frontiers in Pharmacology

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/fphar.2019.01301

HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction *via* Chou's Five-Step Rule

Ping Xuan[1], Hui Cui[2], Tonghui Shen[1], Nan Sheng[1] and Tiangang Zhang[3]**

[1] School of Computer Science and Technology, Heilongjiang University, Harbin, China, [2] Department of Computer Science and Information Technology, La Trobe University, Bundoora, VIC, Australia, [3] School of Mathematical Science, Heilongjiang University, Harbin, China

Identifying new treatments for existing drugs can help reduce drug development costs and explore novel indications of drugs. The prediction of associations between drugs and diseases is challenging because their similarities and relations are complicated and non-linear. We propose a HeteroDualNet model to address this issue. Firstly, three types of matrices are extracted to represent intra-drug similarities, intra-disease similarity and drug-disease associations. The intra-drug similarities consider three drug features and a newly introduced drug-related disease correlation. Secondly, an embedding mechanism is proposed to integrate these matrices in a heterogenous drug-disease association layer (hetero-layer). Further, a neighbouring heterogeneous layer (hetero-layer-N) is constructed to incorporate the biological premise that similar drugs can often treat related diseases. Finally, a dual convolutional neural network is built with hetero-layer and hetero-layer-N as two branches to learn from characteristics of drug-disease and the relations of their neighbours simultaneously. HeteroDualNet outperformed the other four methods in comparison over a public dataset of 763 drugs and 681 diseases in terms of Areas Under the Curves of Receiver Operating Characteristics and Precision-Recall, and recall rate at top *k*. Case study of five drugs further proved the capacity of HeteroDualNet in finding reliable disease candidates of drugs as validated by database records or literature. Our findings show that the embedded heterogenous layers of original and neighbouring drug-disease representations in a dual neural network improved the association prediction performance.

Keywords: drug-disease association prediction, multiple kinds of similarities, neighbouring heterogeneous layer, deep learning, dual convolutional neural network

OPEN ACCESS

Edited by:
Jianfeng Pei,
Peking University,
China

Reviewed by:
Feng Zhu,
Zhejiang University,
China
Kuo-Chen Chou,
The Gordon Life Science Institute,
United States

****Correspondence:***
Tonghui Shen
shen.tonghui@163.com
Tiangang Zhang
zhang@hlju.edu.cn

Specialty section:
This article was submitted to Translational Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *01 July 2019*
Accepted: *11 October 2019*
Published: *08 November 2019*

Citation:
Xuan P, Cui H, Shen T, Sheng N and Zhang T (2019) HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction via Chou's Five-Step Rule. Front. Pharmacol. 10:1301. doi: 10.3389/fphar.2019.01301

INTRODUCTION

The research and development (R&D) processes of new drugs are time-consuming and expensive. Stringent drug testing and approvals are required for an invented new drug to make it to market. For instance, it takes an average of 15 years from preliminary examination of compounds to clinical trials of drug candidates, and finally to drug marketing, while the estimated investment cost is about 800 million dollars (Adams and Brantner, 2006; Tamimi and Ellis, 2009; Pushpakom et al.,

2018). However, even in the case of a significant amount of time and capital investment, the R&D of new drugs still faces high failure risks (Li et al., 2016). Meanwhile, the number of new drugs approved by major drug regulatory agencies around the world is decreasing year by year (Grabowski, 2004; Nosengo, 2016). According to the statistics of the US Food and Drug Administration (FDA), the average success rate of new drugs approved from 2003 to 2011 was less than 10% (Padhy and Gupta, 2011; Hay et al., 2014; Pritchard et al., 2017). Therefore, the conventional R&D productivity of new drugs has been stagnant in the last few decades (Paul et al., 2010).

Given the challenges faced by conventional drug R&D techniques, there are significant needs of innovative drug development strategies to increase R&D productivity, which is one of the essential priorities in the pharmaceutical industry. Drug repositioning techniques, or the so-called reuse of existing drugs, have been proved of its advantages over the conventional drug R&D strategies. (Hurle et al., 2013) Drug repositioning is the process to identify new indications for existing drugs and is playing an essential role in the state-of-the-art drug R&D process. Drug repositioning can be applied to drugs which have been approved to market. Because those drugs have passed the procedures of laboratory, pharmacokinetics, toxicology and safety testing, drug developers can use these drugs in clinical trials directly. In this way, drug repositioning skips those procedures and will significantly reduce the time and financial costs in drug development. At the same time, it also reduces the risks of drug development failure. Thus, drug repositioning has attracted great interests in the pharmaceutical industry and research community (Hurle et al., 2013).

Drug repositioning aims to find potential indications for existing drugs (Shim and Liu, 2014; Chen et al., 2016). Computational methods in biology are playing increasingly important roles in the stimulation, development and finding of new drugs (Chou, 2015). To develop useful predictors for biological systems *via* computing models, Chou's 5-steps (Chou, 2011; Chou, 2019b) are used by recent publications (Chou, 2019a; Awais et al., 2019; Ehsan et al., 2019; Hussain et al., 2019). These steps provide guidance in the development and validation of computerized methods, which include selection of a valid benchmark dataset for training and testing, representation of samples by effective formulation to reflect intrinsic correlations with the target, development of algorithms for prediction, objective performance evaluation by cross-validation, and consideration of public accessibility by web-server.

Several methods have been proposed to predict drug-disease associations. For example, Chiang and Butte proposed a technique based on the internal correlation of networks to predict the potential drug-disease associations (Chiang and Butte, 2009). Sirota et al. developed a prediction method by integrating the common gene expressions of drugs and diseases (Sirota et al., 2011). Besides, Yang and Agarwal et al. proposed to infer the new drug-disease associations by using the phenotypic information on drug side effects (Yang and Agarwal, 2011). Most of these methods are designed for early-stage drugs which have multiple uses and treatment plans. They cannot be used for association prediction when there are no common gene expressions and side effects information between drugs and diseases.

With the increasing amount and variety of drug-related data, recent research has been focusing on integrating multimodality information to investigate the potential uses of drugs. Gottlieb et al. proposed a classification model which used various associations of drug and disease as distinguish signatures. A logistical regression model was then used to predict the indications of drugs (Gottlieb et al., 2011). A kernel-based strategy was proposed to integrate molecular structure, molecular activity, and phenotypic information for drug repositioning (Wang et al., 2013). Heterogenous networks have also been investigated to predict drug indications. Heterogeneous networks are constructed by associating drugs, diseases, targets and genes. The prediction can be achieved by approaches such as network clustering (Wu et al., 2013), priority ranking (Martinez et al., 2015), network topology measurement (Chen et al., 2015), or iteration (Wang et al., 2014b). Given these heterogeneous networks, some other models integrated multiple chemical features such as chemical phenotype of drugs and molecular characteristics of diseases. Then the prediction of new drug indications can be achieved by proteochemometric models (Dakshanamurthy et al., 2012; Yu et al., 2015), statistical (Iwata et al., 2015) or sparse subspace learning (Liang et al., 2017; Xuan et al., 2019) models.

Most of the above existing methods for drug-disease association predictions are shallow models. The associations between drugs and diseases, however, are non-linear and complicated. It is challenging for these shallow models to dig out advanced level while hidden drug-disease relations. Thus, there are great necessities to develop models to learn the deep representations of drug-disease associations for improved drug indication prediction.

In this work, we propose a novel convolutional network with heterogeneous layers and dual branches, referred to as HeteroDualNet, for drug-disease association prediction. Our first unique contribution is the extraction of three types of matrices for the representation and indexing of intra-drug similarity, drug-disease similarity and drug-disease associations. When constructing intra-drug similarity matrices, we consider both regular drug features, including chemical substructures, domains and annotations of target proteins, and a newly introduced feature calculated by drug-related disease correlations. The second contribution is that we construct a new heterogenous drug-disease association layer (hetero-layer) to associate three types of matrices by a proposed embedding mechanism. Further, a drug-disease association layer with neighbouring information (hetero-layer-N) is constructed by the embedding mechanism to reflect the biological premise that similar drugs can often treat related diseases. Finally, HeteroDualNet is built to predict drug-disease associations with hetero-layer and hetero-layer-N as two branches to learn from both original and neighbouring characteristics of drugs and diseases simultaneously. We also investigate the prediction capacity of the proposed model in therapeutic drug indications by case studies of five drugs.

MATERIALS AND METHODS

Dataset

We obtained the data of drugs and diseases from a published work (Wang et al., 2014a). There are 763 drugs, 681 diseases and 3051 known drug-disease associations. The characteristics of each drug include 881 chemical substructures which were initially derived from the chemical fingerprints extracted from the PubChem database (Wang et al., 2009); 1,426 target protein domains from the InterPro database (Mitchell et al., 2015); and 4,447 target protein annotations obtained from the UniProt database (Uniprot, 2010). The similarities among diseases were calculated by (Wang et al., 2010) and provided in the dataset.

