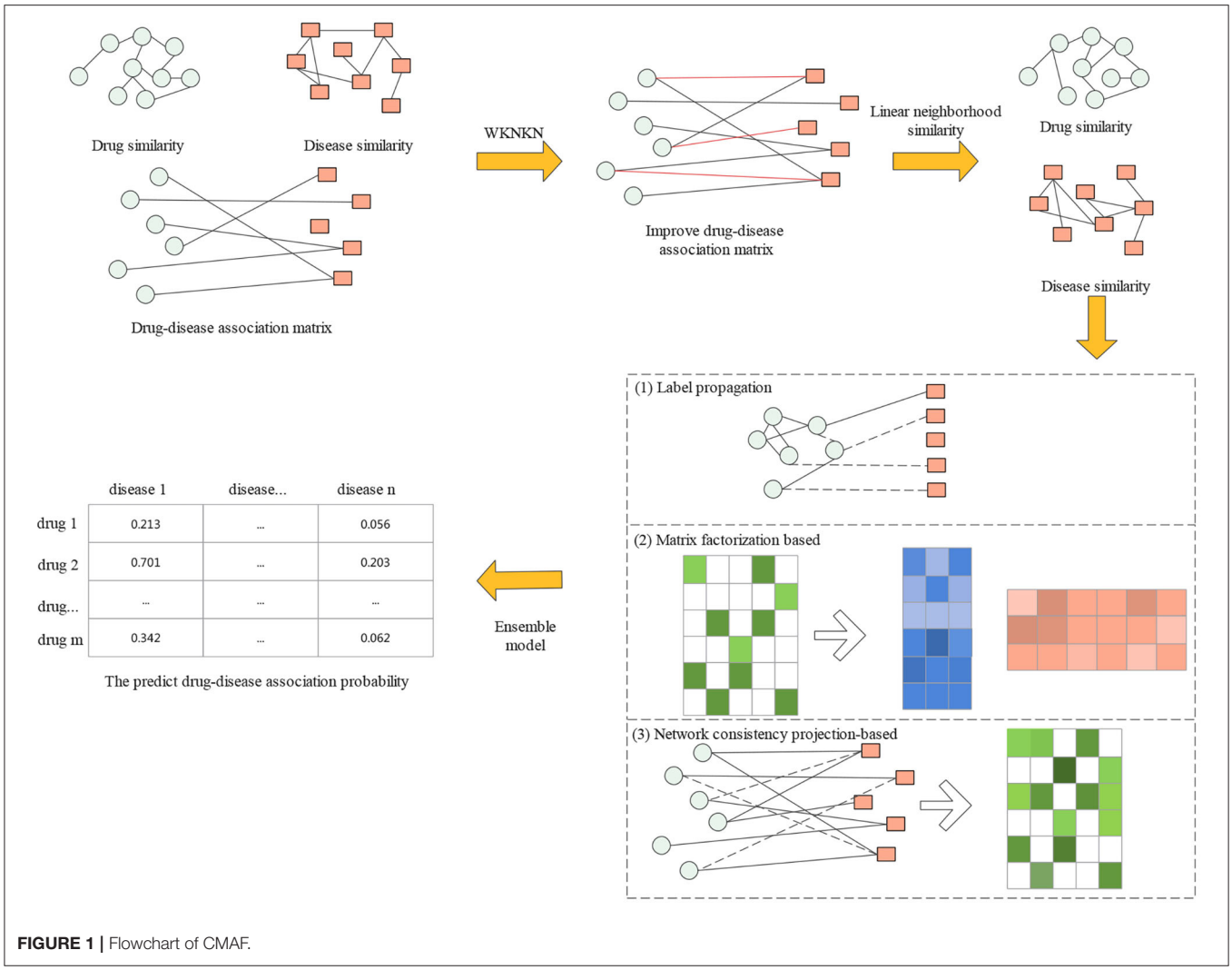

FIGURE 1 | Flowchart of CMAF.

likelihood profile of the drug r_i as follows:

$$Y_r(p) = \left(\sum_{i=1}^{K} w_i Y(r_i)\right) / Q_r \tag{1}$$

where r_i to r_k represent the K known nearest neighbors of drug r_p; the weight coefficient is $w_i = T^{i-1} S^r(r_i, r_p)$ where $T \leq 1$ is the decay term, and here, we set T to 0.5; and $S^r(r_i, r_p)$ is the similarity between r_i and r_p. Moreover, $Q_r = \sum_{i=1}^{K} S^r(r_i, r_p)$ is the normalization term. For the same reason, the interaction likelihood profile of disease d_j is as follows:

$$Y_d(q) = \left(\sum_{j=1}^{K} w_j Y(d_j)\right) / Q_d \tag{2}$$

where d_1 to d_k represent the K known nearest neighbors of disease d_q, the weight coefficient is $w_j = T^{j-1} S^d(d_j, d_q)$, the decay term T is 0.5, $S^d(d_j, d_q)$ is the similarity between d_j and d_q, and the normalization term is $Q_d = \sum_{j=1}^{K} S^d(d_j, d_q)$.

