OXFORD

Briefings in Bioinformatics, 2022, 1–13
https://doi.org/10.1093/bib/bbab581
Problem Solving Protocol

A weighted bilinear neural collaborative filtering approach for drug repositioning

Yajie Meng, Changcheng Lu, Min Jin, Junlin Xu, Xiangxiang Zeng and Jialiang Yang

Corresponding authors. Min Jin, College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China. E-mail: jinmin@hnu.edu.cn; Xiangxiang Zeng, College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China. E-mail: xzeng@foxmail.com; Jialiang Yang, Geneis Beijing Co., Ltd, Beijing, 100102, China, E-mail: yangjl@geneis.cn

Abstract

Drug repositioning is an efficient and promising strategy for traditional drug discovery and development. Many research efforts are focused on utilizing deep-learning approaches based on a heterogeneous network for modeling complex drug–disease associations. Similar to traditional latent factor models, which directly factorize drug–disease associations, they assume the neighbors are independent of each other in the network and thus tend to be ineffective to capture localized information. In this study, we propose a novel neighborhood and neighborhood interaction-based neural collaborative filtering approach (called DRWBNCF) to infer novel potential drugs for diseases. Specifically, we first construct three networks, including the known drug–disease association network, the drug–drug similarity and disease–disease similarity networks (using the nearest neighbors). To take the advantage of localized information in the three networks, we then design an integration component by proposing a new weighted bilinear graph convolution operation to integrate the information of the known drug–disease association, the drug's and disease's neighborhood and neighborhood interactions into a unified representation. Lastly, we introduce a prediction component, which utilizes the multilayer perceptron optimized by the α-balanced focal loss function and graph regularization to model the complex drug–disease associations. Benchmarking comparisons on three datasets verified the effectiveness of DRWBNCF for drug repositioning. Importantly, the unknown drug–disease associations predicted by DRWBNCF were validated against clinical trials and three authoritative databases and we listed several new DRWBNCF-predicted potential drugs for breast cancer (e.g. valrubicin and teniposide) and small cell lung cancer (e.g. valrubicin and cytarabine).

Keywords: drug, disease, drug–disease association prediction, drug repositioning, neighborhood interactions

Introduction

Drugs are bioactive compounds that act on protein targets to cure/decelerate a specific disease or to promote the health of a living being [1]. Traditional *de novo* drug discovery has three steps: discovery stage, preclinical stage and clinical stage [2], which usually spans >10 years [3]. A recent study estimated that it costs $2.6 billion on average to develop a new drug approved by the Food and Drug Administration in 2015, as compared with $802 million in 2003 [4]. Biological experimental approaches pose considerable difficulties (e.g. time-consuming, costly and high-risk). Hence, repurposing of 'old' drugs to treat both common and rare diseases is becoming more and more attractive because it involves the use of de-risked compounds, with potentially lower overall development costs and shorter development timelines [5–7]. Computational drug repurposing narrows down the search space for drug–disease interactions by suggesting drug candidates for wet-lab validation [8]. There is a pressing need, therefore, for novel computational drug repurposing methodologies to facilitate drug discovery.

The drug repositioning problem can be modeled computationally as a recommendation system that recommends new indications based on known drug–disease associations. As the most typical drug repositioning method, matrix factorization projects drugs and diseases into a shared latent space, using a vector of latent features to represent a drug or a disease, and thereafter the drug–disease association is modeled as the inner product of their latent vectors. Matrix factorization and completion algorithms have been widely and successfully used in bioinformatics research, such as uncovering lncRNA–disease associations [9], predicting microRNA–disease associations [10–12], discovering potential anti-COVID-19 drugs [13, 14], identifying drug–drug interaction prediction [15], predicting drug side effects [16] and handling the dropouts problem by modeling single-cell

Yajie Meng is a doctoral student at Hunan University. Her research interests include bioinformatics and data mining.
Changcheng Lu is a graduate student at Hunan University. His research interests include bioinformatics and deep learning.
Min Jin is a professor at Hunan University. Her research interests include bioinformatics, artificial intelligence and data mining.
Junlin Xu is a doctoral student at Hunan University. His research interests include clustering, single cell, bioinformatics and computational biology.
Xiangxiang Zeng is a professor at Hunan University. His research interests include bio-computing and bioinformatics.
Jialiang Yang is the vice president of Geneis Beijing Co., Ltd. His research interests include bioinformatics, artificial intelligence and genomics.
Received: October 13, 2021. **Revised:** November 25, 2021. **Accepted:** December 19, 2021

RNA-sequencing imputation [17]. Many studies have suggested that matrix factorization and completion methods become promising computational strategies for drug repositioning [18–21]. For instance, Luo *et al.* presented a drug repositioning recommendation system (called DRRS) based on the singular value thresholding (SVT) algorithm to complete the large drug–disease adjacency matrix of a heterogeneous network, which integrated the disease–disease, drug–drug and drug–disease networks [18]. Zhang *et al.* proposed a similarity constrained matrix factorization method called SCMFDD, for the drug–disease association prediction [22]. Different from the conventional matrix factorization techniques, SCMFDD considers the biological context of the problem by introducing drug–drug feature-based similarity and disease–disease semantic similarity as constraints for drugs and diseases. In order to optimize the fusion process of multiple drug–drug and disease–disease similarities, Yang *et al.* developed a novel matrix factorization method for drug repositioning, called MSBMF. MSBMF concatenates multiple similarity matrices of drug and disease and decomposes the drug–disease association matrix into a drug-feature matrix and a disease-feature matrix, which are constrained by non-negative factorization. Zhang *et al.* designed a novel drug repositioning method by using Bayesian inductive matrix completion, termed DRIMC [23]. The aforementioned methods can be regarded as a linear multiplication of latent features. Although these methods have achieved strong performance, a deficiency is that they cannot effectively capture the complex structure of drug–disease association data and efficiently handle the high-complexity matrix operations on large-scale data.

To tackle the problem, some pioneering studies developed deep-learning-based models for drug repositioning, such as deepDR [24], LAGCN [25] and PADME [26]. Existing deep-learning techniques mainly constructed a heterogeneous network by using drug's and disease's side information and exploited deep-learning technologies to the heterogeneous network to learn better representation of drugs and diseases, which enhances the learning of drug–disease associations, and finally improves the prediction accuracy. Nevertheless, similar to the matrix factorization and completion algorithms, they generally utilize the global structure of the heterogeneous network and assume that the neighbors are independent of each other, i.e. considering all similar neighbors and ignoring the possible interactions between them. In some cases, the interactions between neighbor nodes could strengthen the target node's characteristics in a network. For example, an intuition in a transaction network is that a customer who has close business relations with rich friends would have a higher chance to repay a loan. Modeling such interactions between neighbors highlights the common properties, which could be helpful for the representation of the target node in a network. However, existing deep-learning models may be ineffective to capture the interactions between neighbors and thus lower the drug repositioning quality.

In this study, we proposed a novel drug repositioning approach based on weighted bilinear neural collaborative filtering, called DRWBNCF. To take advantage of localized topology information in different domains, we first constructed three networks, i.e. the known drug–disease association network, the drug–drug similarity network and the disease–disease similarity network to characterize nearest neighbors and their interactions information. Note that we utilized the neighborhood effects from most similar drugs and most diseases to create drug–drug and disease–disease similarity networks. In this way, our model only uses nearest neighbors instead of all similar neighbors, and thus gets more accurate results by filtering out noisy information. Then, we designed an integration component by proposing a new weighted bilinear graph convolution operation to encode the known drug–disease association together with the neighborhood and neighborhood interactions of the drug and disease. Finally, unlike the previous latent factor models that linearly factorize drug–disease associations, we introduced a prediction component, which uses the multi-layer perceptron (MLP) optimized by the α-balanced focal loss function and graph regularization to model the complex drug–disease associations. The basic concept of DRWBNCF is to incorporate both the known drug–disease association as well as neighborhood and neighborhood interactions in a unified MLP neural network. To evaluate the effectiveness of DRWBNCF, we compared it with four state-of-the-art methods over three real-world datasets of drug–disease associations. Experimental results show that DRWBNCF achieves the best performance for drug repositioning.