Hypothesis and Framework

We hypothesize that a dual neural network which integrates features of drugs, drug-related disease correlations, and the biological premise of drugs and diseases will improve the performance of drug-disease association predictions. The overview of the proposed method is shown in **Figure 1**. Given the input dataset, the drugs and diseases information is firstly extracted and indexed by three types of similarity matrices in terms of intra-drug, intra-disease and drug-disease. Then, a heterogenous drug-disease association layer, referred by hetero-layer, is constructed by a proposed embedding mechanism to associate those matrices among drugs and diseases. Another heterogeneous layer with neighbouring information, denoted by hetero-layer-N, is built to represent the biological premise that similar drugs can often treat related diseases. Lastly, the dual convolutional neural network is constructed by integrating hetero-layer and hetero-layer-N using a fully connected layer.

Drug and Disease Similarity and Association Representation

We define three types of matrices to represent and index the information of drugs and diseases in terms of intra-drug similarity, intra-disease similarity and drug-disease associations.

Intra-Disease Similarity Matrix

Intra-disease similarities were calculated and provided by (Wang et al., 2010) based on semantic information of diseases (Wang et al., 2010). This information was also used in published work such as Liang et al. (2017) and Zhang et al. (2018). The similarity between disease d_i and the disease d_j is denoted by $\boldsymbol{D}(i,j) \in [0,1]$. where is the intra-disease similarity matrix and N^{DI} is the number of diseases. The greater $D(i,j)$ is, the higher similarity between diseases d_i and d_j.

Intra-Drug Similarity Matrix

Four intra-drug similarity matrices are obtained by calculating the similarities between drugs from four perspectives, including the chemical substructures, target protein domain information, target protein annotations and the related disease information of drugs.

The first three intra-drug similarity matrices of chemical substructure, domain and annotation information of target proteins represent that if two drugs have more common chemical substructures, target protein domains or gene ontology information, the more similar they are. Thus, we calculate these three intra-drug similarity matrices by cosine similarity measurement (Liang et al., 2017).

To calculate the first three intra-drug similarity matrices, we firstly obtain matrices of features and drugs. The feature matrix of chemical feature and all the drugs is denoted by $\boldsymbol{F}_1 \in \mathbb{R}^{N_1^F \times N^{DR}}$ where N_1^F is the number of chemical substructure features, and N^{DR} is the number of drugs. Similarly, the feature matrix of protein domain and drugs is $\boldsymbol{F}_2 \in \mathbb{R}^{N_2^F \times N^{DR}}$ and that of protein annotation and drugs is $\boldsymbol{F}_3 \in \mathbb{R}^{N_3^F \times N^{DR}}$, where N_2^F is the number of target protein domain feature and N_3^F is the number of target protein annotation. Each element of the vectors is 1 or 0 according to whether the drug has such a feature. Given the dataset used in this paper, $N_1^F = 881$, $N_2^F = 1426$ and $N_3^F = 4{,}447$. Let $\boldsymbol{f}_{t,i}$ be the feature vector of i-th drug r_i in the t-th feature matrix F_t ($1 \le t \le 3$), the similarity $\boldsymbol{R}_t(i,j)$ between drugs r_i and r_j in terms of feature t is calculated by cosine similarity measurement as

$$\boldsymbol{R}_t(i,j) = \frac{f_{t,i} \cdot f_{t,j}}{||f_{t,i}||\,||f_{t,j}||}. \quad (1)$$

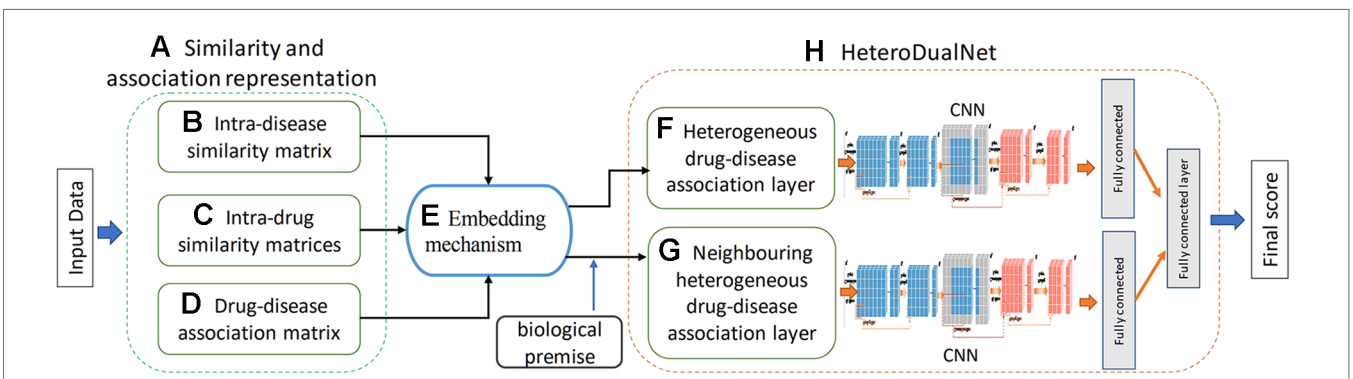

FIGURE 1 | Overview of the proposed HeteroDualNet model for drug-disease association prediction. Given input data, **(A)** similarity and association representations are extracted including **(B)** intra-disease similarity, **(C)** intra-drug similarity, and **(D)** drug-disease association. Then **(E)** an embedding mechanism is proposed to embed these matrices. The final drug-disease association score is obtained by **(H)** HeteroDualNet with **(F)** heterogeneous and biological premise enhanced **(G)** neighboring heterogeneous drug-disease association layers.

where $R_t(i,j)\in[0,1]$ and higher values indicate higher similarity between a pair of drugs.

The fourth intra-drug similarities matrix $\boldsymbol{R}_4 \in \mathbb{R}^{N^{DR}\times N^{DR}}$ is obtained based on the idea that if two drugs are associated with similar diseases, the drugs are more likely to be correlated. Given the dataset of diseases $\boldsymbol{DI}=\{d_k|k\in[1,N^{DI}]\}$ and intra-disease similarity matrix D if i-th drug r_i is associated with a subset of diseases $\boldsymbol{DI}_m \subset \boldsymbol{DI}$, and drug r_j is related to a disease subset $\boldsymbol{DI}_n$, the similarity $\boldsymbol{R}_4(i,j)$ between i-th and j-th drugs can be obtained by calculating the similarity between $\boldsymbol{DI}_m$ and $\boldsymbol{DI}_n$ as proposed in our previous work (Xuan et al., 2019) by

$$\boldsymbol{R}_4(i,j)=\frac{\sum_{k=1}^{num(DI_m)}\max\left(D\left(d_{i,k},d_{j,*}\right)\right)+\sum_{k=1}^{num(DI_n)}\max\left(D\left(d_{j,k},d_{i,*}\right)\right)}{num\left(DI_m\right)+num\left(DI_n\right)} \quad (2)$$

where $num(DI_m)$ denotes the number of elements in DI_m. $d_{i,k}$ represents the kth disease related with drug r_i, $d_{j,*}$ denotes all the related diseases of drug r_j, and $max(D(d_{i,k},d_{j,*}))$ is the maximum similarity between drug r_j's kth related disease and all the related diseases of r_j. Similarly, $max(D(d_{i,k},d_{j,*}))$ denotes the maximum similarity between drug r_j's kth related disease and all the associated diseases of r_i. The final similarity between r_i and r_j is obtained by the average maximum similarities between diseases in their relevant disease subsets DI_m and DI_n.

Drug-Disease Association Matrix

The drug-disease association matrix is denoted by $\boldsymbol{A}\in\mathbb{R}^{N^{DR}\times N^{DI}}$ where an element can be 0 or 1. 1 indicates that a drug and a disease are related, and the association is available; while 0 represents that the relation between a drug and a disease is unknown. Among all the 763 drugs and 681 diseases in the dataset, 3051 drug-disease associations are available. The remaining unknown associations are to be predicted.

HeteroDualNet Architecture

The sparsity of drug-disease associations makes it challenging to dig out the hidden characteristics and relations between drugs and diseases. We construct HeteroDualNet, a dual convolutional neural network with heterogeneous layers, to predict drug-disease associations. One branch integrates the three matrices of drugs and diseases by a heterogeneous association layer (hetero-layer); the other branch incorporates the neighbouring information in a neighbouring heterogenous layer (hetero-layer-N). The two heterogeneous layers are learnt by passing through convolutional and pooling layers and joint by a connection module. The final association score is obtained by weighted voting of association scores from two branches.