Then, we fuse Y_r and Y_d to replace $Y_{ij} = 0$ by taking the average of the two values mentioned above and denote it as Y_{rd}; we can then obtain a new adjacency matrix Y.

$$Y = max(Y, Y_{rd}) \tag{3}$$

where, $Y_{rd} = (Y_r + Y_d)/2$.

2.3. Improved Similarity of Drugs and Diseases

Similarity-based methods are widely used to find similar drugs (Vilar and Hripcsak, 2017). Some studies have shown that the use of similarity measures in drug repositioning often shows high predictive power (Azad et al., 2020). Therefore, similarity measurement is always regarded as an important step in drug repositioning research. The improvement of similarity can improve the prediction performance (Wang and Kurgan, 2019), reduce the computation cost, and make the similarity-based method more attractive and promising (Ding et al., 2014).

Relevant studies found that each data point can be linearly reconstructed from its neighborhood (Wang and Zhang, 2008),

we can calculate the pairwise drug similarity and pairwise disease similarity, which is the same method as in previous works (Zhang et al., 2017).

Here, we use drug data points as an example. Let x_i represent the feature vector of the i-th drug. The optimization problem is expressed as:

where $N(x_i)$ denotes the set of $K(0 < K < n)$ nearest neighbors. Here, we set K to 100.

$$\begin{aligned} \min_{\omega_i} \varepsilon_i &= \left\| x_i - \sum_{i_j : x_i \in N(x_i)} \omega_{i,i_j} x_{i_j} \right\|^2 \\ &= \sum_{i_j, i_k : x_{i_j}, x_{i_k} \in N(x_i)} \omega_{i,i_j} G^i_{i_j,i_k} \omega_{i,i_k} = \omega_i^T G^i \omega_i \\ \text{s.t. } & \sum_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} = 1, \omega_{i,i_j} \geq 0, j = 1, 2, \ldots, K \end{aligned} \tag{4}$$

$G^i_{i_j,i_l} = \left(x_i - x_{i_j}\right)^T \left(x_i - x_{i_l}\right)$. ω_{i,i_j} are the weights x_{i_j} for rebuilding x_i and can be seen as the similarity of x_i and x_{i_j}.

To avoid over-fitting, we add the regularization term for the rebuilt weight w_i and the objective function can be transformed as follows:

$$\begin{aligned} &\min_{\omega_i} \varepsilon_i = \omega_i^T G^i \omega_i + \lambda \left\| \omega_i \right\|^2 = \omega_i^T \left(G^i + \lambda I\right) \omega_i \\ &\text{s.t. } \sum_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} = 1, \omega_{i,i_j} \geq 0, j = 1, 2, \ldots, K \end{aligned} \tag{5}$$

where λ denotes the regularization parameter. Here, we set $\lambda = 1$.

We adopt standard quadratic programming to solve Equation (5), and its solution is called the *linear neighborhood similarity*. Here, a weight matrix W can be obtained, which we regard as the drug linear neighborhood similarity S^{r^*}.

Likewise, we can obtain the disease linear neighborhood similarity S^{d^*}.

2.4. Prediction Method

In this section, we use the drug linear neighborhood similarity and disease linear neighborhood similarity S^{d^*} to carry out three classical approaches to predict unobserved drug-disease interactions separately and ensemble their prediction results to obtain the final association possibility of drug-disease pairs.

2.4.1. Label Propagation

Label propagation (LP) methods perform the following task: given a weighted network, in which a small part of the nodes are labeled (with labels, such as positive), calculate the labels of the remaining unlabeled nodes (Zhang et al., 2015).

We formulate S^{d^*} as a directed graph, where drugs are nodes and the edge between drug r_i and drug r_j is weighted by the linear neighborhood similarity between the two drugs.

After constructing the graph, we utilize a label propagation approach to predict the unknown drug-disease pair association score (LPRIA). The known drug-disease associations are considered the initial node label information, and then the label information is updated. In each step, each drug node absorbs its neighbor's label information with probability α and maintains the initial state with probability $1 - \alpha$. Here, we set α as 0.5. The updated process can be written as:

$$Y_j^{t+1} = \alpha S^{r^*} Y_j^t + (1 - \alpha) Y_j^0 \tag{6}$$

where, Y_j^0 denotes the j-th column of the initial drug-disease interaction matrix Y (i.e., the initial states of all drugs for disease d_j). Furthermore, taking all diseases into account, the update process can be formulated in matrix form as:

$$Y^{t+1} = \alpha S^{r^*} Y^t + (1 - \alpha) Y^0 \tag{7}$$

Equation (7) will be used to update the label matrix until it converges, and Equation (7) will converge to:

$$Y^{r^*} = (1 - \alpha) \left(I - \alpha S^{r^*}\right)^{-1} Y^0 \tag{8}$$

where I represents the identity matrix and Y^{r^*} represents the predicted drug-disease pair probability from the drug side. For the convergence analysis of this update process, please refer to Wang and Zhang (2008).