Materials and Methods

Datasets

In this study, we adopted three real-world datasets that are used in previous works to evaluate the effectiveness of DRWBNCF. The first one is Fdataset [18, 27], which is corresponding to the gold standard drug–disease dataset reported in the work of Gottlieb *et al.* [28]. It includes 1933 known drug–disease associations between 593 drugs collected in DrugBank database [29] and 313 diseases obtained in OMIM database [30]. The second one is Cdataset [31], containing 2532 known drug–disease associations between 663 drugs extracted from DrugBank database and 409 diseases listed in the OMIM database. The third one is LRSSL, which includes 763 drugs taken from DrugBank database, 681 diseases obtained in MeSH database and 3051 validated drug–disease associations totally [32]. The simple statistics for the three datasets are shown in Table 1.

In our study, the pairwise drug similarity is calculated based on the chemical structure of simplified molecular input line entry system (SMILES) format [33], and is

Table 1. Details of the three benchmark datasets

Datasets	No. of drugs	No. of diseases	The known associations
Fdataset	593	313	1933
Cdataset	663	409	2532
LRSSL	763	681	3051

represented as the Tanimoto index of chemical fingerprints of the drug pair via the Chemical Development Kit [34]. The pairwise disease similarity is measured based on the semantic similarity of disease phenotypes via the text mining analysis of medical descriptions information of the disease pair.

The construction of three networks

We construct three networks, including the known drug–disease association network, drug–drug similarity network and disease–disease similarity network. Here, we denote the known drug–disease association network G by a binary matrix $A \in \mathbb{R}^{n*m}$, where each entry $A_{ij} \in \{0, 1\}$, n and m are the number of drugs and diseases, respectively. If the drug r_i has been experimentally confirmed to associate with the disease d_j, $A_{ij} = 1$, otherwise, $A_{ij} = 0$.

The drug–drug similarity network G^r is represented by the matrix $A^r \in \mathbb{R}^{n*n}$, where each entry of A^r is constructed based on the similarity of each pair of drugs. These similarities of drugs are denoted by a $n \times n$ matrix $\mathbf{S}^r$, where the (i, j) entry $\mathbf{S}^r(i, j)$ is the similarity between the drug r_i and the drug r_j. In order to get more accurate results by avoiding noisy information, our model only exploits$k$-nearest neighbors instead of all similar neighbors as considered in previous studies. The entry A^r_{ij} of A^r is defined as

$$A^r_{ij} = \begin{cases} \mathbf{S}^r(i,j) & \text{if } r_j \in \tilde{N}_k(r_i) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathbf{S}^r(i, j)$ is the drug similarity matrix, $\tilde{N}_k(r_i) = \{r_i\} \cup N_k(r_i)$ is a set of r_i's extended k-nearest neighbors including r_i, and $N_k(r_i)$ is the k-nearest neighbors of drug r_i. In the same way, we denote the disease–disease similarity network G^d by the matrix $A^d \in \mathbb{R}^{m*m}$. The entry A^d_{ij} of matrix A^d is defined as

$$A^d_{ij} = \begin{cases} \mathbf{S}^d(i,j) & \text{if } d_j \in \tilde{N}_k(d_i) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{S}^d$ is the disease similarities matrix, the (i, j) entry $\mathbf{S}^d(i, j)$ of matrix $\mathbf{S}^d$ is the similarity between the disease d_i and the disease d_j, $\tilde{N}_k(d_i) = \{d_i\} \cup N_k(d_i)$ is a set of d_i's extended k-nearest neighbors including disease d_i, and $N_k(d_i)$ is the k-nearest neighbors of disease d_i.

For convenience, we represent the degree of drug r by d_r, i.e. $d_r = |N(r)|$, where $|N(r)|$ is the number of the drug r's all neighbors and accordingly $\tilde{d}_r = |\tilde{N}(r)| = d_r + 1$. Similarly, we represent the degree of disease d by d_d, i.e. $d_d = |N(d)|$, where $|N(d)|$ is the number of the disease d's all neighbors, and accordingly $\tilde{d}_d = |\tilde{N}(d)| = d_d + 1$.

Model architecture

Problem formulation

Assume that we have a set of drugs and diseases, denoted by $\mathcal{R}$ and $\mathcal{D}$ respectively. Let$r \in \mathcal{R}$ denote a drug and $d \in \mathcal{D}$ denote a disease, and $\mathcal{Y}_{r,d}$ be the association label between r and d. If r has associated with d, $\mathcal{Y}_{r,d} = 1$, otherwise $\mathcal{Y}_{r,d} = 0$. We call a drug–disease pair $<r, d>$ an association pair when $\mathcal{Y}_{r,d} = 1$. The drug repositioning can be defined as a prediction problem that infers the value of the association label $\mathcal{Y}_{r,d}$ based on association pairs.

The conventional graph convolution operation

To date, existing efforts on GNN have largely defined the graph convolution operations as a weighted sum (i.e. linear aggregation) over features of the neighbors of the target node to form the representation of the target node [35]. To encode the neighbors information of the drug r and the disease d by the conventional graph convolution operation, the representations of the drug r and the disease d are as follows:

$$\mathbf{h}_r = AGG\left(\{\mathbf{h}_i\}_{i \in \tilde{N}_k(r)}\right) = \sum_{i \in \tilde{N}_k(r)} \alpha_{ri} \mathbf{h}_i \mathbf{W}_r \quad (3)$$

$$\mathbf{h}_d = AGG\left(\{\mathbf{h}_i\}_{i \in \tilde{N}_k(d)}\right) = \sum_{i \in \tilde{N}_k(d)} \alpha_{di} \mathbf{h}_i \mathbf{W}_d \quad (4)$$

where $\mathbf{h}_r$ is the representation of neighbors information of the drug r, $\mathbf{h}_d$ is the representation of neighbors information of the disease d, $AGG(\cdot)$ is the linear aggregator, α_{xi} is the weight of neighbor i and is defined as $\frac{1}{\sqrt{\tilde{d}_x \tilde{d}_i}}$, i is the index from the extended k-nearest neighbors $\tilde{N}_k(x)$, and $\mathbf{W}_r \in \mathbb{R}^{|\mathcal{R}|*K_1}$ and $\mathbf{W}_d \in \mathbb{R}^{|\mathcal{D}|*K_2}$ is the weight matrix.

Although the conventional graph convolution operation improves the representation of the target node, it is built upon a summation operation and naturally assumes that the neighbors are independent and ignores the interactions among neighbors. Most existing graph convolution operations forgo the importance of interactions among neighbors. When such interactions exist, such as the co-occurrence of two neighbor nodes is a strong signal of the target node's characteristics,

