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

Ping Xuan¹, Hui Cui², Tonghui Shen¹*, Nan Sheng¹ and Tiangang Zhang³*

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

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.,
Drug repositioning is playing an essential role in the state-of-the-art drug R&D strategies. (Paul et al., 2010). 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, toxicity 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 data set, 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 heterogenous layer with neighboring 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.

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 $D(i,j) \in [0,1]$, where $D$ is the intra-disease similarity matrix and $N_d^d$ 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 $F_i \in \mathbb{R}^{N_d^f \times N_F}$, where $N_d^f$ is the number of chemical substructure features, and $N_F$ is the number of drugs. Similarly, the feature matrix of protein domain and drugs is $F_i \in \mathbb{R}^{N_d^p \times N_F}$, and that of protein annotation and drugs is $F_i \in \mathbb{R}^{N_d^a \times N_F}$, where $N_d^p$ is the number of target protein domain feature and $N_d^a$ 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, $N_d^f = 881$, $N_d^p = 1426$ and $N_d^a = 4,447$. Let $f_i$, be the feature vector of $i$-th drug $r_i$ in the $t$-th feature matrix $F_t (1 \leq t \leq 3)$, the similarity $R_{t}(i,j)$ between drugs $r_i$ and $r_j$ in terms of feature $t$ is calculated by cosine similarity measurement as

$$R_t(i,j) = \frac{f_{i,t} \cdot f_{j,t}}{\|f_{i,t}\| \cdot \|f_{j,t}\|}$$ (1)
where \( R_{(i,j)} \in [0,1] \) and higher values indicate higher similarity between a pair of drugs.

The fourth intra-drug similarities matrix \( R_{\alpha} \in \mathbb{R}^{N_{\alpha} \times N_{\alpha}} \) 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 \( D_{r} = \{d_{r} | k \in [1,N_{r}]\} \) and intra-disease similarity matrix \( D \) if \( i \)-th drug \( r_{i} \) is associated with a subset of diseases \( D_{i} \subseteq D \), and drug \( j \) is related to a disease subset \( D_{j} \), the similarity \( R_{\alpha}(i,j) \) between \( i \)-th and \( j \)-th drugs can be obtained by calculating the similarity between \( D_{i} \) and \( D_{j} \) as proposed in our previous work (Xuan et al., 2019) by

\[
R_{\alpha}(i,j) = \frac{\sum_{k \in \text{num}(D_{i})} \max(D(d_{r_{i},d_{r_{j}}})) + \sum_{k \in \text{num}(D_{j})} \max(D(d_{r_{i},d_{r_{j}}}))}{\text{num}(D_{i}) + \text{num}(D_{j})}
\]

where \( \text{num}(D_{i}) \) denotes the number of elements in \( D_{i} \), \( d_{r_{i}} \) denotes the \( k \)-th disease related with drug \( r_{i} \), \( d_{r_{j}} \) denotes all the related diseases of drug \( r_{j} \), and \( \max(D(d_{r_{i},d_{r_{j}}})) \) is the maximum similarity between drug \( r_{i} \) and \( k \)-th related disease and all the related diseases of \( r_{j} \). Similarly, \( \max(D(d_{r_{i},d_{r_{j}}})) \) denotes the maximum similarity between drug \( r_{i} \) and \( k \)-th related disease and all the associated diseases of \( r_{j} \). The final similarity between \( r_{i} \) and \( r_{j} \) is obtained by the average maximum similarities between diseases in their relevant disease subsets \( D_{i} \) and \( D_{j} \).

**Drug-Disease Association Matrix**
The drug-disease association matrix is denoted by \( A \in \mathbb{R}^{N_{\alpha} \times N_{\alpha}} \) 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 \( R_{\alpha} \), drug-disease association matrix \( A \) and intra-disease matrix \( D \), the heterogeneous matrix \( X_{L} \) of drug \( r_{i} \in [1,N_{r}] \) and disease \( d_{k} \in [1,N_{d}] \) is obtained by the following embedding procedures.

