Briefings in Bioinformatics, 22(2), 2021, 2141–2150

doi: 10.1093/bib/bbaa044
Advance Access Publication Date: 4 May 2020
Problem Solving Protocol

Identifying drug–target interactions based on graph convolutional network and deep neural network

Tianyi Zhao, Yang Hu, Linda R. Valsdottir, Tianyi Zang and Jiajie Peng

Corresponding authors: Tianyi Zang, Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China. E-mail: tianyi.zang@hit.edu.cn; Jiajie Peng, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: jiajiepeng@nwpu.edu.cn

Abstract

Identification of new drug–target interactions (DTIs) is an important but a time-consuming and costly step in drug discovery. In recent years, to mitigate these drawbacks, researchers have sought to identify DTIs using computational approaches. However, most existing methods construct drug networks and target networks separately, and then predict novel DTIs based on known associations between the drugs and targets without accounting for associations between drug–protein pairs (DPPs). To incorporate the associations between DPPs into DTI modeling, we built a DPP network based on multiple drugs and proteins in which DPPs are the nodes and the associations between DPPs are the edges of the network. We then propose a novel learning-based framework, 'graph convolutional network (GCN)-DTI', for DTI identification. The model first uses a graph convolutional network to learn the features for each DPP. Second, using the feature representation as an input, it uses a deep neural network to predict the final label. The results of our analysis show that the proposed framework outperforms some state-of-the-art approaches by a large margin.

Key words: drug–target interaction prediction; graph convolutional network; deep neural network; biological networks

Introduction

The identification of drug–target interactions (DTI) is an important step in developing new drugs and understanding their side effects [1]. Two experimental methods are widely used to identify DTIs [2]: affinity chromatography [3] and protein microarrays [4]. Due to the increasing number of synthesized compounds developed to target a large number of proteins and disease processes, identifying DTIs using biological experiments is time-consuming and costly [5], and very few true DTIs are found using such methods [6]. Therefore, in recent years, researchers have sought to identify DTIs using computational approaches [7]. The existing computational DTI identification methods can

Tianyi Zhao is a PhD student in Department of Computer Science at Harbin Institute of Technology. He currently works as a bioinformatician in Beth Israel Deaconess Medical Center.
Yang Hu is an associate professor in Department of Life Science at Harbin Institute of Technology. His expertise is bioinformatics.
Linda R. Valsdottir holds an MS in Biology and works as a scientific writer at the Smith Center for Outcomes Research in Cardiology at Beth Israel Deaconess Medical Center in Boston, MA. Her work is focused on helping researchers communicate their findings in an effort to translate novel analytical approaches and clinical expertise into improved outcomes for patients.
Tianyi Zang is a professor with the School of Computer Science and Technology at Harbin Institute of Technology (HIT), China. Before joining HIT in 2009, he was a research fellow at the Department of Computer Science at University of Oxford, UK. His current research is concerned with biomedical bigdata computing and algorithms, deep-learning algorithms for network data, intelligent recommendation algorithms, and modeling and analysis methods for complex systems.
Jiajie Peng is an associate professor in the School of Computer Science at Northwestern Polytechnical University. His expertise is computational biology and machine learning. Availability and implementation: https://github.com/zty2009/GCN-DNN/.
Submitted: 31 October 2019; **Received (in revised form):** 5 March 2020

be classified into three categories: text-mining-based methods, biological-feature-based methods and network-based methods.

Identifying DTIs using text mining methods

Text-mining-based methods identify DTIs by extracting information from the literature and using the descriptions of the drugs and their targets as features [8]. Several methods, such as 'MAM' [9], 'PharmGKB' [10] and 'Chem2bio2rdf' [11] measure the associations between drugs and targets based on semantic similarity. Recently, researchers have begun to take advantage of machine-learning methods to identify DTIs using text-based features. Fu *et al.* [12] proposed a semantic similarity framework using random forest (RF) and support vector machine (SVM) methods. However, the diversity of language expression and conflicting information found in the literature limit the performance of text-mining-based methods [13].

Identifying DTIs using biological feature-based methods

Biological feature-based methods apply machine-learning methods to extracted biological features of drugs and targets to identify DTIs [14–16]. These methods usually include two key components: feature extraction and DTI prediction. 'SimBoost' [17] trains a gradient-boosting machine model on the similarities between drugs and proteins to learn their binding affinities. 'NRLMF' [18] uses the similarities between drugs and proteins to model the probability that a drug will interact with a target by logistic matrix factorization, a type of collaborative filtering method. 'BLM–NII' [19] integrates neighbor-based interaction-profile inference into a bipartite local model (BLM). These methods improve the accuracy of DTI prediction to some extent. However, these methods do not take drug–drug or protein–protein interactions into account [20]. Of course, the relationships between a disease process and a drug are far more complex than the 'one gene, one drug, one disease' paradigm [21].

Identifying DTI using network-based methods

Network-based methods have gained broader attention in recent years [22]. They mainly consist of two steps: network construction and DTI identification. Network-based methods calculate the similarities between drugs and targets based on network topology. Networks including drugs, proteins or both are built to identify novel DTIs [7, 23, 24]. The bipartite graph is the most common network structure in this type of method [25]. Drugs and proteins are the nodes in the network, and edges are known DTIs. The aim of this method is to predict unknown edges based on the known edges. The basic idea is that drugs tend to bind to similar targets and vice versa [22]. Therefore, calculating the similarities between drugs and proteins plays a vital role in this type of method. 'DDR' [26] constructs a drug–drug interaction network and a protein–protein interaction network based on the similarities between drugs and proteins. Then, they use an RF method to infer combinations of drugs and proteins. Several other methods also use this concept to predict DTIs [25, 27, 28]. Network-based methods generally achieve good prediction accuracy and consider the associations between proteins and between drugs [29]. However, these methods do not take associations between drug–protein pairs (DPPs) into account.