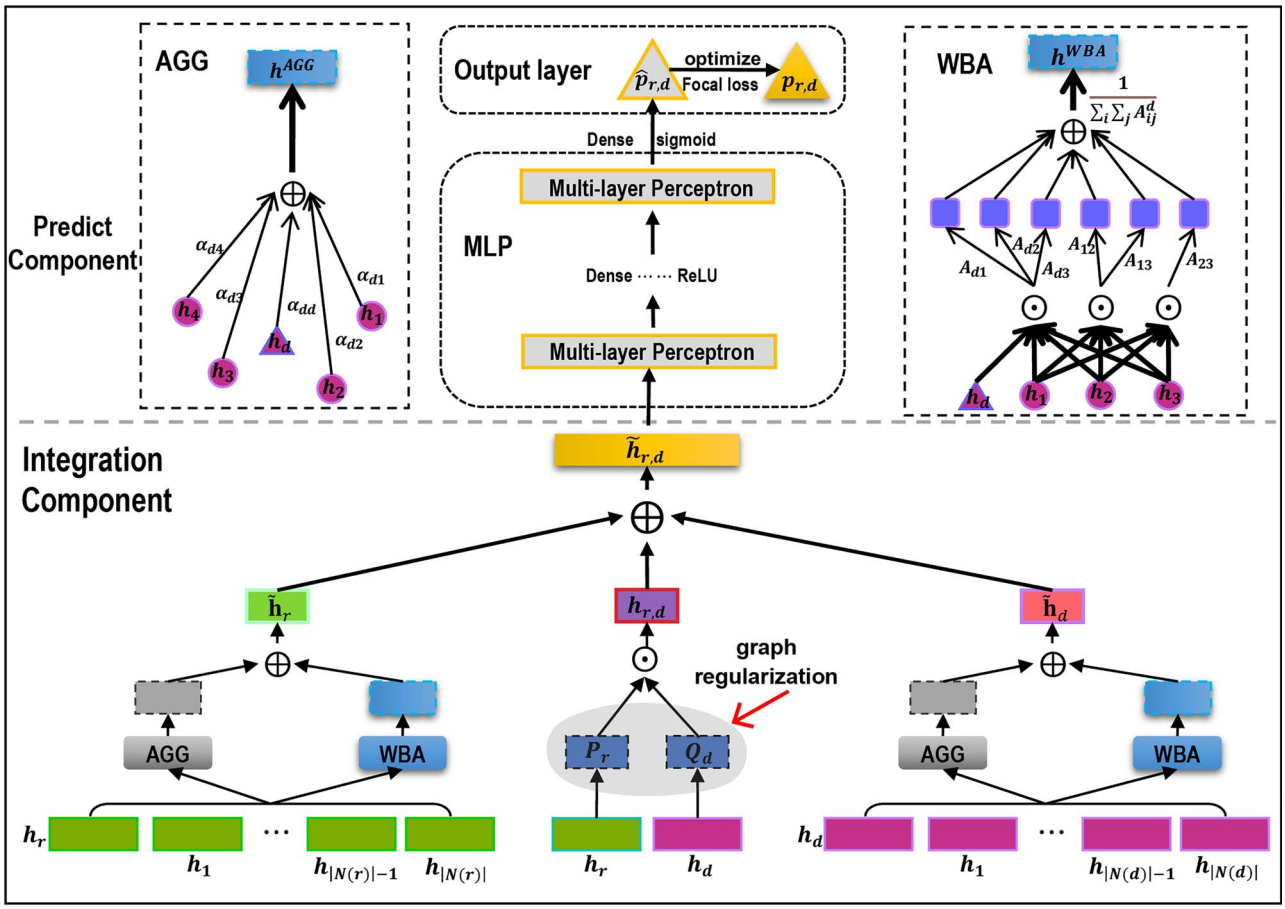

Figure 1. The overall architecture of DRWBNCF. DRWBNCF involves two components: (i) an integration component, which describes how to encode the known drug–disease association together with neighborhood and neighborhood interactions of the drug and disease; (ii) a prediction component, which utilizes the multi-layer perceptron optimized by the α-balanced focal loss function and graph regularization to model the complex drug–disease associations.

existing graph convolution operation may fail to capture the signal.

Our proposed DRWBNCF model

In this section, the details related our proposed models are described. The workflow of DRWBNCF is briefly shown in Figure 1.

Encoding the drug–disease known association

For a given drug–disease pair <r, d>, we represent the original feature matrix of drugs and diseases by $\mathbf{S}^r$ and $\mathbf{S}^d$, that encode the involved drug r and disease d with a vector $s_r \in \mathbb{R}^{1\times|\mathcal{R}|}$ and $s_d \in \mathbb{R}^{1\times|\mathcal{D}|}$ (a row), respectively. By applying two projection matrices $\mathbf{P} \in \mathbb{R}^{|\mathcal{R}|\times K_3}$ and $\mathbf{Q} \in \mathbb{R}^{|\mathcal{D}|\times K_3}$, the encoded drug r or disease d (i.e. s_r or s_d) vector will be transformed into a latent vector as below:

$$\mathbf{p}_r = s_r \cdot \mathbf{P}, \mathbf{q}_d = s_d \cdot \mathbf{Q} \tag{5}$$

where $\mathbf{p}_r \in \mathbb{R}^{1\times K_3}$ and $\mathbf{q}_d \in \mathbb{R}^{1\times K_3}$ are the drug's and disease's latent features vectors, respectively. To effectively capture the overall structure of drug–disease association, we adopt the generalized matrix factorization method and define the drug–disease association function $\varnothing(\mathbf{p}_r, \mathbf{q}_d)$ as follows:

$$\mathbf{h}_{r,d} = \varnothing(\mathbf{p}_r, \mathbf{q}_d) = \mathbf{p}_r \odot \mathbf{q}_d \tag{6}$$

where $\mathbf{h}_{r,d} \in \mathbb{R}^{1\times K_3}$ is the latent vector for association information of drug–disease pair <r, d>, "$\odot$" denotes the element-wise product.

Encoding neighborhood and neighborhood interactions

Inspired by the bilinear aggregator [36], To learn the characteristics representation of drug r more comprehensively, we consider all pairwise weighted interactions between the extended k-nearest neighbors (including the drug r), and define a weighted bilinear aggregator (WBA):

$$\mathbf{h}_r = WBA\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(r)}\right) = \frac{1}{\sum_i\sum_j A^r_{ij}} \sum_{i\in\tilde{N}_k(r)} \sum_{j\in\tilde{N}_k(r)\&i<j} \times \left(\mathbf{h}_i\mathbf{U}_r \odot \mathbf{h}_j\mathbf{U}_r\right)\cdot A^r_{ij} \tag{7}$$

where $\mathbf{h}_r$ is the latent vector for the weighted interactions information between neighbors of drug r, i and j are the indexes from the extended k-nearest neighbors $\tilde{N}_k(r)$—they are constrained to be different to avoid self-interactions that are meaningless and may even introduce additional noises, and $\mathbf{U}_r \in \mathbb{R}^{|\mathcal{R}|*K_1}$ is the weight matrix. We denote the interaction weight between i and j by A^r_{ij}. Our proposed WBA aggregates all pair-wise weighted interactions between neighbors and is capable of capture the complex patterns of information propagation, which are hard to reveal by the conventional graph convolution operation individually.

Similarly,

$$\mathbf{h}_d = WBA\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(d)}\right) = \frac{1}{\sum_i\sum_j A^d_{ij}} \sum_{i\in\tilde{N}_k(d)} \sum_{j\in\tilde{N}_k(d)\&i<j} \times \left(\mathbf{h}_i\mathbf{U}_d \odot \mathbf{h}_j\mathbf{U}_d\right)\cdot A^d_{ij} \tag{8}$$

where $\mathbf{h}_d$ is the latent vector for weighted interactions information between the extended k-nearest neighbors of disease d, and $\mathbf{U}_d \in \mathbb{R}^{|\mathcal{D}|*K_2}$ is the weight matrix.

In order to enhance the characteristics representations of drug r and disease d, we further adopt a linear combination scheme consisting of the conventional graph convolution operator (i.e. weighted sum) and our proposed WBA, formally defining a new graph convolution operator as follows:

$$\tilde{\mathbf{h}}_r = \lambda\cdot\sigma\left(WBA\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(r)}\right)\right) + (1-\lambda) \cdot\sigma\left(AGG\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(r)}\right)\right) \tag{9}$$

$$\tilde{\mathbf{h}}_d = \lambda\cdot\sigma\left(WBA\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(d)}\right)\right) + (1-\lambda) \cdot\sigma\left(AGG\left(\{\mathbf{h}_i\}_{i\in\tilde{N}_k(d)}\right)\right) \tag{10}$$

where $\tilde{\mathbf{h}}_r \in \mathbb{R}^{1\times K_1}$ and $\tilde{\mathbf{h}}_d \in \mathbb{R}^{1\times K_2}$ are the latent vectors for the neighborhood and neighborhood interactions information of the drug r and the disease d, respectively, $\sigma(\cdot)$ is a nonlinear activation function, $WBA(\cdot)$ and $AGG(\cdot)$ denote our proposed weighted bilinear and the traditional linear aggregators respectively, λ is a hyperparameter, which measures the WBA and traditional linear aggregator. The representations of drug r and disease d are initialized to $\mathbf{s}_r$ and $\mathbf{s}_d$, respectively.