Likewise, we constructed the label propagation approach from the disease side to obtain the predicted drug-disease interaction score matrix Y^{d^*}. The final association score Y^* is obtained according to the average of Y^{r^*} and Y^{d^*}.

2.4.2. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is an unsupervised model (Fujita et al., 2018). Its goal is to obtain two non-negative matrices and take their product as the optimal approximation to the original matrix. From the perspective of drug repositioning, the drug-disease association matrix $Y \in R^{m \times n}$ is factorized into two non-negative matrices, $W \in R^{m \times k}$ and $H \in R^{n \times k}$ ($k \ll \min(m, n)$), here, we set k to 100, and $Y \approx WH^T$.

To avoid over-fitting and increase the learning performance, Tikhonov and graph regularization terms are added to the standard NMF model to predict novel drug-disease pairs (NMFRIA). NMFRIA's objective function is as follows:

$$\begin{aligned} \min_{W,H} \left\| Y - WH^T \right\|_F^2 + \lambda_l \left(\|W\|_F^2 + \|H\|_F^2 \right) + \lambda_r \operatorname{Tr}\left(W^T L_r W\right) \\ + \lambda_d \operatorname{Tr}\left(H^T L_d H\right) \\ \text{s.t.} \quad W \geq 0, H \geq 0 \end{aligned} \tag{9}$$

where λ_l, λ_r, and λ_d represent the regularization coefficients; $Tr(\cdot)$ denotes the trace of a matrix, $L_r = D_r - S^{r^*}$ is the graph Laplacian matrix for the drug similarity matrices, S^{r^*} and $L_d = D_d - S^{d^*}$ are the graph Laplacian matrices for the disease similarity matrices S^{d^*} (Liu et al., 2014); and D_r and D_d represent the diagonal matrices whose entries are the row sums of S^{r^*} and S^{d^*}, respectively.

The method proposed by Xiao et al. (2018) is adopted to solve the minimization problem, and W and H are updated with an iterative equation. Here, the updating rules can be defined as:

$$w_{ik} \leftarrow w_{ik} \frac{\left(YH + \lambda_r S^{r^*} W\right)_{ik}}{\left(WH^T H + \lambda_l W + \lambda_r D_r W\right)_{ik}} \tag{10}$$

$$h_{jk} \leftarrow h_{jk} \frac{\left(Y^T W + \lambda_d S^{d^*} H\right)_{jk}}{\left(HW^T W + \lambda_l H + \lambda_d D_d H\right)_{jk}} \tag{11}$$

where w_{ik} represents the i-th row and the k-th column of non-negative matrix W, and h_{jk} represents the j-th row and the k-th column of non-negative matrix H.

According to Equations (10) and (11) the two non-negative matrices W and H are updated until convergence, and then we can obtain the predicted drug-disease interaction matrix as $Y^{**} = WH^T$. Here, we set λ_l to 2, and $\lambda_r = \lambda_d = 0.0001$.

2.4.3. Network Consistency Projection

Network consistency projection (NCP) considers drugs r_i that have a higher similarity to other drugs in the drug similarity matrix; the more drugs are associated with disease d_j, the higher the spatial similarity of drug r_i with disease d_j (and vice versa). Here, we use the NCP approach (Gu et al., 2016) for drug-disease association (NCPRIA) to obtain the predicted association scores between unknown drug-disease pairs.

NCPRIA computes the association probability between drug r_i and disease d_j by fusing two network consistency projection scores (the drug and disease space projection scores). Considering that unknown drug-disease pairs are not confirmed by experiment, which cannot prove that they are unrelated, and to prevent 0 from being the denominator, we replace 0 in the matrix Y with 10–30.

The drug space projection is the projection of the drug similarity network S^{r^*} on the drug-disease interaction network Y, which can be described as follows:

$$NCP\_R(i,j) = \frac{S^{r^*}(i,:)^* Y(:,j)}{|Y(:,j)|} \tag{12}$$

where $S^{r^*}(i,:)$ denotes the similarities between drug r_i and all other drugs in the i-th row of matrix S^{r^*} and $Y(:,j)$ denotes the associations between disease d_j and all drugs. $|Y(:,j)|$ represents the length of the vector $Y(:,j)$. $NCP\_R(i,j)$ represents the network consistency projection score of $S^{r^*}(i,:)$ on $Y(:,j)$. It is worth noting that the smaller the angle is between $S^{r^*}(i,:)$ and $Y(:,j)$, the more drugs are related to disease j and the more similar drugs there are to drug i, the larger the network consistency projection score $NCP\_R(i,j)$.

Similarly, we can obtain the disease space projection score as follows:

$$NCP\_D(i,j) = \frac{Y(i,:)^* S^{d^*}(:,j)}{|Y(i,:)|} \tag{13}$$

where $S^{d^*}(:,j)$ denotes the j-th column of matrix S^{d^*} and $Y(i,:)$ denotes the i-th row of drug-disease association Y. $NCP\_D(i,j)$ represents the network consistency projection score of $S^{d^*}(:,j)$ on $Y(i,:)$.