Embedding Mechanism for Heterogeneous Drug-Disease Association Matrix

The heterogenous drug-disease association layer is built upon an embedded matrix of afore-extracted matrices. An embedding mechanism is proposed based on the idea that if two drugs are more similar, the more likely they are associated with related diseases, whereas two similar diseases tend to be associated with similar drugs. Given intra-drug matrices $\boldsymbol{R}_t$, drug-disease association matrix A and intra-disease matrix D, the heterogeneous matrix $\boldsymbol{X}_L$ of drug $r_i(i\in[1,N^{DR}])$ and disease $d_k(k\in[1,N^{DI}])$ is obtained by the following embedding procedures.

Firstly, row vectors $R_t(i,*)$ are combined sequentially as $\boldsymbol{X}_{L,11}=[\boldsymbol{R}_1(i,*);\boldsymbol{R}_2(i,*);\boldsymbol{R}_3(i,*);\boldsymbol{R}_4(i,*)]$ where $\boldsymbol{R}_t(i,*)$ denotes the i-th row in an intra-drug similarity matrix R_t which records the t-th type of similarities between r_i and all drugs, t = 1,2,3,4 denotes chemical substructures, target protein domains, target protein annotations and related disease information respectively. Secondly, the transposed column vector $\boldsymbol{A}^{\mathrm{T}}(*,k)$ is concatenated under $\boldsymbol{R}_4(i,*)$ as $\boldsymbol{X}_{L,21}$ where $\boldsymbol{A}(*,k)$ is the kth column of A which contains the associations between d_k and all the drugs. Thirdly, $A(i,*)$ is repeated four times and spliced to the right of each row in $\boldsymbol{X}_{L,11}$ as $\boldsymbol{X}_{L,12}=[\boldsymbol{A}(i,*);\boldsymbol{A}(i,*);\boldsymbol{A}(i,*);\boldsymbol{A}(i,*)]$ where $\boldsymbol{A}(i,*)$ denotes the ith row of A which includes the associations between r_i and all the diseases. Lastly, $\boldsymbol{D}(k,*)$ is spliced under $X_{L,12}$ where $\boldsymbol{D}(k,*)$ is the kth row of $\boldsymbol{D}$ containing the similarities between d_k and all the diseases. The final embedded matrix $\boldsymbol{X}_L\in\mathbb{R}^{5\times(N_r+N_d)}$ of drug r_i and disease d_k is formed as

$$\boldsymbol{X}_L=\begin{bmatrix}\boldsymbol{X}_{L,11} & \boldsymbol{X}_{L,12}\\ \boldsymbol{X}_{L,21} & \boldsymbol{X}_{L,22}\end{bmatrix}=\begin{bmatrix}\boldsymbol{R}_1(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_2(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_3(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{R}_4(i,*) & \boldsymbol{A}(i,*)\\ \boldsymbol{A}^T(*,k) & \boldsymbol{D}(k,*)\end{bmatrix} \quad (3)$$

Given such a heterogeneous matrix X_L, the unknown drug-disease relations can be inferred *via* the correlations between diseases. In the meanwhile, the unavailable associations can be derived upon the similarities between drugs. In **Figure 2**, we illustrate the embedding procedure and use drug r_2 and disease d_1 whose association is unknown as an example. If r_2 is very similar to r_3 and r_4 (as shown in **Figure 2A**),r_3 and r_4 are closely associated with d_1(**Figure 2B**), it can be inferred that r_2 is more likely to be associated with d_1. Alternatively, if d_1 is similar to d_4 (shown in **Figure 2C**), and d_4 is related with r_2 (**Figure 2B**), a high possibility that r_2 is associated with d_1 can be derived.

Neighbouring Heterogeneous Association Matrix

The neighbouring heterogeneous drug-disease association matrix X_{L-N} embeds the neighbours of drug r_i and disease d_k. The embedding is proposed based on the biological premise that if the neighbours of a drug are associated with the neighbours of a disease, there is a high probability that the drug and the disease are associated. The embedding procedures considering the neighbours of r_i and d_k is: Firstly, we find drugs r_m, r_n, r_p,

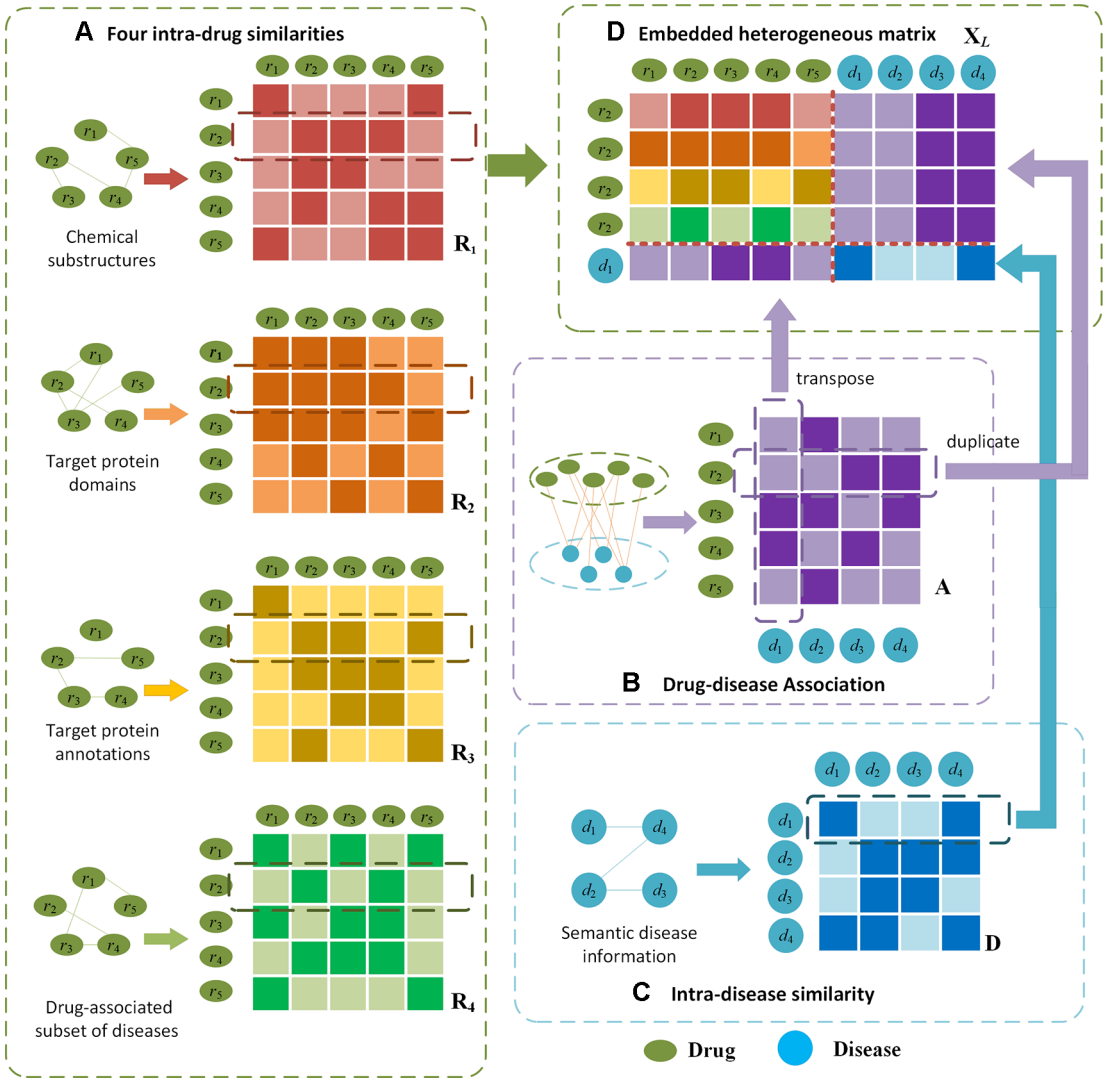

FIGURE 2 | Illustration of the proposed embedding mechanism for heterogenous drug-disease association matrix. Given drug r_2 and disease d_1 as an example, **(D)** the heterogeneous matrix is obtained by integrating **(A)** four types of intra-drug similarities, **(B)** drug-disease associations and **(C)** intra-disease similarities. In **(A)** and **(C)**, darker colours indicate higher similarities; in **(B)** darker colour represents the drug-disease association is available.