By Equations (3), (4), (7) and (8), Equations (9) and (10) are modified as follows:

$$\tilde{\mathbf{h}}_r = \lambda\sigma\left(\frac{1}{\sum_i\sum_j A^r_{ij}} \sum_{i\in\tilde{N}_k(r)} \sum_{j\in\tilde{N}_k(r)\&i<j} \left(\mathbf{h}_i\mathbf{U}_r \odot \mathbf{h}_j\mathbf{U}_r\right)\cdot A^r_{ij}\right) + (1-\lambda)\,\sigma\left(\sum_{i\in\tilde{N}_k(r)} \alpha_{ri}\mathbf{h}_i\mathbf{W}_r\right) \tag{11}$$

$$\tilde{\mathbf{h}}_d = \lambda\sigma\left(\frac{1}{\sum_i\sum_j A^d_{ij}} \sum_{i\in\tilde{N}_k(d)} \sum_{j\in\tilde{N}_k(d)\&i<j} \left(\mathbf{h}_i\mathbf{U}_d \odot \mathbf{h}_j\mathbf{U}_d\right)\cdot A^d_{ij}\right) + (1-\lambda)\,\sigma\left(\sum_{i\in\tilde{N}_k(d)} \alpha_{di}\mathbf{h}_i\mathbf{W}_d\right) \tag{12}$$

Integrating the known association with neighbors and weighted interactions between neighbors

Once we have obtained the representations for the drug–disease known association and the drug's and disease's neighborhood and neighborhood interactions, we further integrate these latent vectors into a unified representation as below:

$$\tilde{\mathbf{h}}_{r,d} = \mathbf{h}_{r,d} \oplus \tilde{\mathbf{h}}_r \oplus \tilde{\mathbf{h}}_d \tag{13}$$

where "$\oplus$" denotes the vector concatenation operation.

MLP-based prediction

The association of a drug–disease pair can be very complex. Previous approaches usually assume a linear relation by decomposing the drug–disease matrix, e.g. standard matrix factorization. Since they directly factorize drug–disease associations, it is weak in identifying strong associations among a set of closely related drugs or diseases, especially when data are highly sparse. We would like to endow our model with a higher level of flexibility and nonlinearity to better characterize the association with the incorporation of neighborhood and neighborhood interactions. Hence, we apply the MLP to predict drug–disease associations. Generally, an MLP component can be constructed layer by layer. For $j = 1, \ldots, L$, the MLP component under our DRWBNCF model is defined as:

$$\mathbf{z}_1 = f^{(1)}\left(\mathbf{W}_1^T z_0 + \mathbf{b}_1\right),$$

$$\mathbf{z}_2 = f^{(2)}\left(\mathbf{W}_2^T z_1 + \mathbf{b}_2\right),$$

$$\mathbf{z}_L = f^{(L)}\left(\mathbf{W}_L^T z_{L-1} + \mathbf{b}_L\right),$$

$$\hat{p}_{r,d} = sigmoid\left(\mathbf{v}^T\mathbf{z}_L\right) \tag{14}$$

where $f^{(x)}(\cdot)$ is the nonlinear activation function for the xth layer. We choose rectifier linear unit (ReLU) as the activation function to avoid the oversaturation. $\mathbf{W}_x$ and $\mathbf{b}_x$ represent the weight matrix and the bias vector, respectively. To feed the MLP, we set $\mathbf{z}_0 = \tilde{\mathbf{h}}_{r,d}$ (defined in Equation (13)). For the output layer, we utilized the sigmoid function as activation function to limit the output of our model in range of 0 to 1. $\mathbf{v}^T$ is the weights and $sigmoid(x) = \frac{1}{1+\exp(-x)}$. $\hat{p}_{r,d}$ is the conditional probability of the association class being 1.

Optimization

We take drug–disease association pairs as positive samples $\mathcal{Y}^+$ and take other pairs (i.e. unobserved or unknown drug–disease pairs) as negative samples $\mathcal{Y}^-$. Herein, that known drug–disease association pairs have been validated manually, which are highly reliable and important for improving prediction performance. However, the number of positive samples is far less than that of negative samples. This extreme class imbalance causes two problems: (i) training is inefficient because easily classified negatives involve the majority of the loss that contribute no useful learning information; (ii) the easy negative samples can overwhelm training and lead to poor performance of models. A common method for addressing the extreme class imbalance is to introduce a weighting factor $\alpha \in [0, 1]$ for class $\mathcal{Y} = 1$ and $1-\alpha$ for class $\mathcal{Y} = -1$, such as the weighted binary cross-entropy loss function. Although α balances the importance of positive and negative samples, it does not differentiate between easy negative samples and hard positive samples. Here,

we adopt α-balanced focal loss function [37] and graph regularization terms as the DRWBNCF's objective function:

$$\begin{aligned}\text{Loss} = -\frac{1}{n \times m}\Bigg(\alpha \times \sum_{(r,d)\in\mathcal{Y}^+}\left(1-\hat{p}_{r,d}\right)^{r}\cdot\log\hat{p}_{r,d} \\ +(1-\alpha)\times\sum_{(r,d)\in\mathcal{Y}^-}\hat{p}^{r}_{r,d}\log\left(1-\hat{p}_{r,d}\right)\Bigg) \\ +\lambda_r Tr\left(P^T\mathcal{L}_r P\right)++\lambda_d Tr\left(Q^T\mathcal{L}_d Q\right)\end{aligned} \tag{15}$$

where α is a weighting factor for class 1, λ_r and λ_d control the regularization strength to avoid overfitting, $\hat{p}_{r,d}$ (defined in Equation (13)) is the conditional probability, $Tr(\cdot)$ is the trace of a matrix, $\mathcal{L}_r = D^r - A^r$ and $\mathcal{L}_d = D^d - A^d$ are the graph Laplacians for A^r and A^d, respectively, and $D^r_{ii} = \sum_s A^r_{is}$ and $D^d_{jj} = \sum_t A^d_{jt}$ are diagonal matrices. We optimized the objective function in Equation (15) through the Adam optimizer [38] and initialized parameters as described in [39]. DRWBNCF's objective function balances several goals. The first term (i.e. The α-balanced focal loss function) reshapes the standard cross-entropy loss function, and focuses training on the sparse set of hard positives, and thus down-weights the loss assigned to the vast number of easy negatives. The second term is for drug graph regularization, which minimizes the distance between latent feature vectors of two neighboring drugs. The third term is for disease graph regularization.

Experimental

Evaluation metrics

To evaluate the performances of DRWBNCF, we adopted 10-fold cross-validation (10-CV) to predict the drug–disease association. Specifically, both the known drug–disease associations and the unobserved drug–disease associations were randomly divided into 10 equal-sized folds. Each fold of the known associations and the unobserved associations was regarded as the testing samples in turn, while the remaining 9-folds were used as the training samples to train our proposed model. Notably, the known associations were taken as positive samples and the unobserved associations were taken as negative samples. After all associations have been tested, we calculate both True-Positive Rate (TPR), False-Positive Rate (FPR) and Precision as follows:

$$TPR\ \ (or\ \textbf{Recall}) = \frac{TP}{TP+FN} \tag{16}$$

where TP is the number of positive samples identified correctly, and FN represents the number of negative samples identified incorrectly.

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

$$\textbf{Precision} = \frac{TP}{TP+FP} \tag{18}$$

where FP is the number of negative samples identified incorrectly, and TN represents the number of negative samples identified correctly.

By varying the rank threshold, TPR, FPR and Precision can be calculated to construct the receiver operating characteristic (ROC) curve and the Precision-recall (PR) curve. For the ROC curve, FPR is plotted on the *x*-axis and TPR is plotted on the *y*-axes. For the PR curve, Recall and Precision are plotted on the *x* and *y* axes, respectively. The area under ROC curve (AUROC) and the area under the precision–recall curve (AUPR) are utilized to evaluate the overall performance of DRWBNCF. To obtain convincing results, we repeat 10 times 10-fold CV like the work of Cai *et al.* [40] and report the average value as the final result.