Finally, the projection score for the drug space and disease space are fused and normalized as follows:

$$Y^{***}(i,j) = \frac{NCP\_R(i,j) + NCP\_D(i,j)}{\left|S^{r^*}(i,:)\right| + \left|S^{d^*}(:,j)\right|} \tag{14}$$

where Y^{***} represents the predicted drug-disease association matrix and $Y^{***}(i,j)$ is the final predicted score of drug r_i and disease d_j.

2.4.4. Integrating the Prediction Results

According to the three aforementioned computational drug repositioning methods, to obtain better performance, a fusion model is adopted to integrate their predicted results, and the final prediction score between drugs and diseases is computed as follows:

$$Rt = 1 - \left(1 - Y^*\right)\left(1 - Y^{**}\right)\left(1 - Y^{***}\right) \tag{15}$$

In particular, Y^* is the predicted drug-disease association probability of the LPRIA method, Y^{**} is the predicted association probability of the NMFRIA method, Y^{***} is the predicted association probability of the NCPRIA method, and Rt stands for the final predicted drug-disease association probability.

3. EXPERIMENTS AND RESULTS

In this section, the performance of our approach, CMAF, is systematically evaluated. First, we describe the evaluation metrics. Based on a gold standard dataset, we compare our approach with several recent prediction algorithms and present the results in this section. In addition, the effectiveness of the developed method is further confirmed by case studies.

3.1. Evaluation Metrics

To evaluate the prediction performance of the proposed CMAF method, 10-fold cross-validation was conducted on the gold standard dataset. In each round of 10-fold cross-validation, all the recorded drug-disease pairs were randomly divided into 10 equal-sized parts. Each part was taken as a test set in turn, while the remaining nine parts of the data were merged as the training set, thus generating 10 pairs of training sets and test sets. To obtain convincing results, 10-fold cross-validation was repeated 10 times, and the average value of 10-folds was taken as the final result. After performing association prediction based on the training set, we can obtain the prediction values for each association. Then, for each drug, the test drug-disease associations are ranked together with all unconfirmed drug-disease pairs (candidate associations) in descending order according to the predicted values. For each specific ranking threshold, four metrics: true positive (TP), false negative (FN), false positive (FP), and true negative (TN), can be obtained based on the ranking results. If a test association has a higher rank value than the given threshold, it is considered as a correctly identified positive sample. Likewise, a candidate association is considered a correctly identified negative sample if it has a lower rank than the given threshold.

To provide an intuitive explanation of the evaluation metrics, a confusion matrix is first defined, which is built by comparing actual values with predicted outcomes. The two classes are constructed with positives and negatives, as shown in **Table 1**.

Next, the evaluation metrics of the true positive rate (TPR) and false positive rate (FPR) can be defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

Where TP and FP represent the numbers of correctly and wrongly identified positive samples and TN and FN represent the numbers of correctly and wrongly identified negative samples; TPR and FPR are calculated based on these four metrics. Furthermore, TPR is the ratio of known drug-disease pairs that are correctly predicted, and FPR is the proportion of unconfirmed drug-disease pairs that are predicted.

After that, the receiver operating characteristic (ROC) curve can be drawn based on TPR and FPR at different thresholds. Meanwhile, the area under ROC (AUC) can be calculated to evaluate the prediction performance. The larger the value of the AUC, the better the prediction performance. For instance, if the value of the AUC is equal to 1, it means the best performance.

3.2. Comparison With Other Methods

In this section, to evaluate the ability of the proposed approach, we compare CMAF with four other recently proposed computational drug repositioning approaches: NBI (Cheng et al., 2012), BNNR (Yang et al., 2019), HGBI (Wang et al., 2013), and NGRHMDA (Huang et al., 2017). NBI is based on a bipartite network and constructs a two-step diffusion model for drug repositioning (Cheng et al., 2012). BNNR was developed to utilize a bounded nuclear norm regularization approach to construct the drug-disease matrix under the low-rank assumption (Yang et al., 2019). HGBI was proposed according to the guilt-by-association principle and an intuitive interpretation of information flow on a heterogeneous graph (Wang et al., 2013). NGRHMDA uses neighbor-based collaborative filtering and a graph-based scoring method to obtain the association score (Huang et al., 2017). Although HGBI and NBI were originally used to predict potential drug-target associations and NGRHMDA was originally used to predict new microbe-disease associations, they can also be used to predict new drug-disease associations. The parameter values used in NBI, BNNR, HGBI, and NGRHMDA are set based on their corresponding literature.

The predictive ability of all drug repositioning approaches is evaluated in terms of the AUC specified in section 3.1. As shown in **Figure 2**, the results demonstrate that our developed approach, CMAF, is superior to the other four drug repositioning approaches. In detail, CMAF obtains an AUC value of 0.941, while BNNR, HGBI, NBI, and NGRHMDA achieve inferior results of 0.931, 0.832, 0.583, and 0.503, respectively.