Firstly, row vectors \( R_{\alpha}(i,*) \) are combined sequentially as \( X_{L,11} = [R_{\alpha}(i,*)] \), \( R_{\alpha}(i,*) \), \( R_{\alpha}(i,*) \), \( R_{\alpha}(i,*) \) where \( R_{\alpha}(i,*) \) denotes the \( i \)-th row in an intra-drug similarity matrix \( R_{\alpha} \) which records the \( i \)-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 \( A^{T}(\cdot,k) \) is concatenated under \( R_{\alpha}(i,*) \) as \( X_{L,21} \) where \( X_{L,21} \) is the \( k \)-th column of \( A \) which contains the associations between \( d_{k} \) and all the drugs. Thirdly, \( A(\cdot,k) \) is repeated four times and spliced to the right of each row in \( X_{L,21} \) as \( X_{L,22} = [A(\cdot,k)] \), \( A(\cdot,k) \), \( A(\cdot,k) \), \( A(\cdot,k) \) where \( A(\cdot,k) \) denotes the \( k \)-th row of \( A \) which includes the associations between \( r_{i} \) and all the diseases. Lastly, \( D(k,\cdot) \) is spliced under \( X_{L,11} \) where \( D(k,\cdot) \) is the \( k \)-th row of \( D \) containing the similarities between \( d_{k} \) and all the diseases. The final embedded matrix \( X_{L} \in \mathbb{R}^{5d(N_{d}+N_{\alpha})} \) of drug \( r_{i} \) and disease \( d_{k} \) is formed as

\[
X_{L} = \left[ \begin{array}{cc}
X_{L,11} & X_{L,12} \\
X_{L,21} & X_{L,22}
\end{array} \right] = \left[ \begin{array}{cc}
R_{\alpha}(i,*) & A(i,*) \\
R_{\alpha}(i,*) & A(i,*) \\
R_{\alpha}(i,*) & A(i,*) \\
R_{\alpha}(i,*) & A(i,*) \\
A^{T}(\cdot,k) & D(k,\cdot)
\end{array} \right]
\]

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 \( r_{3} \) and disease \( d_{4} \) whose association is unknown as an example. If \( r_{3} \) is very similar to \( r_{4} \) and \( r_{4} \) (as shown in Figure 2A), \( r_{3} \) and \( r_{4} \) are closely associated with \( d_{4} \) (Figure 2B), it can be inferred that \( r_{3} \) is more likely to be associated with \( d_{4} \). Alternatively, if \( d_{4} \) is similar to \( d_{1} \) (shown in Figure 2C), and \( d_{4} \) is related to \( r_{3} \) (Figure 2B), a high possibility that \( r_{3} \) is associated with \( d_{4} \) 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 \( r_{\alpha} \), \( r_{\alpha} \), \( r_{p} \),
and \( r_q \) which are the most similar neighbours of drug \( r_i \) in \( R_1 \), \( R_2 \), \( R_3 \) and \( R_4 \) respectively. We also find \( d_p \), the most similar neighbour of \( d_k \), in \( D \). Similar with \( X_{L,N,11} \), the \( m \)-th row of \( R_1 \), \( n \)-th row of \( R_2 \), \( p \)-th row of \( R_3 \) and \( q \)-th row of \( R_4 \) are combined from top to bottom to form \( X_{L,N,11} \). Secondly, the \( l \)-th column of \( A \) indicating the association between the most similar disease \( d_l \) and all the drugs is transposed and concatenated under \( X_{L,N,11} \) as \( X_{L,N,21} \). Thirdly, row vectors \( A(m,*), A(n,*), A(p,*), A(q,*) \) are spliced to the right of each row in \( X_{L,N,11} \), where \( A(m,*), A(n,*), A(p,*), A(q,*) \) indicate the associations between drugs \( r_m, r_n, r_p, r_q \) 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 \( X_{L,N,21} \). In such a way, the final embedding of most similar neighbours of \( r_i \) and \( d_k \) is formed as 

\[
X_{L,N} \in \mathbb{R}^{(n_{m}+n_{n}+n_{p}+n_{q})}:
\]

\[
X_{L,N} = \left[ \begin{array}{cc} X_{L-N,11} & X_{L-N,12} \\ X_{L-N,21} & X_{L-N,22} \end{array} \right] \end{array} \right] = \left[ \begin{array}{cc} R_1(m,*) & A(l,*) \\ R_2(n,*) & A(l,*) \\ R_3(p,*) & A(l,*) \\ R_4(q,*) & A(l,*) \\ A^T(*,l) & D(l,*) \end{array} \right]
\]