Yamanashi *et al.* [30] classified target proteins into four categories: enzymes, ion channels, G protein-coupled receptors and nuclear receptors. Many DTI identification methods have been tested on this dataset, and most achieved high precision in terms of both area under the curve (AUC) and area under the precision recall curve (AUPR) [25, 31]. However, since most existing methods do not consider associations between different DPPs, these methods do not perform well when applied to the Food and Drug Administration (FDA)-approved drugs in DrugBank. 'DrugE-Rank' [32] achieved more than 30% improvement in AUPR compared to previous methods when tested in DrugBank, but the highest AUPR was only 0.2831 among several tests, which means that the false-positive rate was quite high. Recently, Olayan *et al.* [26] developed the 'DDR' method, which achieved an AUPR of 0.63, 0.42 and 0.4 in three respective tests. Although DDR showed great improvements compared with previous methods, it still does not model the associations among different drug–target pairs to reduce the false-positive rate.

Our aims

To address the pitfalls of these previous approaches, we propose GCN-DTI, which combines a graph convolutional network (GCN) [33] and a deep neural network (DNN) [34] to predict DTIs. GCN-DTI transforms the edge prediction problem into a DPP classification problem. Here, a DPP is a combination of any drug and any protein. If the drug and protein in a specific DPP can interact with one another, it is labeled a true DPP and we can call it a 'DTI'.

In our GCN-DTI model, drug networks and protein networks are used to generate a DPP network. In the DPP network, each node is a DPP, and the edges of the DPP network are inferred by the respective drug and protein networks. Therefore, our DPP network contains information on the individual drugs and proteins, drug–drug interactions, protein–protein interactions, drug–protein interactions, and most importantly, associations between DPPs. The GCN can extract the features of each DPP according to the topology of the DPP network. After extracting the features from this large network with a GCN layer, a DNN is used to predict the labels of the DPPs.

The major contributions of this research are as follows:

(i) By integrating multiple types of interactions, we built a DPP network in which the nodes are DPPs and the edges represent the associations between DPPs.
(ii) We employed a GCN-based model to combine drug and protein features with the structural information of the DPP network.
(iii) The results of our evaluation of this model show that GCN-DTI outperforms some state-of-the-art approaches for drug–target interaction prediction.

Methodology

There are three steps in the GCN-DTI method (Figure 1): construction of the DPP network (section Construction of the DPP network), encoding by GCN (section GCN-based feature representation) and classification by DNN (section Classification by DNN).

Construction of the DPP network

Nodes and edges of DPP network

To analyze the relationships between DPPs, we first construct a DPP network based on drug and protein networks, which have