and r_q which are the most similar neighbours of drug r_i in $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, $\boldsymbol{R}_3$ and $\boldsymbol{R}_4$ respectively. We also find d_l, the most similar neighbour of d_k, in D. Similar with $\boldsymbol{X}_{L,11}$, the m-th row of $\boldsymbol{R}_1$, nth row of $\boldsymbol{R}_2$, p-th row of $\boldsymbol{R}_3$, and qth row of $\boldsymbol{R}_4$ are combined from top to bottom to form $\boldsymbol{X}_{L-N,11}$. Secondly, the l-th column of $\boldsymbol{A}$ indicating the association between the most similar disease d_l and all the drugs is transposed and concatenated under $\boldsymbol{X}_{L-N,11}$ as $\boldsymbol{X}_{L-N,21}$. Thirdly, row vectors $\boldsymbol{A}(m,^*)$, $\boldsymbol{A}(n,^*)$, $\boldsymbol{A}(p,^*)$, $\boldsymbol{A}(q,^*)$ are spliced to the right of each row in $\boldsymbol{X}_{L-N,11}$, where $\boldsymbol{A}(m,^*)$, $\boldsymbol{A}(n,^*)$, $\boldsymbol{A}(p,^*)$, $\boldsymbol{A}(q,^*)$ indicate the associations between drugs r_m,r_n,r_p and r_q and all the diseases. Lastly, the l-th row of D containing the similarities between disease d_l and all the other diseases is concatenated under $\boldsymbol{X}_{L-N,21}$. In such a way, the final embedding of most similar neighbours of r_i and d_k is formed as $\boldsymbol{X}_{L-N} \in \mathbb{R}^{5\times\left(N^{DR}+N^{DI}\right)}$:

$$\boldsymbol{X}_L = \begin{bmatrix} \boldsymbol{X}_{L-N,11} & \boldsymbol{X}_{L-N,12} \\ \boldsymbol{X}_{L-N,21} & \boldsymbol{X}_{L-N,22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_1(m,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_2(n,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_3(p,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{R}_4(q,*) & \boldsymbol{A}(l,*) \\ \boldsymbol{A}^T(*,l) & \boldsymbol{D}(l,*) \end{bmatrix} \tag{4}$$

In $\boldsymbol{X}_{L-N}$, the most similar neighbours of drugs and diseases serve as the bridge to propagate associations. In **Figure 3**, we use drug r_2 and disease d_1 whose association is unknown as an example to illustrate the embedding procedure and information

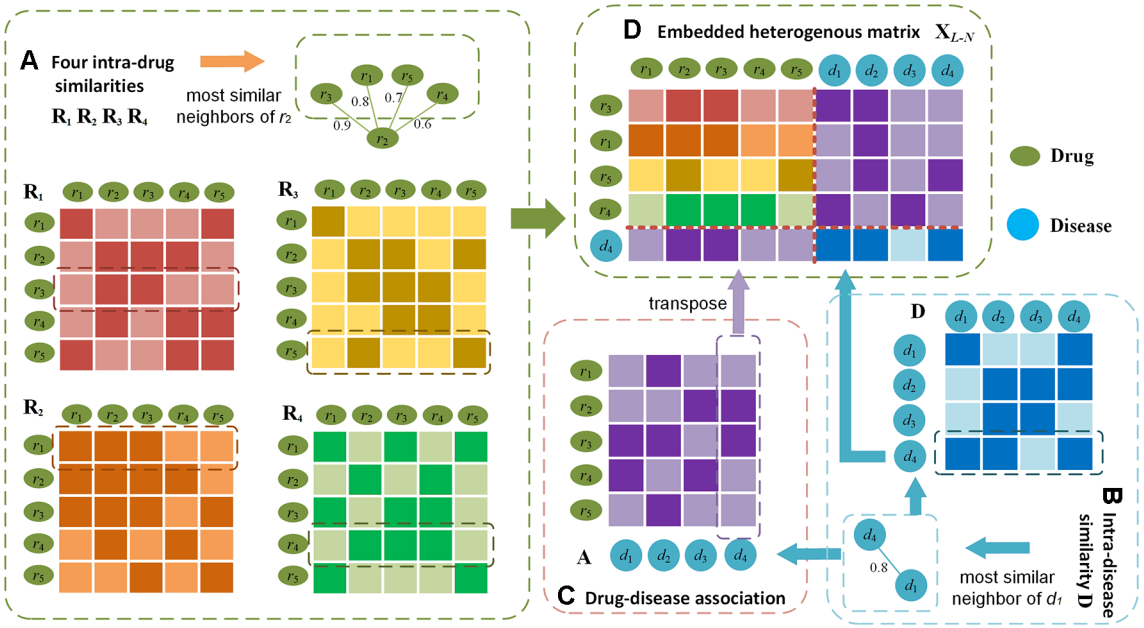

FIGURE 3 | Illustration of the embedding procedure for neighbouring heterogeneous matrix. Using drug r_2 and disease d_1 as an example, **(D)** the final matrix is obtained by finding the most similar neighbours (e.g. r_3,r_1,r_5,r_4) of r_2 calculated from **(A)** four intra-drug similarities respectively, the most similar neighbour (e.g. d_4) of drug d_1 by **(B)** intra-drug similarity matrix, and **(C)** drug-disease associations. In **(A)** and **(B)**, darker colours indicate higher similarities; in **(C)** darker colour represents the drug-disease association is available.

propagations. For instance, assume we find that drug r_2 likes r_3 the most in $\boldsymbol{R}_1$, r_1 in $\boldsymbol{R}_2$,r_5 in $\boldsymbol{R}_3$, and r_4 in $\boldsymbol{R}_4$(**Figure 3A**), and d_1 likes d_4 the most in D (as shown in **Figure 3B**). In the embedded matrix $\boldsymbol{X}_{L\text{-}N}$, the left part indicates that all $r_2^{'}s$ most similar neighbours (r_3, r_1, r_5, r_4) are very similar to r_2 and r_3. Because d_4 is associated with bridging drugs r_2 and r_3 based on A (**Figure 3C**), it can be inferred that there is a high probability that r_2 and d_1 are associated. The right part shows that the majority of $r_2^{'}s$ most similar neighbours are related with d_2. As $d_1^{'}s$ most similar neighbour d_4 is closely related to the bridging disease d_2 by $\boldsymbol{D}$, it can be derived that d_1 is probably related with r_2.

HeteroDualNet for Association Prediction

The architecture of HeterDualNet is given in **Figure 4**. The hetero-layer and hetero-layer-N are obtained by zero padding heterogenous matrices $\boldsymbol{X}_L$ and $\boldsymbol{X}_{L\text{-}N}$. One branch in the dual CNN model alternates two convolution and two pooling operations over hetero-layer (**Figure 4A**), the other branch is built where hetero-layer-N is convolved and pooled for neighbouring feature representations (**Figure 4B**). These two branches are connected by a fully connected network to achieve the final association score between r_i and d_k (**Figure 4C**). Same network settings are used in the two branches, thus we introduce the branch with hetero-layer in detail.

Convolutional module on hetero-layer. The heterogeneous matrix $\boldsymbol{X}_L$ is firstly padded with zeros to preserve the marginal information of matrices. In the first convolutional layer, we set N_1 filters where each filter is with width and length of $n_{w_{c1}}$ and $n_{l_{c1}}$. The hetero-layer is thus denoted as $\boldsymbol{V}_1 \in \mathbb{R}^{(5+2l)\times(N_r+N_d+2p)}$ where $l=\left(n_{w_{c1}}-1\right)/2$ $p=\left(n_{l_{c1}}-1\right)/2$. The case when N_1=3, $n_{w_{c1}}=3$, and $n_{l_{c1}}=5$ is illustrated as an example in **Figure 4A**. The weight parameter matrix of a n-th filter in the first convolutional layer is denoted by $\boldsymbol{W}_{1,n} \in \mathbb{R}^{n_{w_{c1}}\times n_{l_{c1}}}$, $n\in[1,N_1]$. The step size of a sliding window is set to be 1×1. The output of the first convolutional layer is obtained as $\boldsymbol{S}_1 \in \mathbb{R}^{N_1\times 5\times(N_r+N_d)}$ where $\boldsymbol{S}_{1,n} \in \mathbb{R}^{5\times(N_r+N_d)}$ is the n-th output after V_1 is scanned by the n-th filter as

$$\boldsymbol{S}_{1,n}=\begin{bmatrix} \boldsymbol{S}_{1,n(1,1)} & \boldsymbol{S}_{1,n(1,2)} & \cdots & \boldsymbol{S}_{1,n(1,N_r+N_d)} \\ \boldsymbol{S}_{1,n(2,1)} & \boldsymbol{S}_{1,n(2,2)} & \cdots & \boldsymbol{S}_{1,n(2,N_r+N_d)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{1,n(5,1)} & \boldsymbol{S}_{1,n(5,2)} & \cdots & \boldsymbol{S}_{1,n(5,N_r+N_d)} \end{bmatrix} \tag{5}$$

where $S_{1,n(i,j)}$ is the element in the i-th row and the j-th column of $S_{1,n}$ as:

$$\boldsymbol{S}_{1,n(i,j)}=g\left(\boldsymbol{V}_1^{'}\cdot\boldsymbol{W}_{1,n}+b_{1,n}\right) \tag{6}$$

where $b_{1,n}$ is the bias, "g" denotes the dot product, and g is a ReLu function. $V_{1(i,j)}$ is the element in the i-th row and the j-th column of V_1. When the filter slides to the position where $V_{1(i,j)}$ is the center point, $\boldsymbol{V}_{1(i,j)}^{'} \in \mathbb{R}^{n_{w_{c1}}\times n_{l_{c1}}}$ is formed by all the elements in the filter window as follow

$$\boldsymbol{V}_{1(i,j)}^{'}=\begin{bmatrix} \boldsymbol{V}_{1(i-1,j-2)} & \boldsymbol{V}_{1(i-1,j-1)} & \boldsymbol{V}_{1(i-1,j)} & \boldsymbol{V}_{1(i-1,j+1)} & \boldsymbol{V}_{1(i-1,j+2)} \\ \boldsymbol{V}_{1(i,j-2)} & \boldsymbol{V}_{1(i,j-1)} & \boldsymbol{V}_{1(i,j)} & \boldsymbol{V}_{1(i,j+1)} & \boldsymbol{V}_{1(i,j+2)} \\ \boldsymbol{V}_{1(i+1,j-2)} & \boldsymbol{V}_{1(i+1,j-1)} & \boldsymbol{V}_{1(i+1,j)} & \boldsymbol{V}_{1(i+1,j+1)} & \boldsymbol{V}_{1(i+1,j+2)} \end{bmatrix} \tag{7}$$

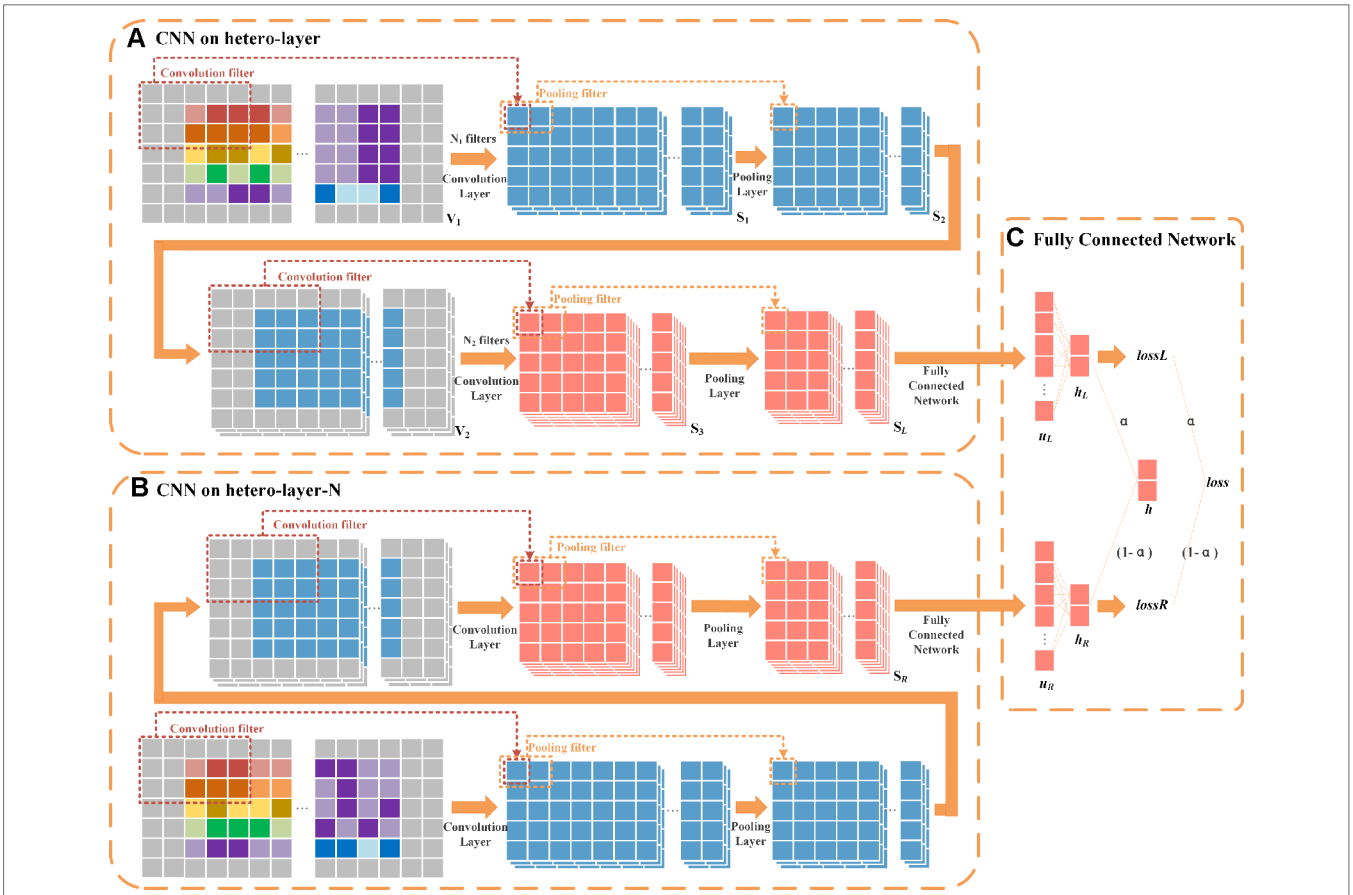

FIGURE 4 | Schematic diagram of HeteroDualNet. **(A)** One branch over hetero-layer of drug-disease characteristics and **(B)** one branch over the neighbouring heterogeneous layer (hetero-layer-N) are connected by **(C)** an integration module for final association score prediction. Three 3×5 filters in 1st convolution, six 3×5 filters in 2nd convolution, a sliding window of 1 × 2 in 1st and 2nd pooling are used for illustration.

We set the width and length of the sliding window in the first pooling layer as $n_{w_{p1}}$ and $n_{l_{p1}}$ ($n_{w_{p1}}=1$ and $n_{l_{p1}}=2$ as an example in **Figure 4**) and the step size as. The output of the first pooling $\boldsymbol{S}_2 \in \mathbb{R}^{N_1\times 5\times (N_r+N_d)/2}$. is obtained by a max-pooling operation where the n-th output $\boldsymbol{S}_{2,n} \in \mathbb{R}^{5\times (N_r+N_d)/2}$ is

$$\boldsymbol{S}_{2,n}=\begin{bmatrix} \boldsymbol{S}_{2,n(1,1)} & \boldsymbol{S}_{2,n(1,2)} & \cdots & \boldsymbol{S}_{2,n(1,(N_r+N_d)/2)} \\ \boldsymbol{S}_{2,n(2,1)} & \boldsymbol{S}_{2,n(2,2)} & \cdots & \boldsymbol{S}_{2,n(2,(N_r+N_d)/2)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{2,n(5,1)} & \boldsymbol{S}_{2,n(5,2)} & \cdots & \boldsymbol{S}_{2,n(5,(N_r+N_d)/2)} \end{bmatrix} \tag{8}$$

where $\boldsymbol{S}_{2,n(i,j)}$ is the maximum value between $\boldsymbol{S}_{1,n(i,2j-1)}$ and $\boldsymbol{S}_{1,n(i,2j)}$ defined as

$$\boldsymbol{S}_{2,n(i,j)}=max\left(\boldsymbol{S}_{1,n(i,2j-1)},\boldsymbol{S}_{1,n(i,2j)}\right) \tag{9}$$

By padding $S_{2,n}$ with zeros, V_2 is obtained as $\boldsymbol{V}_2 \in \mathbb{R}^{(5+2l)\times(N_r+N_d+2p)}$ where $l=\left(n_{w_{c2}}-1\right)/2$ and $p=\left(n_{l_{c2}}-1\right)/2$. The number of filters is set as N_2 in the second convolution. The output of the second convolution is obtained as $\boldsymbol{S}_3 \in \mathbb{R}^{N_2\times 5\times (N_r+N_d)/2}$. In the second pooling layer, we set the width and length of the sliding window as $n_{w_{p2}}$ and $n_{l_{p2}}$, and the step size as $n_{w_{p2}}\times n_{l_{p2}}$. For instance, the case when $N_2 = 6, n_{w_{p2}}=1$ and $n_{l_{p2}}=2$ is illustrated as an example in **Figure 4**. The output of the second pooling is obtained as $\boldsymbol{S}_4 \in \mathbb{R}^{N_2\times 5\times (N_r+N_d)/4}$ which is also the final output. Let $\boldsymbol{S}_L$ represent the final output of this branch, $\boldsymbol{S}_L = \boldsymbol{S}_4$.