In \( X_{L,N} \), 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 \( d_2 \) as an example, (D) the final matrix is obtained by finding the most similar neighbours (e.g. \( r_2, r_3, r_5, r_6 \)) of \( r_2 \) calculated from (A) four intra-drug similarities respectively, the most similar neighbour (e.g. \( d_4 \)) of drug \( d_2 \) 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 that drug \( r_2 \) likes \( r_3 \) the most in \( R_1 \), \( r_1 \) in \( R_2 \), \( r_5 \) in \( R_3 \), and \( r_6 \) in \( R_4 \) (Figure 3A), and \( d_1 \) likes \( d_4 \) the most in \( D \) (as shown in Figure 3B). In the embedded matrix \( X_{L,N} \), the left part indicates that all \( r_5 \) most similar neighbours (\( r_3, r_1, r_5, r_6 \)) are very similar to \( r_2 \) and \( r_5 \). 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 \( d_4 \) are associated. The right part shows that the majority of \( r_5 \) most similar neighbours are related with \( d_4 \). As \( d_5 \) most similar neighbour \( d_4 \) is closely related to the bridging disease \( d_4 \) by \( D \), 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 \( X_L \) and \( X_{L,N} \). 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 \( d_i \) (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 \( X_L \) is firstly padded with zeros to preserve the marginal information of matrices. In the first convolutional layer, we set \( N_l \) filters where each filter is with width and length of \( n_{w,l} \) and \( n_{l,h} \). The hetero-layer is thus denoted as \( V_l \in \mathbb{R}^{(n_{w,l} \times n_{l,h}) \times (n_L \times n_L + p)} \) where \( l = (n_{w,l} - 1) / 2 \) and \( p = (n_{l,h} - 1) / 2 \). The case when \( N_l = 3 \), \( n_{w,l} = 3 \), and \( n_{l,h} = 5 \) 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 \( W_{l,n} \in \mathbb{R}^{n_{w,l} \times n_{l,h}} \), \( n \in [1, N_l] \). The step size of a sliding window is set to be 1 × 1. The output of the first convolutional layer is obtained as \( S_L \in \mathbb{R}^{N_L \times (N_L \times N_l) \times n} \) where \( S_L \in \mathbb{R}^{3 \times (3 \times 3)} \) is the \( n \)-th output after \( V_l \) is scanned by the \( n \)-th filter as

\[
S_{L,n} = \left[ \begin{array}{cccc}
S_{L,n(1,1)} & S_{L,n(1,2)} & \cdots & S_{L,n(1,3)} \\
S_{L,n(2,1)} & S_{L,n(2,2)} & \cdots & S_{L,n(2,3)} \\
\vdots & \vdots & \ddots & \vdots \\
S_{L,n(3,1)} & S_{L,n(3,2)} & \cdots & S_{L,n(3,3)}
\end{array} \right]
\]

(5)

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

\[
S_{L,n(i,j)} = g(V_l \cdot W_{l,n} + b_{l,n})
\]

(6)

where \( b_{l,n} \) is the bias, “\( g \)” denotes the dot product, and \( g \) is a ReLu function. \( V_{l(i,j)} \) is the element in the \( i \)-th row and the \( j \)-th column of \( V_l \). When the filter slides to the position where \( V_{l(i,j)} \) is the center point, \( V_{l(i,j)} \in \mathbb{R}^{n_{w,l} \times n_{l,h}} \) is formed by all the elements in the filter window as follow

\[
V_{l(i,j)} = \left[ \begin{array}{cccc}
V_{l(i-1,j-1)} & V_{l(i-1,j)} & V_{l(i-1,j+1)} & V_{l(i-1,j+2)} \\
V_{l(i+1,j-1)} & V_{l(i+1,j)} & V_{l(i+1,j+1)} & V_{l(i+1,j+2)} \\
V_{l(i,j-1)} & V_{l(i,j)} & V_{l(i,j+1)} & V_{l(i,j+2)} \\
V_{l(i,j+1)} & V_{l(i,j+2)} & V_{l(i,j+3)} & V_{l(i,j+4)}
\end{array} \right]
\]

(7)
HeteroDualNet for Drug-Disease Association Prediction

Xuan et al.

We set the width and length of the sliding window in the first pooling layer as \( n_{wp1} \) and \( n_{lp1} \) (as an example in Figure 4) and the step size as \( n_{wlpp2} \). The output of the first pooling is \( S_{2,1} \in \mathbb{R}^{5 \times (N + N_d/2)} \) obtained by a max-pooling operation where the \( n \)-th output \( S_{2,1,n} \in \mathbb{R}^{5 \times (N + N_d/2)} \) is

\[
S_{2,1,n} = \max \left( S_{1,1,(i,j-1)}, S_{1,1,(i,j)} \right) \quad (8)
\]

where \( S_{1,1,(i,j-1)} \) is the maximum value between \( S_{1,1,(i,j-1)} \) and \( S_{1,1,(i,j)} \) defined as

\[
S_{2,1,n} = \max \left( S_{1,1,(i,j-1)}, S_{1,1,(i,j)} \right) \quad (9)
\]