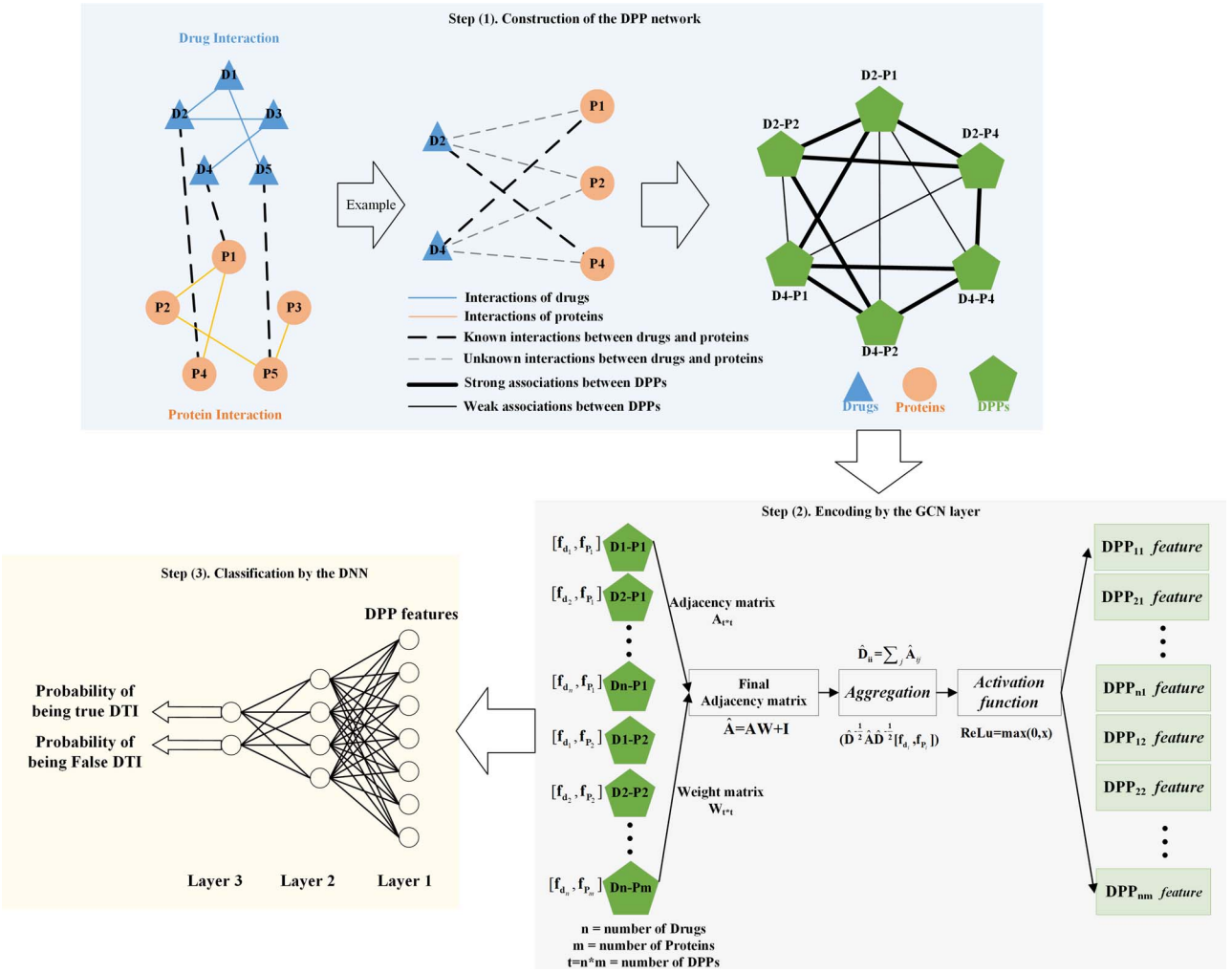

Figure 1. There are three steps in implementing GCN-DTI: construction of the DPP network, encoding by the GCN layer and classification by the DNN. As shown in the left of step 1, drug and protein networks are constructed based on known drug and protein interactions. The blue triangles denote drugs, and brown points denote proteins. Next, black dotted lines denote known interaction and grey dotted lines denote unknown interactions. Two drugs and three proteins are used as an example. The pair made of drug 2 & protein 4 and the pair made of drug 4 & protein 1 can be connected by black dotted lines since they represent known DTIs, while the grey dotted lines connect the unknown DPPs. Finally, in the right part of step 1, the bold lines denote strong connections and normal black lines denote weak connections. Here, six DPPs (including two DTIs: D2-P4 and D4-P1) are selected to show the DPP network. Each node (green pentagon) denotes a DPP and each edge denotes an association between DPPs. Taking D4-P4 as an example, since it shares P4 with D2-P4, the two pairs have a strong connection. Since P4 can interact with P1, D4-P4 can connect weakly with D2-P1. Since D4 cannot interact with D2 and P4 cannot interact with D4, D4-P4 does not have association with D2-P2 (An 'association' is defined in the section Construction of the DPP network.) In step 2, a GCN is used to extract each DPP feature from the DPP network. Each DPP node contains two pieces of information marked $\left[f_{d_i}, f_{p_i}\right]$, where f_{d_i} denotes the drug feature and f_{p_i} denotes the protein feature. If the number of drugs is n and the number of proteins is m, the number of nodes in the DPP network would be $t = m \times n$. The adjacency matrix A can be obtained based on the edges of the network, and the weighted matrix W can be obtained based on three kinds of associations between DPPs. The weighted adjacency matrix $\hat{A}$ can be obtained by $\hat{A} = AW + I$, and the information for each node is integrated with weighted adjacency matrix $\hat{A}$. Finally, the Rectified Linear Unit (ReLU) function is applied as the activation function to obtain the DPP feature. In step 3, three layers have been constructed to map DPP features to their labels. The output of this model is the probability of a DPP being true or false.

corresponding across-network associations based on known interactions between drugs and proteins. These associations, which are obtained from the drug–drug interaction networks and the protein–protein interaction networks, represent the edges of the DPP network. Each DPP contains a drug and a protein, and represents a node of the DPP network. Therefore, the number of nodes in the DPP network is:

$$T = n \times m \tag{1}$$

where T is the number of nodes in the DPP network, n is the number of drugs and m is the number of proteins.

We define associations between DPPs as strong associations, weak associations and non-associations. DPP associations can therefore be inferred as strongly connected and weakly connected as illustrated here:

(i) If two DPPs share a common drug or protein, they are defined as strongly connected.
(ii) If there is an association between the drugs or the proteins in two DPPs, they are defined as weakly connected.
(iii) If the two DPPs do not have a drug or protein in common and their drugs or proteins also cannot interact with one other, they are defined as having a non-association.

If we define any DPP as D_iP_j, the association of DPPs could be represented by adjacency matrix A:

$$A = \begin{vmatrix} f(D_1P_1D_1P_1)f(D_1P_1D_1P_2)\ldots f(D_1P_1D_nP_m) \\ f(D_1P_2D_1P_1)f(D_1P_2D_1P_2)\ldots \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ f(D_1P_mD_1P_1) \quad \ldots \quad f(D_nP_mD_nP_m) \end{vmatrix}_{T\times T} \tag{2}$$

where A represents the adjacency matrix of the DPP network as well as the edges of the DPP network, and $f\left(D_iP_j, D_kP_l\right)$ represents the function used to calculate the associations between DPPs.

The associations between different DPPs can be calculated as follows:

$$f(D_iP_j, D_kP_l) = \begin{cases} 1 & \max(D_iD_k, P_jP_l) = 1 \\ 0.5 & 0 < \max f(D_iD_k, P_jP_l) < 1 \\ 0 & \max(D_iD_k, P_jP_l) = 0 \end{cases} \tag{3}$$

where D_iD_k denotes the interaction between the i_{th} drug and the k_{th} drug, and P_jP_l denotes the interaction between the j_{th} protein and the l_{th} protein.

DPP feature extraction

Next, we extract the biological features of the drugs and targets. Drug features are defined by chemical categories (e.g. adrenal cortex hormones, amides, amines and cardiovascular agents). Protein features are defined by their sequence information and the chemical properties of their amino acids. The features of each DPP are made up of the combined features of its drug and protein molecules.

Protein features. The interaction between a drug and its target is influenced by the hydrophobicity, polarity and tertiary structure of the target protein [20]. Furthermore, the patterns of hydrophobic and hydrophilic residues contribute to a protein's structure. Therefore, the hydrophilicity and hydrophobicity of each amino acid in a protein's sequence will be extracted as a chemical characteristic of that protein. The amino acids are divided into six groups based on their chemical characteristics [35, 36]: strongly hydrophilic or polar acids (R, D, E, N, Q, K, H), strongly hydrophobic acids (L, I, A, V, M, F), weakly hydrophilic or weakly hydrophobic acids (S, T, Y, W), and proline (P), glycine (G), and cysteine (C), which are in their own categories based on their unique characteristics.

In addition, the relative proportion of each amino acid in a sequence is an important consideration when evaluating protein similarity. Therefore, this is also extracted as a feature.

Taken together, each protein contains a 26-dimensional feature consisting of 6 chemical characteristics and the relative proportions of each of the 20 amino acids.

Drug features. Simplified Molecular Input Line Entry Specification (SMILES) is commonly used to extract the chemical structure of drugs [37, 38], since this method can explicitly describe molecular structure in American Standard Code for Information Interchange (ASCII) strings [39]. However, subtle differences in functional groups can lead to significant differences in the chemical properties of drugs, even if their SMILES codes are similar. Therefore, in this paper, the category of a drug is an important basis for judging its similarity to other drugs. There are more than 10 000 categories of drugs in DrugBank, so to avoid the curse of dimensionality, only the most common categories are used to distinguish drugs.

Since any one drug may have dozens of category labels (e.g. a single drug can be an amide, an enzymeinhibitor and a steroid), these categories can accurately reflect many of its characteristics. Compared with its chemical formula and other properties, a drug's category is a feature that is relatively easy to extract, is understandable and can represent the true chemical properties of the drug. Therefore, the categories of each drug are selected as a feature.