Convolutional module on hetero-layer-N. The settings of convolution and pooling operations on hetero-layer-N is the same as the above branch. Let S_R denote the final output given $\boldsymbol{X}_{L-N}$ as inputs, $\boldsymbol{S}_R \in \mathbb{R}^{N_2\times 5\times (N_r+N_d)/4}$.

Final integration module. The integration of two branches is obtained by firstly flattening $\boldsymbol{S}_L$ and $\boldsymbol{S}_R$ as vectors $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathbb{R}^{1\times(N_2\times 5\times (N_r+N_d)/4)}$. u_L and u_R are then fed into a fully connected layer (as shown in **Figure 4C**).

The association score $\boldsymbol{h}_L \in \mathbb{R}^{2\times 1}$ between drug r_i r_i and disease d_k in one branch is obtained as

$$\boldsymbol{h}_L = softmax\left(\boldsymbol{W}_L\boldsymbol{u}_L^{\mathrm{T}}+\boldsymbol{b}_L\right) \tag{10}$$

where $\boldsymbol{W}_L \in \mathbb{R}^{2\times(5\times(N_r+N_d)/4\times n_2)}$ is the weight parameter matrix, and b_L is a bias vector. $h_L(1)$ contains the probability that r_i is

associated with d_k and $h_L(2)$ is the probability that r_i and d_k are not associated. Similarly, the association score h_R of the other branch is calculated by

$$\boldsymbol{h}_R = softmax\left(\boldsymbol{W}_R \boldsymbol{u}_R^{\mathrm{T}} + \boldsymbol{b}_R\right) \tag{11}$$

The final association score $\boldsymbol{h}$ is obtained by a weighted fusion of $\boldsymbol{h}_L$ and $\boldsymbol{h}_R$ as

$$\boldsymbol{h} = \alpha \boldsymbol{h}_L + (1-\alpha)\boldsymbol{h}_R,\ s.t.\ 0 \le \alpha \le 1 \tag{12}$$

where α is a regulation parameter to balance the contributions of two branches. Let *lossL* and *lossR* denote the losses of two branches as:

$$lossL = \min \left\| \boldsymbol{h}_L - \boldsymbol{y} \right\|_F^2,\ lossR = \min \left\| \boldsymbol{h}_R - \boldsymbol{y} \right\|_F^2 \tag{13}$$

where $\boldsymbol{y} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$ is the probability that drug r_i and disease d_k are associated. If r_i and d_k are associated, y_0=0 and y_1=1, otherwise y_0=1 and y_1=0. The final loss *loss* is obtained by

$$loss = min\ \alpha \left\| \boldsymbol{h}_L - \boldsymbol{y} \right\|_F^2 + (1-\alpha) \left\| \boldsymbol{h}_R - \boldsymbol{y} \right\|_F^2 \tag{14}$$

where the regulation parameter α is the same as that in Equation 12. With the network architecture and loss function, the parameters are randomly initialized and adjusted in the training process until the loss function is minimized. Given three types of drug-disease matrices, the final drug-disease association score can be predicted by the trained HeteroDualNet model.

In order to reduce the impact of overfitting which is caused by the number of parameters in the proposed model based on dual CNN, we adopt the widely used dropout strategy to prevent the overfitting of HeteroDualNet. During each iteration process for training the model, HeteroDualNet randomly ignores some neurons to ensure that the trained model will have a good generalization ability.

EXPERIMENTAL EVALUATIONS AND DISCUSSIONS

Experimental Setup

The drug-disease samples with known associations are regarded as one class (L_1), while those pairs with unknown associations are considered as the other class (L_2). In total, there are 3051 L_1 samples, and 763*681-3051 = 516552 L_2 samples. Because L_1 and L_2 samples are largely imbalanced, undersampling strategy is used to address this issue. We divided the data into two subsets. One subset A is composed of 3051 L_1 samples and 3051 L_2 samples, while the second subset B contains the remaining 516552 – 3051 L_2 samples.

Five-fold cross-validation is performed to evaluate the prediction performance of HeteroDualNet and other compared models. The same training and testing data are used for the training and testing of the models. In each round of validation, the samples in subset A are equally divided into five parts where four parts are used as the training dataset, and one part together with subset B are used for testing.

As the calculation of the *4*-th intra-drug similarities matrix $\boldsymbol{R}_4$ involves drug-disease association matrix $\boldsymbol{A}$ and intra-disease matrix $\boldsymbol{D}$ to ensure that there is no testing data information in the training dataset, $\boldsymbol{R}_4$ is recalculated by removing drug-disease samples that appear in training in each round of validation.

Comparison Methods and Evaluation Metrics

To evaluated the contributions of the proposed HeteroDualNet architecture and heterogenous drug-disease similarity representations, our model is compared with other four prediction methods including TL_HGBI (Wang et al., 2014b), MBiRW (Luo et al., 2016), LRSSL (Liang et al., 2017), and SCMFDD (Zhang et al., 2018). LRSSL is based on three drug features without considering neighbouring information and our proposed fourth intra-drug similarity from drug-related disease correlations. MBiRW used only one type of drug feature. SCMFDD and TL_HGBI used matrix decomposition and heterogeneous networks, but they didn't consider neighbouring information and multiple features.

The prediction performance is comprehensively evaluated by true positive rate (*TPR*), false positive rate (*FPR*), the Receiver Operating Characteristic (ROC) area under curve (ROC AUC), the Precision-Recall area under curve (PR AUC) and recall rate under different top *k* values. TPR and FPR are calculated as

$$TPR = \frac{TP}{TP+FN},\ FPR = \frac{FP}{TN+FP}, \tag{15}$$

where *TP* (*FN)* is the number of positive samples that are correctly identified (misidentified), *TN* (*FP)* is the number of correctly identified (misidentified) negative samples. A sample is regarded as a positive sample when its predicted association score is greater than a threshold θ. If the testing sample's score is smaller than θ, it is identified as a negative sample. The values of *FPR* and *TPR* are calculated by setting different values of θ. The average ROC AUC value of all the evaluated drugs is used as the overall prediction performance of a method.

Since two classes are heavily imbalanced, the evaluation by PR AUC is more appropriate than ROC AUC in our study. Thus, PR AUC is also compared among different methods. *Precision* and *Recall* are defined by

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN} \tag{16}$$

where *Precision* represents the ratio between the number of correctly identified positive samples and all samples which are predicted to be positive samples, and *Recall* represents the ratio of the correctly identified positive samples to all the positive samples. Meanwhile, because the top-ranked results are of greater interest in real practices, which are often considered by biologists for further validation, we also calculate the recall rate in top *k* ranked results. The higher the recall rate for the top *k* disease, the more drug-related diseases can be predicted by the model.

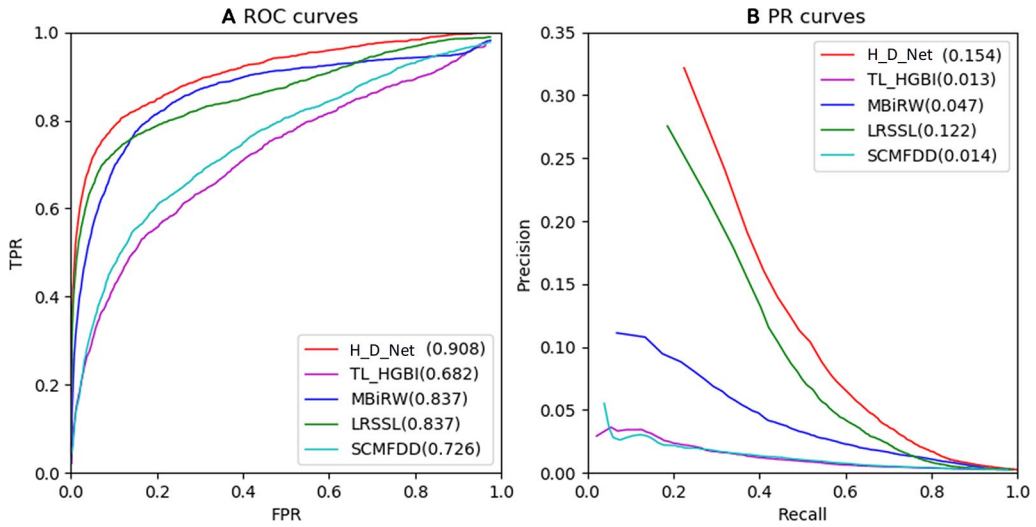

FIGURE 5 | Comparison between the proposed HeteroDudalNet model (H_D_Net) against four other methods by Receiver Operating Characteristic (ROC) **(A)** and Precision-Recall (PR) **(B)** curves.