TABLE 1 | Confusion matrix.

		Actual value	
		Positive	**Negative**
Predicted value	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

3.3. Comparison of the Three Methods With Their Combined Model

The effectiveness of the fusion method is evaluated in this section. We performed drug-disease association prediction on

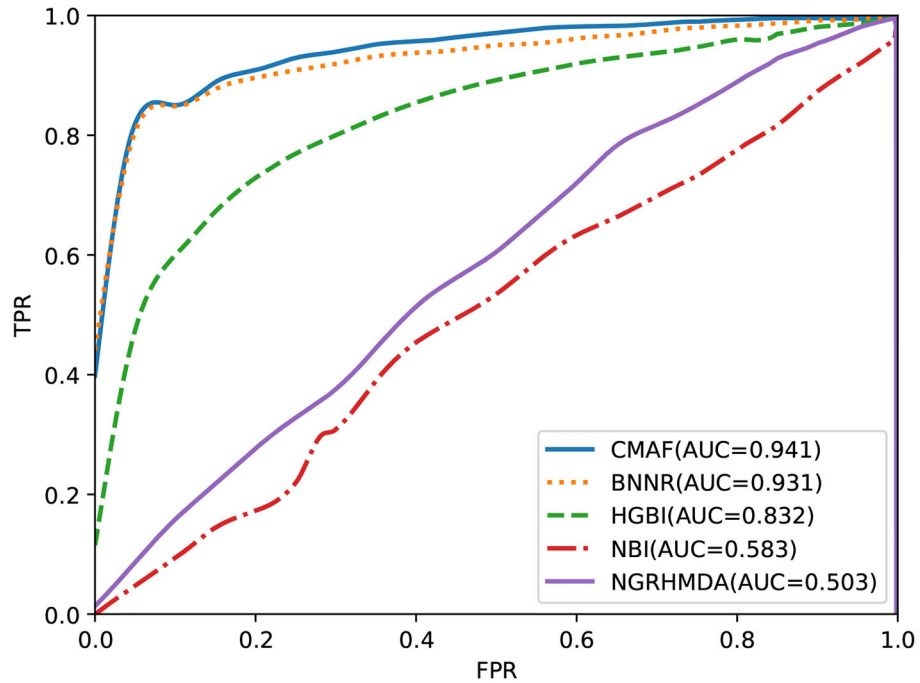

FIGURE 2 | Prediction results of various methods according to ROC curve analysis.

the gold standard dataset by using three methods (i.e., the LPRIA, NMFRIA, and NCPRIA methods) and their combined method. As shown in **Figure 3**, the AUC values of the three methods LPRIA, NMFRIA, and NCPRIA were 0.927, 0.923, and 0.920, respectively; however, the fusion method CMAF obtained an AUC value of 0.941. The experimental results

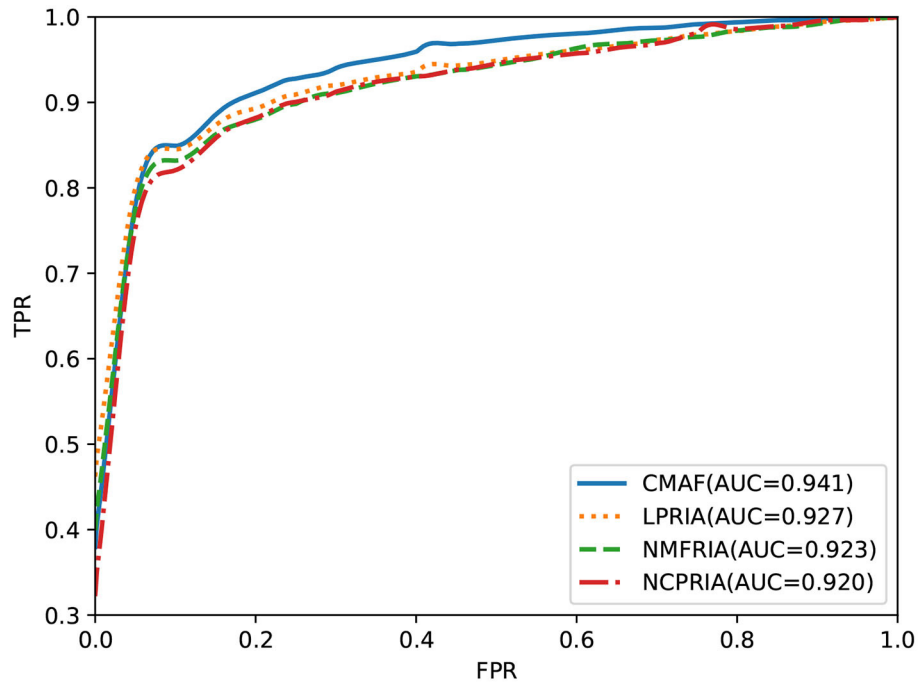

FIGURE 3 | Prediction performance of CMAF and the three individual methods according to the ROC curve.