Since the features of each DPP are obtained by combining the features of its corresponding drug and target molecules, it is important that the dimensions of the drug and protein features are similar in number. This ensures that the DPP features are not biased towards either source. Since each protein's feature is 26-dimensional, 25 or 27 categories are selected to encode the features of the drugs. Changing the number of drugs included in the DPP network will result in different categories being selected (see Supplementary Section 1).

GCN-based feature representation

For a given DPP network, $G = (V, E)$, $V = \{v_1, v_2, \ldots, v_n\}$ denotes the DPP nodes, $E \subseteq V \times V$ is the set of edges (i.e. associations between DPPs) and $E = \{e_1, e_2, \ldots, e_m\}$. Here, n is the number of DPPs and m is the number of edges.

Adjacency matrix A can be binary or weighted [40]. Since we have defined three types of associations between DPPs, the matrix is weighted in this paper such that a strong association = 1, a weak association = 0.5 and a non-association = 0. Therefore, we define $W \in \mathbb{R}^{n\times n}$ as the weighted matrix encoding the weight of the connection between two vertices (i.e. two DPPs). The weighted matrix W, which can be calculated by Formula (3), considers whether two DPPs are connected, and if so, how strongly they are connected. Finally, the weighted adjacency matrix can be obtained by $A' = A \circ W$ (A' is the Hadamard product of A and W).

Generally, each node in a GCN network should contain its own features [40], so an identity matrix is always added to the adjacency matrix:

$$\hat{A} = A + I \tag{4}$$

where A is the network's adjacency matrix and I is the identity matrix. However, when calculating adjacency matrix A in the DPP network, we already incorporate I, so the weighted adjacency matrix in this paper is $\hat{A} = A'$.

Therefore, the Laplacian matrix should be:

$$L = D - \hat{A} = I_n - D^{-\frac{1}{2}}\hat{A}D^{-\frac{1}{2}} \tag{5}$$

where I_n is the identity matrix and D is the inverse degree matrix (see Supplementary Section 2.1).

Finally, the features of each DPP in the network can be extracted by the GCN using the following formula:

$$X' = \mathrm{ReLu}\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X\right) \tag{6}$$

where X is the feature vector of each node:

$$X = [f_d, f_p] \tag{7}$$

in which f_d denotes the feature of the corresponding drug, and f_p denotes the feature of the corresponding protein. The length of f_p is 26 and the length of f_d is 25 or 27. We concatenate f_d and f_p as the feature vector for each DPP, so as a result, the features of each DPP are the combination of the features of its drug and its protein.

After GCN encoding, each node (i.e. each DPP) contains all the information associated with its corresponding drug and protein, as well as its location in the network.

Classification by DNN

After extracting the features of the DPP network using the GCN, the DNN model is used as a supervised learning model to determine the authenticity of the DPPs.

As shown in Figure 2, the DNN model contains three layers. The number of nodes, activation function and dropout rate of each layer is also given. The input for this model is the DPP feature vector, which was extracted using the GCN. For the first layer, 256 nodes are built using an ReLU function. ReLU activation is chosen because of its computational efficiency, sparsity and reduced likelihood of a vanishing gradient. Since this is a two-class problem, we chose a sigmoid activation function for the final layer.

Binary cross entropy was chosen as the loss function since it is the most suitable for two-classification problems. Its output is relatively easily understood: when y_i and $\widehat{y}_i$ are equal, the loss is 0; otherwise, the loss is a positive number. Moreover, the greater the difference between the two probabilities, the greater is the loss. Finally, 'RMSProp' was chosen as the optimizer (see Supplementary Section 2.2).

Results

We briefly introduce the datasets we used in the section Datasets. Details of the experiments conducted using these datasets are described in the section Experiment setup. Next, section Performance evaluation in the DrugBank dataset shows the results of a comparison between GCN-DTI and six existing methods using the DrugBank dataset. Section Performance evaluation in the Yamanashi dataset describes the comparison between the GCN-DTI method and the DDR method, which performed best among the six existing methods in the Yamanashi dataset. In the section Types of associations between drugs and proteins, we evaluate the features used in our method compared with those used in other methods. Finally, case studies were conducted to show the validity of the results obtained using GCN-DTI method.

Datasets

The HIPPIE database [41] was used to obtain information on protein–protein interactions (PPIs). PPIs with scores greater than 0.5 were selected to build a protein network.

We evaluated our model in data from the Yamanashi [30] and DrugBank 5.0.3 datasets, which were also used for a study by Wishart *et al.* [42]. A total of 1481 known drugs and 1408 known proteins with 9880 DTIs were extracted from the databases. The DrugBank database also contains information about drug interactions for building drug networks.

In addition, 450 new potential drugs and 304 new potential targets were obtained from the DrugBank database. Drugs that can interact with more than 500 known drugs were selected as new drugs. Proteins that can interact with more than 15 known proteins were selected as new proteins. Here, new drugs are those that do not have any known targets to interact with, and new targets are those that do not have any known drugs to interact with. The definitions of new drugs and targets are the same as those used in Olayan *et al.* [26].