By padding \( S_{2,1,n} \) with zeros, \( V_2 \) is obtained as \( V_2 \in \mathbb{R}^{(5+2l) \times (N + N_d/2+p)} \) where \( l = (n_{wp1} - 1)/2 \) and \( p = (n_{lp1} - 1)/2 \). The number of filters is set as \( N_2 \) in the second convolution. The output of the second convolution is obtained as \( S_3 \in \mathbb{R}^{N_2 \times 5 \times (N + N_d/2)} \). In the second pooling layer, we set the width and length of the sliding window as \( n_{wp2} \) and \( n_{lp2} \), and the step size as \( n_{wlpp2} \). For instance, the case when \( N_2 = 6, n_{wp2} = 1 \) and \( n_{lp2} = 2 \) is illustrated as an example in Figure 4. The output of the second pooling is obtained as \( S_4 \in \mathbb{R}^{N \times 5 \times (N + N_d/2)} \) which is also the final output. Let \( S_L \) represent the final output of this branch, \( S_L = S_4 \).

**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 \( X_{L-N} \) as inputs, \( S_R \in \mathbb{R}^{N_2 \times 5 \times (N + N_d/2)} \).

**Final integration module.** The integration of two branches is obtained by firstly flattening \( S_L \) and \( S_R \) as vectors \( u_L, u_R \in \mathbb{R}^{N \times (N + N_d/2)} \). \( u_L \) and \( u_R \) are then fed into a fully connected layer (as shown in Figure 4C).

The association score \( h_L \in \mathbb{R}^{2 \times 1} \) between drug \( r_i \) and disease \( d_k \) in one branch is obtained as

\[
h_L = \text{softmax} \left( W_L u_L + b_L \right) \quad (10)
\]

where \( W_L \in \mathbb{R}^{2 \times (N + N_d/2)} \) is the weight parameter matrix, and \( b_L \) is a bias vector. \( h_L(1) \) contains the probability that \( r_i \) is...
The final association score $h$ is obtained by a weighted fusion of $h_t$ and $h_d$ as

$$h = \alpha h_t + (1-\alpha) h_d, \text{ s.t. } 0 \leq \alpha \leq 1$$  \hspace{1cm} (12)$$

where $\alpha$ is a regulation parameter to balance the contributions of two branches. Let $\text{loss}_L$ and $\text{loss}_R$ denote the losses of two branches as:

$$\text{loss}_L = \min \|h_t - y\|_f, \text{ loss}_R = \min \|h_d - y\|_f$$  \hspace{1cm} (13)$$

where $y = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$ 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 $\text{loss}$ is obtained by

$$\text{loss} = \min \alpha \|h_t - y\|_f^2 + (1-\alpha) \|h_d - y\|_f^2$$  \hspace{1cm} (14)$$

where the regulation parameter $\alpha$ 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 $R_i$ involves drug-disease association matrix $A$ and intra-disease matrix $D$ to ensure that there is no testing data information in the training dataset, $R_i$ 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 heterogeneous 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

$$\text{TPR} = \frac{TP}{TP+FN}, \text{ FPR} = \frac{FP}{TN+FP}.$$  \hspace{1cm} (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 $\theta$. If the testing sample's score is smaller than $\theta$, it is identified as a negative sample. The values of FPR and TPR are calculated by setting different values of $\theta$. 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

$$\text{Precision} = \frac{TP}{TP+FP}, \text{ Recall} = \frac{TP}{TP+FN}.$$  \hspace{1cm} (16)$$

where $\text{Precision}$ represents the ratio between the number of correctly identified positive samples and all samples which are predicted to be positive samples, and $\text{Recall}$ 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.
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, HeteroDualNet 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, HeteroDualNet 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 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.

| Methods       | ROC AUC | PR AUC |
|---------------|---------|--------|
| HeteroDualNet | 0.908   | 0.154  |
| TL_HGBI       | 0.723   | 0.031  |
| MBiRW         | 0.855   | 0.045  |
| LRSSL         | 0.845   | 0.089  |
| SCMFDD        | 0.611   | 0.006  |
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

| Drug name     | Rank | Disease name                     | Description               | Rank | Disease name                     | Description               |
|---------------|------|----------------------------------|---------------------------|------|----------------------------------|---------------------------|
| ciprofloxacin | 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, DrugBank            | 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.
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 useful 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/TCCB.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/btx628
Cheng, X., Xiao, X., and Chou, K. C. (2017). p-Loc-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
The use, distribution or reproduction in other forums is permitted, provided the original publisher (s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.