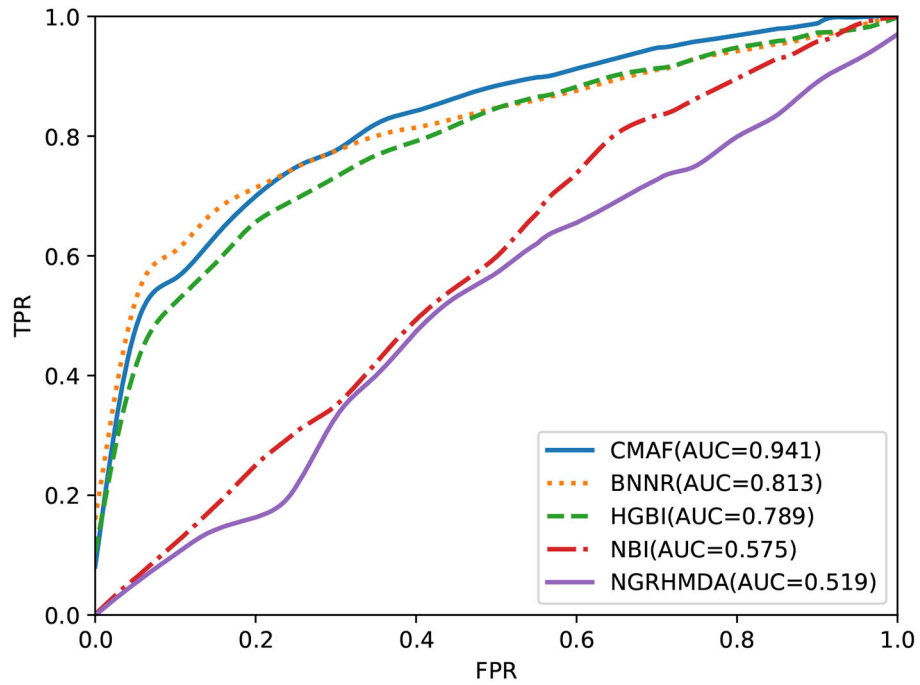

FIGURE 4 | Prediction performance of CMAF and the other four methods in predicting drug-disease associations for new drugs according to the ROC curves.

TABLE 2 | Case studies of four chosen drugs: levodopa, flecainide, zoledronic acid, and amantadine.

Drug (DrugBank IDs)	Top 5 candidate diseases (OMIM IDs)	Evidence
DB01235	168600	KEGG/DB/CTD
Levodopa	125320	DB/CTD
	165199	
	254770	
	190400	
DB01195	608583	CTD
Flecainide	194200	KEGG/CTD
	115000	DB/CTD
	157300	
	608622	CTD
DB00399	166710	KEGG/CTD
Zoledronic acid	102400	
	144700	CTD
	166300	
	114480	CTD
DB00915	168600	KEGG/DB/CTD
Amantadine	125320	DB/CTD
	605055	
	104300	CTD
	607225	

For each drug, the top five ranked predicted drugs are listed below.

illustrated the effectiveness of our fusion approach. Specifically, the CMAF method obtained the best performance among these four methods.

3.4. Prediction for New Drugs

To test the predictive performance of CMAF for new drugs, a de novo prediction test was executed. In de novo drug validation, for each of the drugs, we deleted all of its known associations, and they were used for testing samples in turn; the other known drug-disease association was used as the training sample. The rankings of the removed drug-disease associations relative to the drug candidate associations were obtained by *de novo* testing, which was used to assess the predictive performance. To compare the predictive ability of different methods in *de novo* testing of new drugs, the other four prediction methods also underwent de novo prediction tests. The experimental results are shown in **Figure 4**, and the graph demonstrates that our CMAF is superior to the other approaches. In detail, CMAF obtains an AUC value of 0.941, while the results of BNNR, HGBI, NBI, and NGRHMDA are 0.813, 0.789, 0.575, and 0.519, respectively.

4. CASE STUDIES

After verifying the predicted performance of CMAF in terms of 10-fold cross-validation, the ability of our proposed model to identify new indications for a given drug is further validated here. To predict new drug-disease interactions, all known drug-disease pairs are considered as the training set, and the remaining unknown drug-disease pairs form the candidate set. By applying our CMAF method, we can obtain all the candidates' set prediction scores. According to the prediction scores, for every drug, all the candidate diseases are ranked.