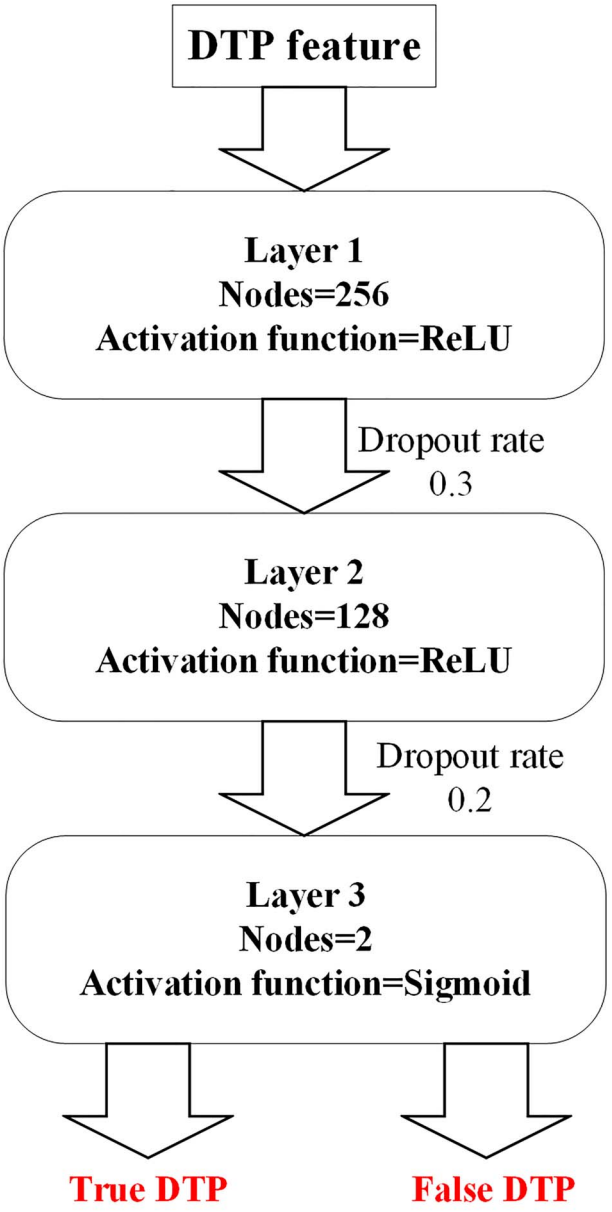

Figure 2. The structure of the proposed DNN model.

Experiment setup

To evaluate the algorithm comprehensively, we tested the performance of our method on three tasks: (1) identifying new drug interactions with known targets, called S_D; (2) identifying new target interactions with known drugs, called S_T and (3) identifying unknown DTIs between known drugs and targets, called S_P (see Supplementary Section 3). In the first case, there were 1931 drugs (450 new) and 1408 proteins, corresponding to 633 600 unknown DPPs and 9880 true DPPs (DTIs). In the second case, there were 1481 drugs and 1712 proteins (304 new), corresponding to 450 224 unknown DPPs and 9880 true DPPs (DTIs). In the third case, there were 1481 drugs and 1408 proteins, corresponding to 2 075 368 unknown DPPs and 9880 true DPPs (DTIs).

Ten-fold cross-validation was used for each of these three tests. Positive and negative sets were divided into ten subsets

each. Then one positive subset and one negative subset were selected as test sets each time, and the remaining data were used as training sets. We repeated each test five times to obtain an average result.

Performance evaluation in the DrugBank dataset

We compared our proposed GCN-DTI method with six existing methods: 'DDR' [26], 'COSINE' [43], 'DNILMF' [27], 'NRLMF' [18], 'KRONRLS-MKL' [25] and 'BLM-NII' [19].

'COSINE,' 'NRLMF,' 'KRONRLS-MKL' and 'BLM-NII' use the similarities between drugs and proteins to infer drug–target interactions using a statistical framework, logistic matrix factorization, improved multiple kernel learning (MKL) and improved bipartite local model (BLM) methods, respectively. Although these methods build connections between drugs and proteins, they do not consider drug and protein networks, which would cause a loss of information. However, 'DDR' and 'DNILMF' construct both a drug network and a protein network, which lead them to extract more chemical and molecular information, and thus derive more information about similarities. 'DDR' uses an RF method to classify DTIs based on different graph-based features, while 'DNILMF' uses logistic matrix factorization. Although these two methods establish drug and protein networks separately, they do not consider associations between different DPPs (see Supplementary Section 4).

Figure 3 shows a comparison of the results obtained using the DrugBank dataset. Note that since 'COSINE' is specifically used to find protein targets for new chemicals, it could only be tested on this task.

As shown in Figure 3, compared with the other methods, GCN-DTI showed a significant improvement in AUPR and also performed well in terms of AUC. Since the AUC resulting from the existing methods was already very good, GCN-DTI only showed a slight improvement in this metric. GCN-DTI performed best at the S_P task in terms of both AUC and AUPR. This demonstrates that establishing connections between different DTIs can effectively improve the ability of the algorithm to distinguish between true and false DTIs. In the S_D and S_T tasks, some drugs or targets were not found in the positive set, so the results were not as good as those for S_P.

Performance evaluation in the Yamanashi dataset

Although previous methods have shown satisfactory outcomes using the Yamanashi dataset, GCN-DTI was implemented in this dataset to show its general applicability. Four S_P tests were performed for enzymes, ion channels, G protein-coupled receptors and nuclear receptors. Ten-fold cross-validation was also used for each test. Detailed information about the experimental setup can be found in the Supplementary Section 5.

Since DDR performed best among the six previous methods in DrugBank, we compared GCN-DTI with DDR in the Yamanashi dataset, and the resulting AUC and AUPR values are shown in Table 1.

Since the improvement of GCN-DTI in the Yamanashi dataset was not very large, we calculated the variance of the results using 10-cross validation repeated five times to show the stability of the results.

In summary, these tests run in two independent datasets—the DrugBank and Yamanashi datasets—showed that GCN-DTI outperformed all other methods tested. These experiments demonstrate the general applicability of GCN-DTI.

Table 1. Comparison of GCN-DTI and DDR in AUC and AUPR

Yamanashi	E	GCN-DTI	**0.98 ± 7.6e-4**	**0.98 ± 4.7e-3**
		DDR	0.97 ± 1.8e-3	0.92 ± 6.3e-3
	IC	GCN-DTI	**0.98 ± 6.8e-4**	**0.92 ± 5.4e-3**
		DDR	0.98 ± 7.7e-4	0.79 ± 8.7e-3
	GPCR	GCN-DTI	**0.97 ± 1.5e-3**	**0.82 ± 7.5e-3**
		DDR	0.96 ± 1.8e-3	0.79 ± 8.3e-3
	NR	GCN-DTI	**0.93 ± 2.4e-3**	**0.85 ± 1.8e-2**
		DDR	0.92 ± 3.3e-3	0.83 ± 2.4e-2

Bold font indicates the best algorithm in this test.
Four S_P tests were performed for enzymes (E), ion channels (IC), G protein-coupled receptors (GPCR) and nuclear receptors (NR), respectively, in Yamanashi dataset.