Baseline methods

To evaluate the performance of our proposed method, we compared DRWBNCF with four state-of-the-art association prediction methodologies listed below:

- DRIMC [23], a novel drug repositioning method, incorporates features associated with drugs and diseases in Bayesian inductive matrix completion. Zhang *et al.* integrated four drug data sources two disease data sources into a fused drug similarity matrix and a fused disease similarity matrix, respectively, and described the feature of each drug or disease by similarity values between it and its nearest neighbors, which is similar to our paper.
- SCMFDD [22] is a novel matrix factorization method for the drug–disease association prediction. Notably, SCMFDD incorporates drug features (i.e. substructures, enzymes, pathways, targets and drug–drug interactions), disease semantic information as similarity constraints for drugs and diseases into matrix factorization, which is different from the classic matrix factorization method.
- NIMCGCN (a novel Neural Inductive Matrix Completion with Graph Convolutional Network method for the miRNA-disease association prediction) [41], is

Table 2. The AUROCs and AUPRs of all compared approaches obtained in 10 times 10-fold cross-validation

Datasets	DRIMC	SCMFDD	NIMCGCN	BNNR	LAGCN	DRWBNCF
AUROC						
Fdataset	0.9131 ± 0.0040	0.7759 ± 0.0011	0.8321 ± 0.0040	0.9330 ± 0.0032	0.8832 ± 0.0246	0.9257 ± 0.0012
Cdataset	0.9341 ± 0.0008	0.7930 ± 0.0009	0.8549 ± 0.0041	0.9410 ± 0.0072	0.9196 ± 0.0052	0.9413 ± 0.0009
LRSSL	0.9321 ± 0.0007	0.7689 ± 0.0008	0.8330 ± 0.0036	0.9284 ± 0.0017	0.9349 ± 0.0013	0.9355 ± 0.0010
Avg.*	0.9261	0.7793	0.8400	0.9341	0.9128	0.9341
AUPR						
Fdataset	0.3136 ± 0.0040	0.0513 ± 0.0003	0.3440 ± 0.0055	0.4410 ± 0.0038	0.1301 ± 0.0113	0.4910 ± 0.0065
Cdataset	0.3923 ± 0.0029	0.0519 ± 0.0005	0.4408 ± 0.0082	0.4730 ± 0.0028	0.1910 ± 0.0061	0.5663 ± 0.0074
LRSSL	0.2672 ± 0.0011	0.0360 ± 0.0002	0.2735 ± 0.0065	0.3214 ± 0.0062	0.1139 ± 0.0030	0.3491 ± 0.0075
Avg.*	0.3244	0.0464	0.3528	0.4118	0.1450	0.4688

Avg.* shows the average AUROC or AUPR over three benchmark datasets.

equipped with both graph convolutional networks to learn miRNA and disease latent feature representations and a novel neural inductive matrix completion to generate an association matrix completion.

- BNNR (bounded nuclear norm regularization) completes the drug–disease matrix under the low-rank assumption [19] Yang *et al.*, constructed a heterogeneous drug–disease network, which integrates the drug–drug, disease–disease and drug–disease networks and incorporated a regularization term to balance the approximation error and the rank properties.
- LAGCN (a Layer Attention Graph Convolutional Network method for the drug–disease association prediction) [25] utilizes the conventional graph convolution operation on the heterogeneous network composed of the known drug–disease associations, drug–drug similarities and disease–disease similarities, to learn the embeddings of drugs and diseases. In order to get more informative representations of drugs and diseases, Yu and Huang *et al.* further introduced the attention mechanism for integrating all useful structural information at multiple graph convolution layers.

Parameters setting

Our proposed DRWBNCF model uses a three-layer architecture with 64 and 32 hidden units, a dropout rate of 0.4, a learning rate of 0.0005, and a training epoch of 64 in all experiments. The number of nearest neighbors k is set to 3, which are selected within [1, 2, 3, ..., 20]. λ that measures the WBA and traditional linear aggregator is set to 0.9, which are selected within [0, 0.1, 0.2, ..., 1.0]. The dimensionality of $\mathbf{h}_{r,d}$, $\tilde{\mathbf{h}}_r$ and $\tilde{\mathbf{h}}_d$ is 64, 32 and 32, respectively. The hyperparameters $\alpha = 0.5$, $\lambda_r = 2^{-3}$, $\lambda_d = 2^{-4}$, $\gamma = 2$. The hyperparameters of DRIMC, SCMFDD, NIMCGCN and LAGCN were chosen as their optimal values.

Results and Discussions

Performance of DRWBNCF in 10 times 10-fold cross-validation

To evaluate the performance of DRWBNCF, we conducted an extensive set of experiments on three benchmark datasets and compared DRWBNCF with five state-of-the-art association prediction approaches by using 10 times 10-fold cross-validation. Table 2 reports that the AUROC obtained by BNNR on Fdataset is higher than that of our proposed model and the average AUROC achieved by BNNR is equal to that of our proposed model. However, the AUPR achieved by BNNR on the three datasets is lower than our proposed model DRWBNCF. Generally, the unknown drug–disease pairs are much larger than the known drug–disease pairs. For such datasets with imbalanced positive and negative samples, AUPR is also an indispensable evaluation indicator. DRWBNCF obtains the best average AUPR of 0.4688, which is 5.7% higher than BNNR. The results show that DRWBNCF improves the prediction performance in 10 times 10-fold cross-validation thanks to combining the information of the known drug–disease association with the drug's and disease's neighborhood and neighborhood interactions information.

Due to the fact that the number of positive associations is far scarce compared to the number of negative associations, it is meaningful to measure the fraction of the correctly predicted true positive associations in the top-k predictions. We further examine the DRWBNCF's recognition ability of true positive associations by using Recall@k metric as reported in the work of Zeng *et al.* [24]. Recall@k refers to the ratio of the number of correctly identified positive associations retrieved from the top-k predictions to the number of all positive associations in the dataset. As shown in Figure 2, the Recall@k values of DRWBNCF are 77.65, 79.54 and 65.29% for the top 10,000 predictions in Fdataset, Cdataset and LRSSL, respectively, significantly outperforming that of DRIMC (66.32, 67.61 and 54.90%), LAGCN (9.98, 63.90 and 42.45%), NIMCGCN (65.34, 66.67 and 53.79%), BNNR (75.24, 77.68 and 63.69%) and SCMFDD (31.09, 28.08 and 20.12%).

Effect of the number of nearest neighbors

In this section, we do experiments in three benchmark datasets to study how the results change in comparison to using different numbers of neighbors in 10-fold cross-validation. Figure 3 plots the AUROCs obtained by DRWBNCF with respect to different numbers k of nearest neighbors. As shown in Figure 3, DRWBNCF achieves best

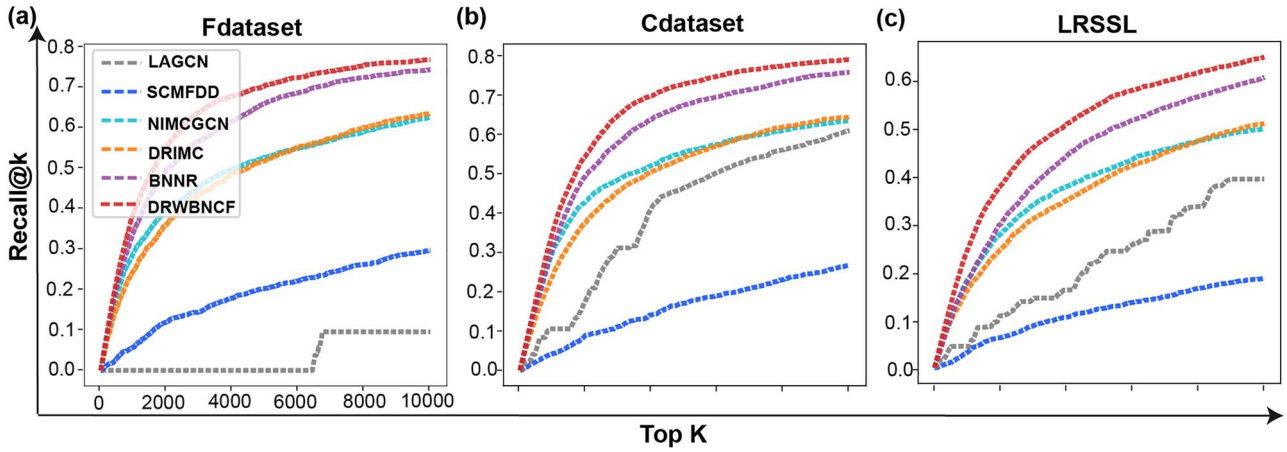

Figure 2. The Recall@k values against the top k predicted list of DRWBNCF and other compared methods during 10-fold cross-validation on Fdataset, Cdataset and LRSSL, respectively.