Experimental Results and Discussion

The ROC and PR of all the methods using all the 763 drugs are shown in **Figure 5**. The AUC results are given in **Table 1**. As shown by **Figure 5A** and **Table 1**, our model achieved the highest AUC of 0.908 among all the methods in comparison, which is 7.1% greater than the second best MBiRW model, 18.2% higher than the SCMFDD method, and 22.6% higher than the TL_HGBI method. As shown by **Figure 5B** and **Table 1**, HeteroDudalNet achieved the best performance where PR AUC reached 0.154, which was 3.2%, 10.7%, 14%, and 14.1% better than the that of LRSSL, MBiRW, SCMFDD and TL_HGBI models respectively.

As shown by the ROC and PR evaluation results, HeteroDudalNet outperformed the second best LRSSL because of the integration of neighbouring information on drugs and diseases and the intra-drug similarity calculated by correlations of drug-related diseases. Compared with LRSSL which considered three types of drug features, the third best model MBiRW considered only one type of drug feature in an adopted a random walk-based model, which resulted in a much lower prediction score. Without considering neighbouring associations and multiple features, SCMFDD and TL_HGBI methods failed to achieve satisfactory prediction performance although they used matrix decomposition and heterogeneous networks.

The average performance over all the 763 drugs in terms of recall rate given different top *k* values is shown in **Figure 6**. The higher the recall rate for the top *k* diseases, the more drug-related diseases can be predicted by a computing model. When increasing the value of *k* from 30 to 240 with a step of 30, the average recall rate of our method is the best among all the models in comparison. When examining the top 30, 60 and 90 diseases, our model achieved recall rates of 69.2%, 77%, and 83.5%, and the second best was obtained by LRSSL with recall rates of 63.4%, 71.3%, and 77.7% respectively. The third-ranked model MBiRW performed slightly worse than LRSSL where the results were 52.9%, 66% and 74.2%. When *k* was increased from 90 to 240, MBiRW started to perform better than LRSSL and achieved its highest recall rate of 88.7% when k was 240, while our model obtained the best rate of 90.9% among all the methods. Overall, the top *k* recall rates of SCMFDD and TL_HGBI were significantly lower than the other techniques in comparison.

As shown by the top *k* recall rate test, our model achieved the best performance, which could be useful for biologists to conduct clinical experiments because the highest ranked list contains more real drug-disease associations. As shown by the results when *k* was smaller than 90, our model and LRSSL outperformed the other methods because of the consideration of multiple drug features. The comprehensive representation of drugs concerning similarities in various perspectives contributes to digging out

TABLE 1 | Receiver Operating Characteristic area under curve (ROC AUC) and Precision-Recall area under curve (PR AUC) of all the methods in comparison.

	Average performance on 763 drugs				
	HeteroDualNet	**TL_HGBI**	**MBiRW**	**LRSSL**	**SCMFDD**
ROC AUC	0.908	0.723	0.855	0.845	0.611
PR AUC	0.154	0.031	0.045	0.089	0.006

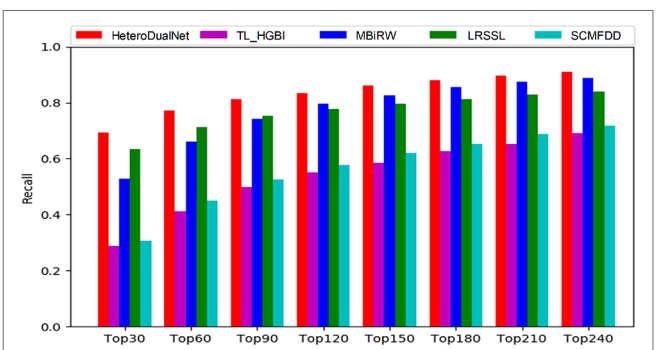

FIGURE 6 | The recalls across all the tested drugs at different top *k* cutoffs.

TABLE 2 | Top 10 related candidate diseases of ciprofloxacin, ceftriaxone, ofloxacin, ampicillin and cefotaxime.

Drug name	Rank	Disease name	Description	Rank	Disease	Description
ciprofloxacn	1	Pneumonia, Bacterial	CTD	6	Gram-Positive Bacterial Infections	CTD
	2	Salmonella Infections	CTD	7	Eye Infections, Bacterial	Literature (Marino et al., 2013)
	3	Bacterial Infections	CTD	8	Soft Tissue Infections	CTD
	4	Streptococcal Infections	DrugBank	9	Enterobacteriaceae Infections	CTD
	5	Gram-Negative Bacterial Infections	CTD	10	Helicobacter Infections	CTD
ceftriaxone	1	Gram-Negative Bacterial Infections	CTD	6	Haemophilus Infections	CTD
	2	Bacterial Infections	CTD, ClinicalTrials	7	Gram-Positive Bacterial Infections	CTD
	3	Septicemia	DrugBank	8	Skin Diseases, Infectious	DrugBank
	4	Respiratory Tract Infections	CTD	9	Wound Infection	ClinicalTrials
	5	Pseudomonas Infections	DrugBank	10	Eye Infections, Bacterial	DrugBank
ofloxacin	1	Eye Infections, Bacterial	ClinicalTrials, DrugBank	6	Pseudomonas Infections	CTD
	2	Gram-Negative Bacterial Infections	DrugBank	7	Bacterial Infections	CTD
	3	Sinusitis	CTD	8	Bacteroides Infections	DrugBank
	4	Streptococcal Infections	CTD	9	Gram-Positive Bacterial Infections	CTD
	5	Pneumonia, Bacterial	CTD	10	Enterobacteriaceae Infections	DrugBank
ampicillin	1	Pseudomonas Infections	unconfirmed	6	Proteus Infections	CTD
	2	Bacterial Infections	CTD	7	Septicemia	DrugBank
	3	Gram-Positive Bacterial Infections	CTD	8	Streptococcal Infections	CTD
	4	Gram-Negative Bacterial Infections	CTD	9	Wound Infection	CTD
	5	Pneumonia, Bacterial	CTD, ClinicalTrials	10	Enterobacteriaceae Infections	DrugBank
cefotaxime	1	Respiratory Tract Infections	CTD, ClinicalTrials	6	Enterobacteriaceae Infections	DrugBank
	2	Pseudomonas Infections	DrugBank	7	Gram-Positive Bacterial Infections	CTD, DrugBank
	3	Gram-Negative Bacterial Infections	CTD, DrugBank	8	Wound Infection	DrugBank
	4	Septicemia	DrugBank	9	Skin Diseases, Infectious	ClinicalTrials
	5	Bacterial Infections	CTD, ClinicalTrials	10	Osteomyelitis	CTD, ClinicalTrials

(1) CTD refers to the Comparative Toxicogenomics Database (CTD), which contains a manually managed drug-disease association. (2) DrugBank refers to the drug-disease association held in the DrugBank database, which collects experimental information of the drug. (3) ClinicalTrials means that the association of drugs with the disease is recorded in the online database ClinicalTrials.gov. (4) literature refers to the literature supporting the association of drugs with the disease. (5) unconfirmed means that there is no evidence that the drug is associated with the disease.

the potential associations between drugs and diseases. When $k > 90$, the number of common features between drug and disease may be decreasing when compared with smaller k values. Thus, considering only multiple features could not always guarantee a good prediction result. MBiRW performed better than LRSSL due to the consideration of global information in a random walk based network. By incorporating three drug characteristics, the calculated correlations between drug-related diseases as intra-drug similarities, and neighbouring information of similar drugs and diseases, our model achieved better results than LRSSL and MBiRW.

Case Studies of Five Drugs

To further evaluate and demonstrate the effectiveness of the proposed HeteroDualNet in finding reliable disease candidates of drugs, we conducted case studies of five drugs, including ciprofloxacin, ceftriaxone, ofloxacin, ampicillin and cefotaxime. Two public drug disease databases, Comparative Toxicogenomics Database (CTD) and DrugBank, were used to verify and confirm the predicted drug-disease associations by the proposed model. CTD is funded by the National Institute of Environmental Health Sciences which contains information of drugs and drugs' effects on diseases extracted from

published literature. DrugBank is supported by the Health Research Institute of Canada, the Alberta Innovation-Health Solutions and Metabolic Innovation Center. Drugs' clinical trial information can be found in DrugBank, which includes drugs and diseases in experiments.