Types of associations between drugs and proteins

There are many ways a drug can interact with a receptor, and different types of associations will lead to different biochemical reactions. For example, agonists usually have high intrinsic activity and can cause biological effects, while antagonists have low intrinsic activity and do not exert biological effects after binding to the receptor, but can instead prevent an agonist from binding to that receptor. The types of associations between a drug and a protein are keys to determining the effect of a drug. These actions are divided into more than 30 types (e.g. inhibitor and antagonist). Due to the importance of the types of associations between drugs and proteins, we tested the ability of GCN-DTI to categorize types of associations based on the known DTIs in DrugBank.

In this database, some actions are less common than others (see Supplementary Section 6). Therefore, we only tested GCN-DTI on inhibitors (1120), antagonists (1006), potentiators (286) and agonists (614), because the majority of DTIs in the dataset are classified as having those actions.

We considered the identification of association types as a multi-classification problem, so the four types were labeled as 0, 1, 2 or 3. The activation function of the DNN model's last layer was changed for this analysis: the 'Softmax' function was chosen because it is suitable for multi-classification neural network outputs. In addition, the loss function was changed to 'categorical cross-entropy' (see Supplementary Section 6).

Finally, the 3026 DTIs were divided into 10 groups for the 10-fold cross-validation.

As shown in Figure 4, GCN-DTI categorized antagonists best with 94.53% accuracy. The mean accuracy of the classification of the 3026 DTIs was 89.76%, which demonstrated the effectiveness of GCN-DTI in classifying different types of associations.

Case study

Thus far, we have demonstrated the effectiveness of our method using test datasets. Finally, we built models using the GCN-DTI method to mine novel interactions.

All positive DTIs from the DrugBank database were used for training in this model. Since there were a total of 9880 positive samples, we randomly selected 9880 negative samples from S_D, S_P and S_T unknown samples, respectively. GCN-DTI was applied to construct three final models. We found several DTIs, which were not included in the DrugBank database, and compared them with other findings from the literature:

(i) An interaction between acetaminophen and DNA polymerase epsilon catalytic subunit A was not previously

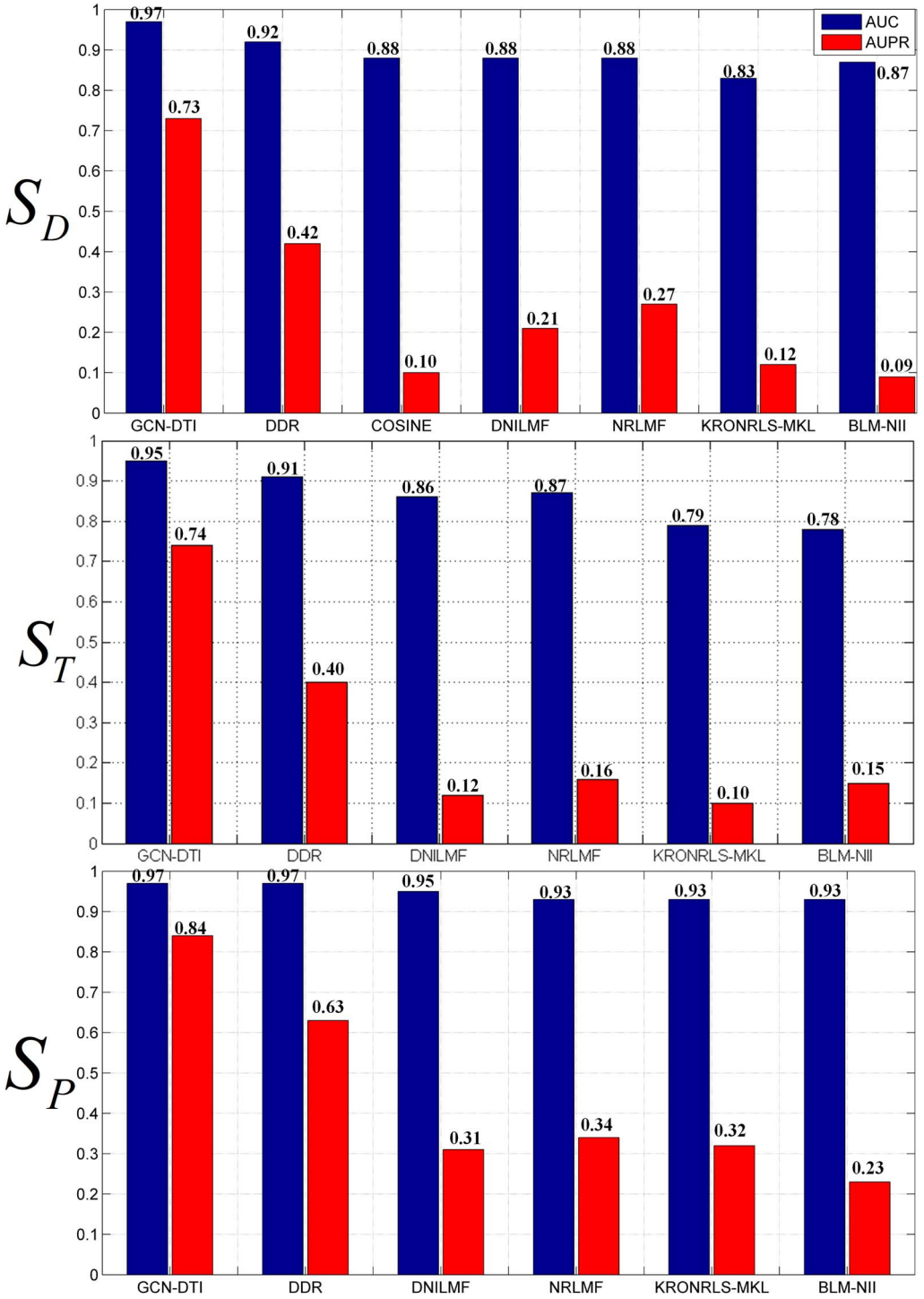

Figure 3. Comparison results of GCN-DTI and previous methods.

reported in DrugBank. Prot *et al.* [44] found that acetaminophen results in increased expression of POLE mRNA using hepatoma cells cultivated inside a microfluidic biochip with or without acetaminophen (APAP). POLE is the coding gene of DNA polymerase epsilon catalytic subunit A.

(ii) Zheng *et al.* [45] and Chandrasekaran *et al.* [46] reported that acetaminophen can increase the expression of myeloperoxidase, but this interaction was not found in DrugBank. Both of those studies investigated the process through which acetaminophen caused liver injury.