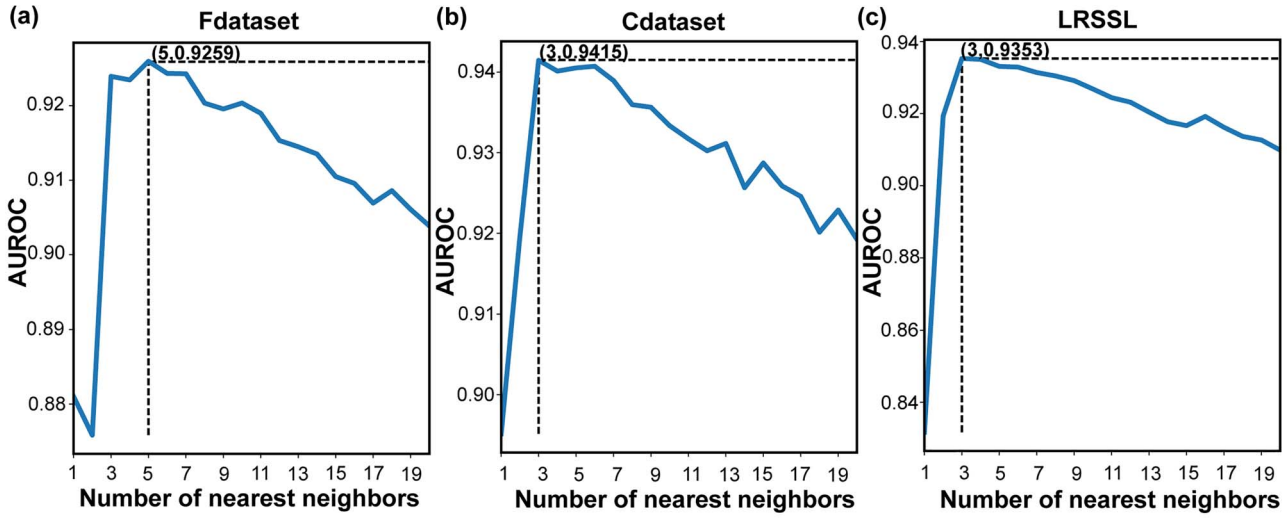

Figure 3. Performance trend of DRWBNCF measured by AUROCs with different numbers of nearest neighbors in 10-fold cross-validation on three benchmark datasets (A) Fdataset, (B) Cdataset and (C) LRSSL.

AUROCs via setting k as 5, 3 and 3, for three benchmark datasets, respectively. Over all datasets (i.e. Fdataset, Cdataset and LRSSL), when $k = 0$ (i.e. without neighbors information), the AUROCs obtained by DRWBNCF are 0.8748, 0.8898 and 0.8343, respectively, while when k is set as the number of all neighbors, these values are 0.7496, 0.7620 and 0.7336. To conclude, DRWBNCF achieves the best performance with the optimal value of k, which is substantially better than that without neighbors information or with all neighbors information. These results demonstrate that the effectiveness of nearest neighbors to predict the association probability for a given drug–disease pair. Consequently, when $k = 3$, we can get reasonably good results for all datasets.

The proposed WBA benefits

In order to analyze the effects among the conventional graph convolution aggregator (i.e. AGG) and our proposed weighted bilinear aggregator (i.e. WBA) in DRWBNCF model clearly, we selected the optimal parameter λ based on nested cross-validation on Fdataset. Nested cross-validation is a common approach to choose hyper-parameters and evaluate model performances. The evaluation results of models corresponding to the different parameter λ on Fdataset in nested cross-validation are shown in Table 3. We find that the AUROC and AUPR of the model with $\lambda = 0.9$ are optimal. Owing to lacking of neighborhood interactions information, the AGG (when $\lambda = 0.0$) does not exhibit excellent performance, whereas our proposed DRWBNCF (when $\lambda = 0.9$) makes full use of both nearest neighbors and weighted interactions between neighbors, showing the best performance.

Discovery of drug candidates for new diseases

To examine the ability of DRWBNCF for discovering novel drug candidates for new diseases without any treatment information, we further implemented the local leave-one-out cross-validation (LOOCV) in Fdataset. Specifically, for each disease d_x, we deleted all the known drug–disease associations about disease d_x as the testing samples and used all the remaining known associations as the training samples. After each disease is repeatedly predicted, we can obtain the predicted

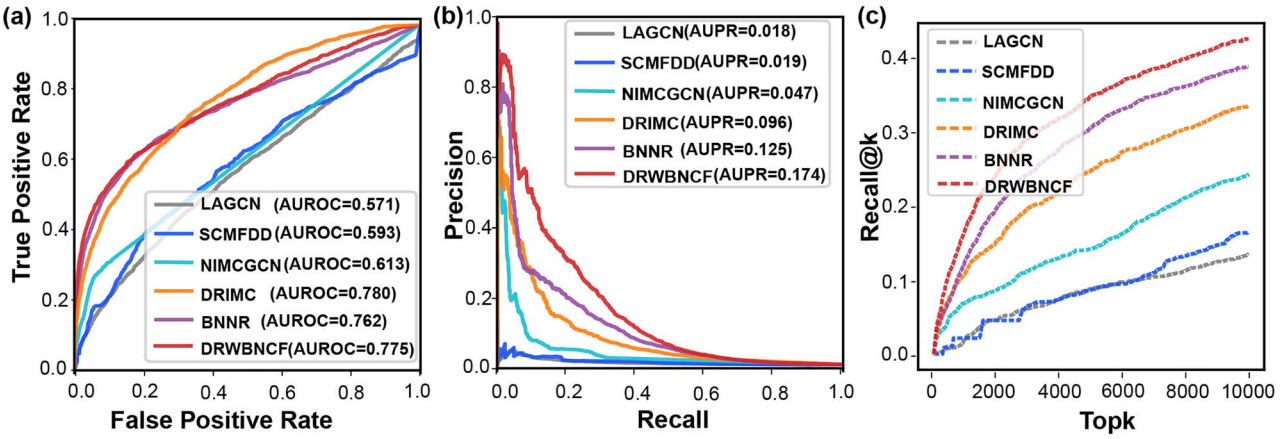

Figure 4. The performance of DRWBNCF and other compared methods in discovering novel drug candidates for new diseases on Fdataset. (a) Receiver operating characteristic (ROC) curves of prediction results obtained by applying DRWBNCF and other compared methods. (b) Precision-recall (PR) curves of prediction results obtained by applying DRWBNCF and other compared methods. (c) The Recall@k in top 10,000 predictions obtained by applying DRWBNCF and other compared methods.

Table 3. The AUROCs and AUPRs of models corresponding to the different parameter λ on Fdataset

λ	AUROCs	AUPRs
$\lambda = 0.0$	0.9238	0.4889
$\lambda = 0.1$	0.9236	0.4918
$\lambda = 0.2$	0.9241	0.4937
$\lambda = 0.3$	0.9237	0.4967
$\lambda = 0.4$	0.9229	0.4936
$\lambda = 0.5$	0.9233	0.4994
$\lambda = 0.6$	0.9238	0.4977
$\lambda = 0.7$	0.9237	0.4973
$\lambda = 0.8$	0.9241	0.4942
$\lambda = 0.9$	0.9247	0.5009
$\lambda = 1.0$	0.9246	0.4982

Note: $\lambda = 0.0$ denotes that a model only includes the conventional graph convolution aggregator (i.e. AGG), $\lambda = 1.0$ represents a model only includes the proposed WBA. $\lambda = 0.4$ means that our proposed WBA accounts for 0.4 while the traditional linear aggregator accounts for 0.6.

probability of associations for each disease. It should be noted that, the known drug–disease associations data are missing for a new disease, our proposed model is able to predict the potential drugs for new diseases by making use of the similarity information of new diseases. As shown in Figure 4A and B, compared with the state-of-the-art drug–disease association prediction methods, e.g. GCN-based, matrix completion-based and matrix factorization-based models, our DRWBNCF achieves 4.9–15.6% absolute AUPR improvements and surpasses two other GCN-based methods by up to 20.4% (AUROC value), which signifies the superior performance of our proposed model on predicting novel drugs for new diseases. Additionally, from the Figure 4C, we can find that the Recall@k value (45.06%) of DRWBNCF for the top 10,000 predictions in Fdataset significantly outperforms that of DRIMC (35.44%), LAGCN (14.28%), NIMCGCN (25.66%), SCMFDD (17.38%) and BNNR (41.38%).