For each of the five drugs, we ranked the predicted diseases according to the relevance scores in descending order. The top 10 ranked diseases are used for verification and listed in **Table 2**. Among all the 50 diseases, 31 disease-drug association information can be found in CTD, and 17 association information can be found in the DrugBank as shown in **Table 2**. The results demonstrated that the predicted candidate diseases are indeed associated with the corresponding drugs. Also, in the CTD database, the association between Ciprofloxacin and Eye Infections, Bacterial can be found in the literature. For the two diseases which cannot be found in CTD and DrugBank, one of them can be verified by ClinicalTrials.gov (https://clinicaltrials.gov/) which records a wealth of clinical research information on various drugs and related diseases by National Institutes of Health (NIH) and the Food and Drug Administration (FDA). Therefore, there is only one disease candidate of drug ampicillin, which is Pseudomonas Infections, cannot be proved by the three databases and is labelled as unconfirmed in **Table 2**. The case studies demonstrated that our model can be used as an effective tool to predict the relations between drugs and diseases. At the same time, it has the capacity to provide computer-aided guidance for biologists in clinical trials.

The future direction for developing userful and powerful computerized prediction methods include establishing web-servers to enable public assessibility (Cheng et al., 2017; Cheng et al., 2018; Xiao et al., 2019; Chou, 2019a; Chou, 2019b). Our future work include providing a web-server for the proposed model to increase the impact of computational model in bioinformatics, medical science and medicinal chemistry.

CONCLUSION

We present a novel HeteroDualNet model for drug-disease association prediction. Our model incorporates three kinds of drug features, a newly introduced intra-drug similarity based on correlations of drug-related diseases, and neighbouring information of drugs and diseases by constructing embedded drug-disease heterogenous matrices and dual branches in a deep neural network. The evaluation of public dataset and comparison with other four published models demonstrated the improved prediction performance in terms of ROC AUC, PR AUC, and recall rate at top *k*. Case studies of five drugs further proved the effectiveness of our model in finding potential relevant diseases of drugs as validated by database records or literature. Our model can be used as an effective tool to predict the associations between drugs and diseases and provide computer-aided guidance for biologists in clinical trials.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found at https://github.com/LiangXujun/LRSSL.

AUTHOR CONTRIBUTIONS

PX, HC, and TS conceived the prediction method, and they wrote the paper. NS and TS developed computer programs. PX, TZ, and TS analyzed the results, and PX, HC, and NS revised the paper.

FUNDING

The work was supported by the Natural Science Foundation of China (61972135), the Natural Science Foundation of Heilongjiang Province (LH2019F049, LH2019A029), the China Postdoctoral Science Foundation (2019M650069), the Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHL-Q18104), the Fundamental Research Foundation of Universities in Heilongjiang Province for Technology Innovation (KJCX201805), and the Fundamental Research Foundation of Universities in Heilongjiang Province for Youth Innovation Team (RCYJTD201805).

REFERENCES

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

Awais, M., Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., and Chou, K. C. (2019). iPhosH-PseAAC: identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 1–19. doi: 10.1109/TCBB.2019.2919025

Chen, H., Zhang, H., Zhang, Z., Cao, Y., and Tang, W. (2015). Network-based inference methods for drug repositioning. *Comput. Math. Methods Med.* 2015, 130620–130620. doi: 10.1155/2015/130620

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PloS Comput. Biol.* 12, e1004975. doi: 10.1371/journal.pcbi.1004975

Cheng, X., Lin, W. Z., Xiao, X., and Chou, K. C. (2018). pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics* 35 (3), 398–406. doi: 10.1093/bioinformatics/bty628

Cheng, X., Xiao, X., and Chou, K. C. (2017). pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. *Bioinformatics* 34 (9), 1448–1456. doi: 10.1093/bioinformatics/btx711

Chiang, A. P., and Butte, A. J. (2009). Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* 86, 507–510. doi: 10.1038/clpt.2009.103

Chou, K. C., and Shen, H. B. (2009). Recent advances in developing web-servers for predicting protein attributes. *Natural Sci.* 1 (02), 63. doi: 10.4236/ns.2009.12011

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273 (1), 236–247. doi: 10.1016/j.jtbi.2010.12.024

Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* 11 (3), 218–234. doi: 10.2174/1573406411666141229162834

Chou, K. C. (2017). An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Topics Med. Chem.* 17 (21), 2337–2358. doi: 10.2174/1568026617666170414145508

Chou, K. C. (2019a). Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.*, 1–16. doi: 10.1007/s10989-019-09893-5

Chou, K. C. (2019b). Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* 26, 4918–4943. doi: 10.2174/0929867326666190507082559

Dakshanamurthy, S., Issa, N. T., Assefnia, S., Seshasayee, A., Peters, O. J., Madhavan, S., et al. (2012). Predicting new indications for approved drugs using a proteochemometric method. *J. Med. Chem.* 55, 6832–6848. doi: 10.1021/jm300576q

Ehsan, A., Mahmood, M. K., Khan, Y. D., Barukab, O. M., Khan, S. A., and Chou, K. C. (2019). iHyd-PseAAC (EPSV): identifying hydroxylation sites in proteins by extracting enhanced position and sequence variant feature via chou's 5-step rule and general pseudo amino acid composition. *Curr. Genomics* 20 (2), 124–133. doi: 10.2174/1389202920666190325162307

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, 496–496. doi: 10.1038/msb.2011.26

Grabowski, H. (2004). Are the economics of pharmaceutical research and development changing?: productivity, patents and political pressures. *Pharmacoeconomics* 22, 15–24. doi: 10.2165/00019053-200422002-00003

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nat. Biotechnol.* 32, 40. doi: 10.1038/nbt.2786

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* 93, 335–341. doi: 10.1038/clpt.2013.1

Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., and Chou, K. C. (2019). SPalmitoylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Analytical Biochem.* 568, 14–23. doi: 10.1016/j.ab.2018.12.019

Iwata, H., Sawada, R., Mizutani, S., and Yamanishi, Y. (2015). Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J. Chem. Inf. Model* 55, 446–459. doi: 10.1021/ci500670q

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). 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-disease associations based on data integration using sparse subspace learning. *Bioinformatics* 33, 1187–1196. doi: 10.1093/bioinformatics/btw770

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

Marino, A., Santoro, G., Spataro, F., Lauriano, E. R., Pergolizzi, S., Cimino, F., et al. (2013). Resveratrol role in Staphylococcus aureus -induced corneal inflammation. *Pathog. Dis.* 68, 61–64. doi: 10.1111/2049-632X.12046

Martinez, 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

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221. doi: 10.1093/nar/gku1243

Nosengo, N. (2016). Can you teach old drugs new tricks? *Nature* 534, 314–316. doi: 10.1038/534314a

Padhy, B. M., and Gupta, Y. K. (2011). Drug repositioning: re-investigating existing drugs for new therapeutic indications. *J. Postgrad. Med.* 57, 153–160. doi: 10.4103/0022-3859.81870

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* 9, 203–214. doi: 10.1038/nrd3078

Pritchard, J.-L. E., O'mara, T. A., and Glubb, D. M. (2017). Enhancing the Promise of Drug Repositioning through Genetics. *Front. Pharmacol.* 8, 896–896. doi: 10.3389/fphar.2017.00896

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2018). Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discovery* 18, 41. doi: 10.1038/nrd.2018.168

Shim, J. S., and Liu, J. O. (2014). Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int. J. Biol. Sci.* 10, 654–663. doi: 10.7150/ijbs.9224

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77. doi: 10.1126/scitranslmed.3001318

Tamimi, N. A., and Ellis, P. (2009). Drug development: from concept to marketing! *Nephron Clin. Pract.* 113, c125–c131. doi: 10.1159/000232592

Uniprot, C. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148. doi: 10.1093/nar/gkp846

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, F., Zhang, P., Cao, N., Hu, J., and Sorrentino, R. (2014a). Exploring the associations between drug side-effects and therapeutic indications. *J. BioMed. Inform* 51, 15–23. doi: 10.1016/j.jbi.2014.03.014

Wang, W., Yang, S., Zhang, X., and Li, J. (2014b). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30, 2923–2930. doi: 10.1093/bioinformatics/btu403

Wang, Y., Chen, S., Deng, N., and Wang, Y. (2013). Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PloS One* 8, e78518. doi: 10.1371/journal.pone.0078518

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633. doi: 10.1093/nar/gkp456

Wu, C., Gudivada, R. C., Aronow, B. J., and Jegga, A. G. (2013). Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol.* 7 Suppl 5, S6–S6. doi: 10.1186/1752-0509-7-S5-S6

Xiao, X., Cheng, X., Chen, G., Mao, Q., and Chou, K. C. (2019). pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics* 111 (4), 886–892. doi: 10.1016/j.ygeno.2018.05.017

Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., and Shen, T. (2019). Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*. 35, 4108–4119. doi: 10.1093/bioinformatics/btz182

Yang, L., and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS One* 6, e28025. doi: 10.1371/journal.pone.0028025

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 Suppl 2, S2. doi: 10.1186/1755-8794-8-S2-S2

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 Bioinf.* 19, 233–233. 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.