(iii) Tesmilifene was found to downregulate the mRNA expression of solute carriers by Walter *et al.* [47]. This is also a novel DTI found using GCN-DTI.

(iv) Riboldi *et al.* [48] found that benzydamine strongly inhibited chemoattractant-induced activation of the mitogen-activated protein kinase (MAPK). The interaction between

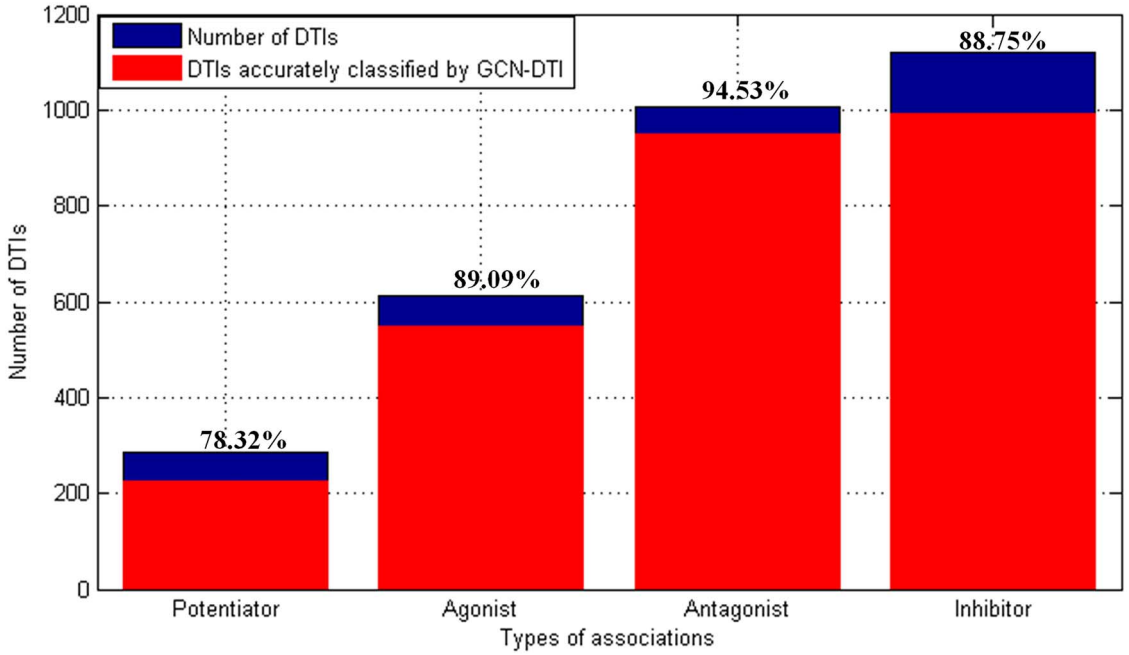

Figure 4. Accuracy of GCN-DTI in identifying types of associations.

benzydamine and MAPK was also identified by GCN-DTI without a positive result in DrugBank.

(v) Beyer *et al.* [49] found that acetaminophen affects the expression of ATP7B mRNA. Furthermore, Jiang *et al.* [50] found that acetaminophen results in a decreased expression of ATP7B mRNA.

These five DTIs are all novel for DrugBank and were identified by GCN-DTI. These case studies show the reliability of the results of GCN-DTI and illustrate its ability to identify interactions between drugs and proteins.

Conclusion

There is a growing body of research that seeks to accurately identify DTIs using computational methods. Although many methods have achieved high precision using the Yamanashi dataset, none of the existing methods have yet achieved satisfactory AUPR for the FDA-approved drugs in DrugBank. We believe that this is because previous methods did not consider associations between DPPs. GCN-DTI was developed to overcome this drawback and was found to achieve improved predictive accuracy, with high AUPR and AUC values.

Most prior research has focused on constructing separate drug and protein networks and predicting the edges connecting the two networks. In contrast, our method constructs a DPP network in which each node contains the information from its corresponding drug and protein sub-networks. The relationships between different DPPs can also be obtained from the corresponding edges of the DPP network. Therefore, our work focuses on distinguishing true and false DPPs from within a very large DPP network. A GCN layer is used to extract the features of each DPP, and then a DNN layer distinguishes between true and false DPP features.

Nearly, a million nodes are included in the network, so the adjacency matrix is too large to process as a whole matrix. Therefore, the encoding process was completed line by line, which places a high demand on computational resources and is time consuming. This is a problem that needs to be addressed in future work.

In conclusion, GCN-DTI substantially improved the accuracy of identification of interactions between drugs and proteins when compared with other methods. Three tests with 10-fold cross-validation were repeated five times to confirm the high AUC and AUPR of GCN-DTI. In addition, five unknown DTIs found using GCN-DTI were supported by the existing literature, which suggests not only the reliability of our results, but also the effectiveness of GCN-DTI in identifying real-world drug–target interactions. The code and results of GCN-DTI are uploaded on Github, which will allow researchers to apply it to other datasets to test DPPs they are interested in.

Conflict of Interest

None declared.

Key Points

- By integrating multiple types of interactions, we built a DPP network in which the nodes are DPPs and the edges represent the associations between DPPs.
- We employed a GCN-based model to combine drug and protein features with the structural information of the DPP network.
- The results of our evaluation of this model show that our method outperforms six state-of-the-art approaches for drug–target interaction prediction.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Natural Science Foundation of China (No. 61702421, U1811262, 61772426); The international Postdoctoral Fellowship Program (no. 20180029); China Postdoctoral Science Foundation (No. 2017M610651); National Key Research and Development Program of China (No. 2016YFC0901605); National Science and Technology Major Project (No. 2016YFC 1202302).