Case studies: computationally predicted potential drugs for breast cancer and small cell lung cancer

To evaluate the practical ability of DRWBNCF, we implemented two case studies for two cancers (i.e. breast cancer and lung cancer), which do not have efficacious drugs available yet. In the process of predicting novel drug candidates for breast cancer and lung cancer, we used all the known drug–disease pairs in Fdataset to train our model and then predicted the association probability of all unobserved pairs. Subsequently, we ranked the pair candidates according to the predicted probabilities in descending order, such that the top-ranked pairs are the most likely to interact. Following the similar settings in previous studies, highly reliable sources and clinical trials, i.e. DrugBank [29], the Comparative Toxicogenomics Database (CTD) [42], DrugCentral [43] and ClinicalTrials.gov, are adopted as references to validate whether the predictions for two cancers are true or not.

Breast cancer is one of the well-known malignant tumors among women. Till now, it still remains the leading cancer-related cause of disease burden for women [44]. We focused on the top 10 DRWBNCF-predicted drugs for potentially treating breast cancer in Table 4. Metigestrona, a long-acting contraceptive, has also been used to treat breast and endometrial neoplasms. Herein, Metigestrona is the top first predicted candidate drug for treating breast cancer. Estramustine, an antineoplastic agent that was primarily used in the treatment of prostatic neoplasms, was predicted by DRWBNCF to be associated with breast cancer. Such a prediction can be supported by a previous clinical trial reporting that estramustine showed encouraging results in the treatment of metastatic breast cancer and was deserved to be studied in earlier clinical situation. In addition, DRWBNCF found that mitoxantrone, an anthracenedione-derived antineoplastic agent, was associated with breast cancer, which is supported by DB, CTD, DrugCentral and ClinicalTrials.gov. To conclude, among the top 10 predicted drugs ranked according to their predicted scores, 8 drugs (80% success rate) have validated by various evidences from authoritative sources and clinical trials.

Small cell lung cancer with high incidence and high mortality worldwide makes this complex neoplasm a notable healthcare issue. We also focused on the top

Table 4. The top 10 DRWBNCF-predicted drugs for potentially treating breast cancer (OMIM: 114480)

Rank	DrugBank IDs	Candidate drugs	Evidences
1	DB00603	Metigestrona	ClinicalTrials.gov, CTD
2	DB01196	Estramustine	ClinicalTrials.gov
3	DB00977	Ethinylestradiol	ClinicalTrials.gov
4	DB00655	Estrone	DrugCentral
5	DB00717	Norethisterone	ClinicalTrials.gov, DrugCentral
6	DB00385	Valrubicin	Unconfirmed
7	DB00694	Daunorubicin	ClinicalTrials.gov
8	DB01204	Mitoxantrone	ClinicalTrials.gov, DB, CTD, DrugCentral
9	DB00007	Leuprolide	ClinicalTrials.gov
10	DB00444	Teniposide	Unconfirmed

Table 5. The top 10 DRWBNCF-predicted drugs for potentially treating small cell lung cancer (OMIM: 182280)

Rank	DrugBank IDs	Candidate drugs	Evidences
1	DB00997	Doxorubicin	ClinicalTrials.gov, DrugCentral, CTD,
2	DB00694	Daunorubicin	ClinicalTrials.gov
3	DB00445	Epirubicin	ClinicalTrials.gov, DB, CTD
4	DB01177	Idarubicin	CTD
5	DB00541	Vincristine	ClinicalTrials.gov, DB, CTD, DrugCentral
6	DB00385	Valrubicin	Unconfirmed
7	DB00762	Irinotecan	ClinicalTrials.gov, DB, CTD, DrugCentral
8	DB00570	Vinblastine	ClinicalTrials.gov
9	DB00987	Cytarabine	Unconfirmed
10	DB01204	Mitoxantrone	ClinicalTrials.gov

10 DRWBNCF-predicted candidate drugs for potentially treating small cell lung cancer in Table 5. Table 5 reports that 8 out of 10 drugs (80% success rate) are validated by the reliable sources and clinical trials. For instance, doxorubicin, an antineoplastic antibiotic used to treat various cancers and Kaposi's Sarcoma, is the top first predicted novel drug for potentially treating small cell lung cancer. Such a prediction is supported by CTD, DrugCentral and ClinicalTrials.gov. Daunorubicin, an anthracycline aminoglycoside antineoplastic, which was primarily used in the treatment of leukemia and other neoplasms, was predicted by DRWBNCF to have potential effect on small cell lung cancer. This prediction is supported by ClinicalTrials.gov. In addition, vincristine and irinotecan predicted by DRWBNCF have been confirmed by three reliable sources and clinical trials for small cell lung cancer promising treatment.

For a clear view, we show the top 10 potential drugs association networks of the two cancers in Figure 5. It indicates that some drugs are usually related to several diseases. For example, daunorubicin, mitoxantrone and valrubicin are associated with both breast cancer and small cell lung cancer. The predictions of daunorubicin and mitoxantrone have been confirmed by ClinicalTrials.gov. In summary, DRWBNCF can aid biological researchers and clinicians in selecting anticancer drugs by accurately repurposing the de-risked and old drugs.

Conclusion

In this study, we proposed a deep-learning methodology DRWBNCF to in silico drug repositioning. The key innovation of DRWBNCF is the explicitly encoding of the local nearest neighbors and their interactions to augment conventional graph convolution operation. Unlike some previous application of neighbor information in drug repositioning approach, DRWBNCF uses nearest neighbors instead of all similar neighbors, which can filter out noisy information to get more accurate results. In order to capture the complex patterns of information propagation, we aggregate all pair-wise weighted interactions between neighbors by proposing a new weighted bilinear graph convolution operation, which is an excellent complement to the conventional graph convolution operation. Through integrating the representations of the drug–disease known association, drug's and disease's neighbors and neighborhood interactions into a unified representation as the input of MLP and differentiating the easy negative samples and hard positive samples by introducing the α-balanced focal loss function to improve the performance of DRWBNCF. To evaluate the performance of DRWBNCF, we conducted an extensive set of experiments on three benchmark datasets, comparing with four state-of-the-art association prediction methods. For example, DRWBNCF achieves the best AUROC and AUPR values under 10 times 10-fold cross-validation over all datasets. In terms of discovering novel drug candidates for new diseases, DRWBNCF obtains the best AUROC, AUPR and Recall@k values under the local leave-one-out cross-validation. The experimental results validate the efficacy of our proposed DRWBNCF model. Case studies demonstrate that DRWBNCF has a high practical predicting ability. For example, on the Fdataset, 80% of