As an example, we selected some drugs and the corresponding top five candidate diseases as verified information, and then we found that some of them were confirmed in the KEGG (Kanehisa et al., 2013), DrugBank and CTD (Davis et al., 2014) databases, as shown in **Table 2**. For example, the effectiveness of levodopa in treating Parkinson's disease (PD) due to its ability to cross the blood-brain barrier can be retrieved from the KEGG, DrugBank, and CTD databases. In addition, relevant literature has shown that levodopa-treated patients have gained improvement in most Parkinsonian features in the past half-century (Lewitt, 2015). Flecainide is helpful for treating atrial fibrillation, as can be retrieved from CTD, and there is literature to prove that in clinical trials and real-world use, flecainide is more effective than other antiarrhythmic drugs (AADs) for the acute termination of recent-onset atrial fibrillation (Echt and Ruskin, 2020). From KEGG and CTD, zoledronic acid can be found to treat and prevent multiple forms of osteoporosis. There is also literature to prove that zoledronic acid administered once yearly for up to 3 years improved bone mineral density (BMD) at several skeletal sites, reduced fracture risk and bone turnover, and/or preserved bone structure and mass relative to placebo in clinical studies in patients with primary or secondary osteoporosis (Dhillon, 2016). Amantadine is an antiviral that can be used to cure PD and can be retrieved from KEGG, DB, and CTD. Relevant literature suggests that amantadine is an old antiviral compound that moderately ameliorates impaired motor behavior in Parkinson's disease (Müller et al., 2019).

5. CONCLUSION

This work proposed a new computational drug repositioning model named CMAF to find new uses for existing drugs. First, the number of known drug-disease interactions is far less than that of unknown drug-disease interactions in practice, which leads to the problem of data sparseness for drug repositioning. Therefore, we used the WKNKN method as a pre-processing step to compute the temporary association scores for these unknown drug-disease interactions in terms of their known neighbors, and then we computed the linear neighborhood similarity for drugs and diseases. After that, the LPRIA, NMFRIA, and NCPRIA methods were adopted to obtain three predictive association possibilities. Finally, we adopted an ensemble strategy to fuse these three prediction models to obtain the hopefully final prediction result. Compared with several recent computational drug repositioning models, our proposed CMAF approach achieves better predictive performance.

Even though our proposed method obtains promising results, it still has some limitations. First, we plan to consider integrating more predictive methods into the ensemble strategy. Second,

CMAF utilizes only single drug-drug similarity and disease-disease similarity to construct prediction methods. In the future, we will compute multiple drug-drug similarities and disease-disease similarities and combine diverse similarities to further improve the predictive performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CY and JW conceived and designed the approach. WW performed the experiments. JL analyzed the data. GZ and WW wrote the manuscript. CY and GZ supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61802113, 61802114, and 61972134), Science and Technology Development Plan Project of Henan Province (Nos. 202102210173 and 212102210091), China Post-doctoral Science Foundation (No. 2020M672212), and Henan Province Post-doctoral Research Project Founding.

ACKNOWLEDGMENTS

This paper was recommended by the 5th Computational Bioinformatics Conference.

REFERENCES

Adams, C. P., and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really $802 million? *Health Affairs* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683. doi: 10.1038/nrd1468

Azad, A. K. M., Dinarvand, M., Nematollahi, A., Swift, J., Lutze-Mann, L., and Vafaee, F. (2020). A comprehensive integrated drug similarity resource for *in-silico* drug repositioning and beyond. *Brief. Bioinform.* doi: 10.26434/chemrxiv.12376505.v1. [Epub ahead of print].

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., et al. (2017). Automatic selection of molecular descriptors using random forest: application to drug discovery. *Expert Syst. Appl.* 72, 151–159. doi: 10.1016/j.eswa.2016.12.008

Chen, H., and Li, J. (2017). "A flexible and robust multi-source learning algorithm for drug repositioning," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, ACM-BCB '17* (New York, NY: Association for Computing Machinery), 510–515. doi: 10.1145/3107411.3107473

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y. P. P., et al. (2019). ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2936476. [Epub ahead of print].

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8:e1002503. doi: 10.1371/journal.pcbi.1002503

Davis, A. P., Grondin, C., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B., et al. (2014). The comparative toxicogenomics database's 10th year anniversary: Update 2015. *Nucleic Acids Res.* 43, D914–D920. doi: 10.1093/nar/gku935

Dhillon, S. (2016). Zoledronic acid (Reclast(®), Aclasta(®)): a review in osteoporosis. *Drugs* 76, 1683–1697. doi: 10.1007/s40265-016-0662-4

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief. Bioinform.* 15, 734–747. doi: 10.1093/bib/bbt056

Echt, D., and Ruskin, J. (2020). Use of flecainide for the treatment of atrial fibrillation. *Am. J. Cardiol.* 125, 1123–1133. doi: 10.1016/j.amjcard.2019.12.041

Ezzat, A., Zhao, P., Wu, M., Li, X. L., and Kwoh, C. K. (2017). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Fu, G., Wang, J., Domeniconi, C., and Yu, G. X. (2017). Matrix factorization based data fusion for the prediction of lncrna-disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep.* 8:9743. doi: 10.1038/s41598-018-28066-w

Gönen, M., and Kaski, S. (2014). Kernelized bayesian matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 2047–2060. doi: 10.1109/TPAMI.2014.2313125

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7:496. doi: 10.1038/msb.2011.26

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6:36054. doi: 10.1038/srep36054

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. (2002). Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52–55. doi: 10.1093/nar/30.1.52

Huang, Y. A., You, Z. H., Huang, Z. A., Zhang, S., and Yan, G. (2017). Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15:209. doi: 10.1186/s12967-017-1304-7

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: lncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3034910. [Epub ahead of print].