Availability and implementation

https://github.com/zty2009/GCN-DNN/

References

1. Tanoli Z, Alam Z, Ianevski A, *et al*. Interactive visual analysis of drug–target interaction networks using drug target profiler, with applications to precision medicine and drug repurposing. *Brief Bioinform* 2020;**21**:211–20.
2. Xue H, Li J, Xie H, *et al*. Review of drug repositioning approaches and resources. *Int J Biol Sci* 2018;**14**:1232–44.
3. Schirle M, Jenkins JL. Identifying compound efficacy targets in phenotypic drug discovery. *Drug Discov Today* 2016;**21**:82–9.
4. Lee H, Lee JW. Target identification for biologically active small molecules using chemical biology approaches. *Arch Pharm Res* 2016;**39**:1193–201.
5. Mathur A, Loskill P, Shao K, *et al*. Human iPSC-based cardiac microphysiological system for drug screening applications. *Sci Rep* 2015;**5**:8883.
6. Paul SM, Mytelka DS, Dunwiddie CT, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203.
7. Chen X, Yan CC, Zhang X, *et al*. Drug–target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2015;**17**:696–712.
8. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015;**74**:97–106.
9. Zhu S, Okuno Y, Tsujimoto G, *et al*. A probabilistic model for mining implicit 'chemical compound–gene'relations from literature. *Bioinformatics* 2005;**21**:ii245–51.
10. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**:163–5.
11. Chen B, Dong X, Jiao D, *et al*. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 2010;**11**:255–5.
12. Fu G, Ding Y, Seal A, *et al*. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinformatics* 2016;**17**:160.
13. Fotis C, Antoranz A, Hatziavramidis D, *et al*. Network-based technologies for early drug discovery. *Drug Discov Today* 2018;**23**:626–35.
14. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 2015;**20**:318–31.
15. Mayr A, Klambauer G, Unterthiner T, *et al*. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* 2018;**9**:5441–51.
16. Lo Y-C, Rensi SE, Torng W, *et al*. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018;**23**:1538–46.
17. He T, Heidemeyer M, Ban F, *et al*. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J Chem* 2017;**9**:24.
18. Liu Y, Wu M, Miao C, *et al*. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol* 2016;**12**:e1004760.
19. Mei J-P, Kwoh C-K, Yang P, *et al*. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2012;**29**:238–45.
20. Li Z-C, Huang M-H, Zhong W-Q, *et al*. Identification of drug–target interaction from interactome network with 'guilt-by-association'principle and topology features. *Bioinformatics* 2015;**32**:1057–64.
21. Pei J, Yin N, Ma X, *et al*. Systems biology brings new dimensions for structure-based drug design. *J Am Chem Soc* 2014;**136**:11556–65.
22. Luo Y, Zhao X, Zhou J, *et al*. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;**8**:573.
23. Lu Y, Guo Y, Korhonen A. Link prediction in drug-target interactions network using similarity indices. *Bmc Bioinformatics* 2017;**18**:39.
24. Vinayagam A, Gibson TE, Lee H-J, *et al*. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci* 2016;**113**:4976–81.
25. Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics* 2016;**17**:46.
26. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 2017;**34**:1164–73.
27. Hao M, Bryant SH, Wang Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci Rep* 2017;**7**:40376.
28. Zong N, Kim H, Ngo V, *et al*. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* 2017;**33**:2337–44.
29. Buza K, Peška L. Drug–target interaction prediction with bipartite local models and hubness-aware regression. *Neurocomputing* 2017;**260**:284–93.
30. Yamanishi Y, Araki M, Gutteridge A, *et al*. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**:i232–40.
31. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**:e1002503.
32. Yuan Q, Gao J, Wu D, *et al*. DrugE-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 2016;**32**:i18–27.
33. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: *Advances in neural information processing systems*, 2016, 3844–52.
34. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. *IEEE Signal processing magazine* 2012;**29**:82–97.
35. Li F-M, Wang X-Q. Identifying anticancer peptides by using improved hybrid compositions. *Sci Rep* 2016;**6**:33910.

36. Pánek J, Eidhammer I, Aasland R. A new method for identification of protein (sub) families in a set of proteins based on hydropathy distribution in proteins. *Proteins: Structure, Function, and Bioinformatics* 2005;**58**:923–34.
37. Aleksandar MV, Jovana BV, Jelena VZ, *et al*. Application of SMILES notation based optimal descriptors in drug discovery and design. *Curr Top Med Chem* 2015;**15**:1768–79.
38. Öztürk H, Ozkirimli E, Özgür A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC Bioinformatics* 2016;**17**:128.
39. Xu X, Zhong L, Xie M, *et al*. ASCII art synthesis from natural photographs. *IEEE Trans Vis Comput Graph* 2016;**23**:1910–23.
40. Kipf TN, Welling M. *Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907*. 2016.
41. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res* 2016; **45**:gkw985.
42. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2017;**46**:D1074–82.
43. Lim H, Gray P, Xie L, *et al*. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci Rep* 2016;**6**:38860.
44. Prot J-M, Bunescu A, Elena-Herrmann B, *et al*. Predictive toxicology using systemic biology and liver microfluidic "on chip" approaches: application to acetaminophen injury. *Toxicol Appl Pharmacol* 2012;**259**:270–80.
45. Zheng Z, Sheng Y, Lu B, *et al*. The therapeutic detoxification of chlorogenic acid against acetaminophen-induced liver injury by ameliorating hepatic inflammation. *Chem Biol Interact* 2015;**238**:93–101.
46. Chandrasekaran VRM, Periasamy S, Liu L-L, *et al*. 17β-estradiol protects against acetaminophen-overdose-induced acute oxidative hepatic damage and increases the survival rate in mice. *Steroids* 2011;**76**:118–24.
47. Walter FR, Veszelka S, Pásztói M, *et al*. Tesmilifene modifies brain endothelial functions and opens the blood–brain/blood–glioma barrier. *J Neurochem* 2015;**134**: 1040–54.
48. Riboldi E, Frascaroli G, Transidico P, *et al*. Benzydamine inhibits monocyte migration and MAPK activation induced by chemotactic agonists. *Br J Pharmacol* 2003;**140**:377–83.
49. Beyer RP, Fry RC, Lasarev MR, *et al*. Multicenter study of acetaminophen hepatotoxicity reveals the importance of biological endpoints in genomic analyses. *Toxicol Sci* 2007;**99**:326–37.
50. Jiang J, Briedé JJ, Jennen DG, *et al*. Increased mitochondrial ROS formation by acetaminophen in human hepatic cells is associated with gene expression changes suggesting disruption of the mitochondrial electron transport chain. *Toxicol Lett* 2015;**234**:139–50.