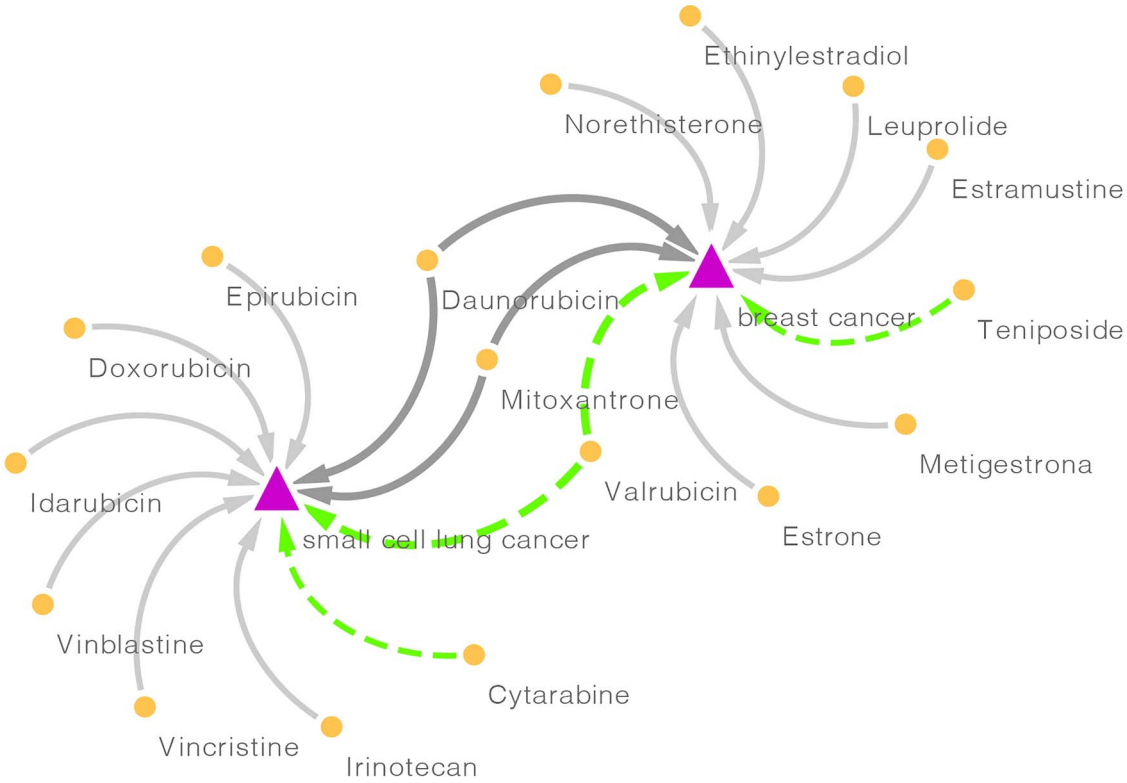

Figure 5. The top 10 potential drugs association networks of breast cancer and small cell lung cancer.

top-10 DRWBNCF-predicted potential drugs for breast cancer and small cell lung cancer have been confirmed by highly reliable sources and clinical trials, including DrugBank, CTD, DrugCentral and ClinicalTrials.gov.

Despite the effectiveness of DRWBNCF to in silico drug repositioning, there are still some limitations to this approach. Since some related studies have demonstrated that integrating multiple data sources may enhance the performance of models, DRWBNCF can incorporate the characteristics of drugs and diseases involved in multiple data sources for better performance in future work. Second, the lack of true negative drug–disease interactions and the high-quality gold-standard dataset may result in an incomplete picture. Third, compared with some novel matrix factorization and completion methods (such as BNNR [19] and DRIMC [23]), DRWNBCF requires more computing resources and is time-consuming (Supplementary Table S2). In the future, we will develop a more efficient deep-learning model with a lighter and more compressed architecture.

In conclusion, our proposed DRWBNCF could significantly benefit from integrating neural collaborative filtering with neighborhood and neighborhood interactions to predict the association probability of a given drug–disease pair. Eventually, DRWBNCF can help accelerate the development of drug discovery by predicting which existing drugs could treat complex diseases to guide the time-consuming and costly wet experiments.

Key Points

- We propose a novel deep-learning approach, called DRWBNCF, which integrate neighborhood and neighborhood interactions information into neural collaborative filtering to in silico drug repositioning.
- Our constructed drug–drug and disease–disease similarity networks employ the nearest neighbors of the drug and disease, not all similar neighbors of them to avoid noisy information.
- The integration component based on our proposed WBA leverages localized information in complementing the drug–disease association data to enhance the prediction ability of DRWBNCF.
- We resort to a multi-layer perceptron optimized by the α-balanced focal loss function and graph regularization to grasp complex associations.

Data availability

The implementation of DRWBNCF and the preprocessed data is available at: https://github.com/luckymengmeng/DRWBNCF.

Conflicts of interest

JY is employed by the company Geneis Beijing Co., Ltd. The other authors have no conflicts of interest to declare.

Funding

This work was supported by the Natural Science Foundation of Hunan Province (Grant 2021JJ30139); the Changsha Key Research and Development Program (Grant KQ2004011) and the National Natural Science Foundation of China (Grant 61773157).

References

1. Rifaioglu AS, Atas H, Martin MJ, *et al.* Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* 2019;**20**:1878–912.
2. Fisher WJ. Alterations in processes and priorities needed for new drug development. *Am Coll Phys* 2006;**145**:793–796.
3. DiMasi JA. New drug development in the United States from 1963 to 1999. *Clin Pharmacol Ther* 2001;**69**:286–96.
4. Avorn J. The $2.6 billion pill–methodologic and policy considerations. *N Engl J Med* 2015;**372**:1877–9.
5. Pushpakom S, Iorio F, Eyers PA, *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**: 41–58.
6. Yang J, Peng S, Zhang B, *et al.* Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 2020;**42**:353–72.
7. Liu C, Wei D, Xiang J, *et al.* An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol Ther Nucleic Acids* 2020;**21**: 676–86.
8. Ezzat A, Wu M, Li X-L, *et al.* Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 2019;**20**:1337–57.
9. Lu C, Yang M, Luo F, *et al.* Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics* 2018;**34**:3357–64.
10. Chen X, Li S-X, Yin J, *et al.* Potential miRNA-disease association prediction based on kernelized Bayesian matrix factorization. *Genomics* 2020;**112**:809–19.
11. Xiao Q, Luo J, Liang C, *et al.* A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 2018;**34**:239–48.
12. Xu J, Zhu W, Cai L, *et al.* LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. *IEEE Access* 2020;**8**: 80728–38.
13. Meng Y, Jin M, Tang X, *et al.* Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl Soft Comput* 2021;**103**:107135.
14. Tang X, Cai L, Meng Y, *et al.* Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front Immunol* 2021;**11**:3824.
15. Zhang W, Chen Y, Li D, *et al.* Manifold regularized matrix factorization for drug-drug interaction prediction. *J Biomed Inform* 2018;**88**:90–7.
16. Zhang W, Liu X, Chen Y, *et al.* Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 2018;**287**:154–62.
17. Xu J, Cai L, Liao B, *et al.* CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 2020;**36**: 3139–47.
18. Luo H, Li M, Wang S, *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12.
19. Yang M, Luo H, Li Y, *et al.* Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 2019;**35**:i455–63.
20. Yang M, Luo H, Li Y, *et al.* Overlap matrix completion for predicting drug-associated indications. *PLoS Comput Biol* 2019;**15**:e1007541.
21. Yang M, Wu G, Zhao Q, *et al.* Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief Bioinform* 2021;**22**:bbaa267.
22. Zhang W, Yue X, Lin W, *et al.* Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics* 2018;**19**:1–12.
23. Zhang W, Xu H, Li X, *et al.* DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion. *Bioinformatics* 2020;**36**:2839–47.
24. Zeng X, Zhu S, Liu X, *et al.* deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**:5191–8.
25. Yu Z, Huang F, Zhao X, *et al.* Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021;**22**:bbaa243.
26. Feng Q. PADME: A deep learning-based framework for drug-target interaction prediction[D]. *Applied Sciences: School of Computing Science*, 2019.
27. Luo H, Li M, Yang M, *et al.* Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform* 2021;**22**:1604–19.
28. Gottlieb A, Stein GY, Ruppin E, *et al.* PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496.
29. Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
30. Hamosh A, Scott AF, Amberger JS, *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**:D514–7.
31. Luo H, Wang J, Li M, *et al.* Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;**32**:2664–71.
32. Liang X, Zhang P, Yan L, *et al.* LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics* 2017;**33**:1187–96.
33. Vidal D, Thormann M, Pons M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model* 2005;**45**:386–93.
34. Steinbeck C, Han Y, Kuhn S, *et al.* The Chemistry Development Kit (CDK): an open-source Java library for chemo-and bioinformatics. *J Chem Inf Comput Sci* 2003;**43**:493–500.
35. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 2016.
36. Zhu H, Feng F, He X, *et al.* Bilinear graph neural network with neighbor interactions. arXiv preprint arXiv:2002.03575 2020.
37. Lin T-Y, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 2980–8.
38. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
39. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010, 249–56.
40. Cai L, Lu C, Xu J, *et al.* Drug repositioning based on the heterogeneous information fusion graph convolutional network[J]. *Brief Bioinform* 2021;**22**(6):bbab319.

41. Li J, Zhang S, Liu T, *et al.* Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020;**36**:2538–46.
42. Davis AP, Grondin CJ, Johnson RJ, *et al.* Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res* 2021;**49**:D1138–43.
43. Avram S, Bologa CG, Holmes J, *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res* 2021;**49**:D1160–9.
44. Britt KL, Cuzick J, Phillips K-A. Key steps for effective breast cancer prevention. *Nat Rev Cancer* 2020;**20**: 417–36.