Lewitt, P. (2015). Levodopa therapy for Parkinson's disease: pharmacokinetics and pharmacodynamics. *Mov. Disord.* 30, 64–72. doi: 10.1002/mds.26082

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2015). A survey of current trends in computational drug repositioning. *Brief. Bioinform.* 17, 2–12. doi: 10.1093/bib/bbv020

Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., et al. (2017). LRSSL: predict and interpret drug-isease associations based on data integration using sparse subspace learning. *Bioinformatics* 33, 1187–1196. doi: 10.1093/bioinformatics/btw770

Lipscomb, C. E. (2000). Medical subject headings (MESH). *Bull. Med. Libr. Assoc.* 88, 265–266.

Liu, H., Song, Y., Guan, J., Luo, L., and Zhuang, Z. (2016a). Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* 17:539. doi: 10.1186/s12859-016-1336-7

Liu, X., Zhai, D., Zhao, D., Zhai, G., and Gao, W. (2014). Progressive image denoising through hybrid graph laplacian regularization: a unified framework. *IEEE Trans. Image Process*. 23, 1491–1503. doi: 10.1109/TIP.2014.2303638

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X. (2016b). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol*. 12:e1004760. doi: 10.1371/journal.pcbi.1004760

Luo, H., Li, M., Wang, S., Liu, Q., Li, Y., and Wang, J. (2018). Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 34, 1904–1912. doi: 10.1093/bioinformatics/bty013

Luo, H., Li, M., Yang, M., Wu, F. X., Li, Y., and Wang, J. (2020). Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief. Bioinform*. 22, 1604–1619. doi: 10.1093/bib/bbz176

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 32, 2664–2671. doi: 10.1093/bioinformatics/btw228

Martìnez, V., Navarro, C., Cano, C., Fajardo, W., and Blanco, A. (2015). Drugnet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif. Intell. Med*. 63, 41–49. doi: 10.1016/j.artmed.2014.11.003

Müller, T., Kuhn, W., and Möhr, J. D. (2019). Evaluating ADS5102 (amantadine) for the treatment of Parkinson's disease patients with dyskinesia. *Expert Opin. Pharmacother*. 20, 1181–1187. doi: 10.1080/14656566.2019.1612365

Napolitano, F., Zhao, Y., M Moreira, V., and Tagliaferri, R. (2013). Drug repositioning: a machine-learning approach through data integration. *J. Cheminform*. 5:30. doi: 10.1186/1758-2946-5-30

Shameer, K., Readhead, B., and Dudley, J. T. (2015). Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Curr. Top. Med. Chem*. 15, 5–20. doi: 10.2174/1568026615666150112103510

Shi, J. Y., Zhang, A. Q., Zhang, S. W., Mao, K. T., and Yiu, S. M. (2018). A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization. *BMC Syst. Biol*. 12:136. doi: 10.1186/s12918-018-0663-x

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., and Willighagen, E. L. (2006). Recent developments of the chemistry development kit (CDK)–an open-source Java library for chemo-and bioinformatics. *Curr. Pharm. Des*. 12, 2111–2120. doi: 10.2174/138161206777585274

Tanimoto, T. T. (1958). *An Elementary Mathematical Theory of Classification and Prediction*. New York, NY: International Business Machines Corporation.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*. 14, 535–542. doi: 10.1038/sj.ejhg.5201585

Vilar, S., and Hripcsak, G. (2017). The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Brief. Bioinform*. 18, 670–681. doi: 10.1093/bib/bbw048

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. doi: 10.1093/bib/bby069

Wang, F., and Zhang, C. (2008). Label propagation through linear neighborhoods. *IEEE Trans. Knowl. Data Eng*. 20, 55–67. doi: 10.1109/TKDE.2007.190672

Wang, W., Yang, S., and Li, J. (2013). Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput*. 18, 53–64. doi: 10.1142/9789814447973_0006

Weininger, D. (1988). Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci*. 28, 31–36. doi: 10.1021/ci00057a005

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). Drugbank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res*. 34, D668–D672. doi: 10.1093/nar/gkj067

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microrna-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455–i463. doi: 10.1093/bioinformatics/btz331

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8:S2. doi: 10.1186/1755-8794-8-S2-S2

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2015). Label propagation prediction of drug-drug interactions based on clinical side effects. *Sci. Rep*. 5:12339. doi: 10.1038/srep12339

Zhang, W., Chen, Y., and Li, D. (2017). Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* 22:2056. doi: 10.3390/molecules22122056

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 19:233. doi: 10.1186/s12859-018-2220-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.