Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 584546, 14 pages
http://dx.doi.org/10.1155/2015/584546

Research Article

Prediction of Drug Indications Based on Chemical Interactions and Chemical Similarities

Guohua Huang,[1,2] Yin Lu,[3] Changhong Lu,[4] Mingyue Zheng,[3] and Yu-Dong Cai[1]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Mathematics, Shaoyang University, Shaoyang, Hunan 422000, China*
[3] *State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*
[4] *Department of Mathematics, East China Normal University, Shanghai 200241, China*

Correspondence should be addressed to Mingyue Zheng; myzheng@simm.ac.cn and Yu-Dong Cai; cai_yud@126.com

Received 9 August 2014; Accepted 11 September 2014

Academic Editor: Tao Huang

Discovering potential indications of novel or approved drugs is a key step in drug development. Previous computational approaches could be categorized into disease-centric and drug-centric based on the starting point of the issues or small-scaled application and large-scale application according to the diversity of the datasets. Here, a classifier has been constructed to predict the indications of a drug based on the assumption that interactive/associated drugs or drugs with similar structures are more likely to target the same diseases using a large drug indication dataset. To examine the classifier, it was conducted on a dataset with 1,573 drugs retrieved from Comprehensive Medicinal Chemistry database for five times, evaluated by 5-fold cross-validation, yielding five 1st order prediction accuracies that were all approximately 51.48%. Meanwhile, the model yielded an accuracy rate of 50.00% for the 1st order prediction by independent test on a dataset with 32 other drugs in which drug repositioning has been confirmed. Interestingly, some clinically repurposed drug indications that were not included in the datasets are successfully identified by our method. These results suggest that our method may become a useful tool to associate novel molecules with new indications or alternative indications with existing drugs.

1. Background

The biopharmaceutical industry has a problem: its output has not kept pace with the enormous increases in pharmaceutical R&D spending [1]. After nearly two decades of focusing on developing highly selective ligands, the clinical attrition figures challenge the hypothesis "one gene, one drug, one disease" [2]. In addition, there has been a significant investment by pharmaceutical companies on the optimization of drug discovery pipeline using advanced techniques such as structure-based drug design, combinatorial chemistry, HTS, and genomics. However, the impact of these techniques does not change the predicament [3]. Computational approaches may play significant roles in reducing the developmental costs and shortening the paths to approval, for example, to facilitate drug repositioning.

Drug repositioning is "the process of finding new uses outside the scope of the original medical indications for existing drugs or compounds" [4]. In modern computational biology, there are two general approaches to drug repositioning: discovering new indications for an existing drug (drug-centric) and identifying effective drugs for a disease (disease-centric) [5]. The former hypothesizes that "similar drugs" have the same therapeutic effects and are equally effective for a disease, whereas the latter assumes that "similar diseases" need the same therapies and can thus be treated with the same drugs. Different computational approaches related to the drug repositioning problem have been proposed, ranging from clustering drugs either based on their pharmacophore descriptors [6] or based on connectivity map-based networks [7] to predicting drug-target interactions [8–10] and drug-disease associations [11–15].

On the other hand, drug repositioning by computational approaches can be classified into small-scaled applications which analyze specific classes of drugs or drugs for specific diseases [6, 13, 14] and large-scale applications which analyze

a relatively large number of drugs and diseases [7, 11, 12, 15, 16]. The datasets vary among different research subjects. Generally, the drugs can be derived from Drugbank [11, 12] or KEGG [17] or FDA approved and practiced drug [15]; the drug indications may originate from the Online Mendelian Inheritance in Man (OMIM) database [11], Drugbank therapeutic categories [12], or DRUGEX system [15]. For the methods allowing large-scale indication predictions, transcriptional responses towards drugs were typically utilized to calculate drug-drug similarity, then the connectivity map was constructed for clustering, and the categories of query drugs were determined by the nearest distance to the clustered communities [7]. Similarly, the integration of the chemical, bimolecular, and clinical information was made to design a general framework based on bipartite network projections, and the drug ranking was calculated by kernelized score functions [12]. From the view of disease pairs, a network-based and guilt-by-association method was applied to predict novel drug indication [15]. In addition to network methods, a logistic regression classifier was built from the classification features originating from drug-drug similarity and disease-disease similarity [11].

In this study, we presented an approach for large-scale identification of drug indications based on a large drug-indication library and the information of chemical interactions in STITCH [18] and chemical similarities in structure. For a given drug, a K-Nearest Neighbor (KNN) ranking strategy was used to predict the indications according to its interactive drugs or similar drugs, based on the assumption that interactive chemicals or similar chemicals in structure are more likely to share similar biological functions [16, 19, 20]. An important merit of the method is that, given a query drug, it can provide a series of candidate indications, ranging from the most likely one to the least likely one. Obviously, the quality and the size of the datasets play a significant role in the predictive ability of a model. We constructed the benchmark dataset from a commercial database, Comprehensive Medicinal Chemistry (CMC) database of Accelrys company [21] that is derived from the Drug Compendium in Pergamon's Comprehensive Medicinal Chemistry, which contains 1,573 drug compounds and 56 indications. The size of dataset in our method is larger than those investigated in most of previous approaches [7, 11, 12]. The performance of the method on this dataset suggests that it can identify the potential disease indications of a query drug.

2. Methods

2.1. Materials

2.1.1. Dataset. Altogether, 1,944 drug compounds and their indications were retrieved from CMC database. By collecting indications of these drugs, 231 indications recorded in CMC database were obtained. Accordingly, 231 categories were used to label these 1,944 drugs. To yield statistically meaningful result, the categories containing less than 8 drug compounds were disregarded, 1,733 drugs were obtained, and then indications were refined to avoid any inclusion relation between two indications by manual adjustment of the medical terminology mainly based on ATC classification system (http://www.whocc.no/atc_ddd_index/), thereby obtaining 56 categories of indications. For formulation, let DS_1 denote a dataset consisting of these drugs, and the codes of these drugs and their indications were available in Supplementary Material I (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/584546).

In addition, since some drugs whose structures are very similar may be derived from the same drug, these drugs can be easily correctly predicted by any proper method. To strictly examine the proposed method, these similar drugs should be excluded. For this purpose, a graph was constructed, where nodes represented drugs and two nodes were adjacent if and only if the similarity score of the corresponding drugs based on fingerprint ECFP_4 was at least 0.7 (the reason to select ECFP_4 is explained in Section 3.1). A maximal independent set of 1,573 nodes was extracted from this graph and the corresponding 1,573 drugs in this independent set comprised the dataset DS_2. These 1,573 drugs were also classified into 56 categories and the similarity score of any two drugs was less than 0.7. Shown in column 3 of Supplementary Material II is the number of drug compounds in each category for dataset DS_2. For convenience, we used tags $D_1, D_2, \ldots, D_{56}$ to represent 56 kinds of indication, where D_1 represented "Antihypertensive," D_2 "Uterine stimulant," and so forth (see columns 1 and 2 of Supplementary Material II for details). Accordingly, the dataset DS_2 can be formulated as follows:

$$DS_1 = S_1 \cup S_2 \cup \cdots \cup S_{56}, \quad (1)$$

where S_i is a subset of DS_2 containing drugs labeled by indication D_i. The detailed codes of drug compounds in each S_i are available in Supplementary Material III.

It is observed from the last row of Supplementary Material II that the sum of the number of drug compounds in each category is 2,005, which is much larger than 1,573 that is the total number of individual drug compounds investigated in this study, indicating that some drug compounds possess more than one indication; that is, they are present in more than one category. Of the 1,573 drug samples, 1,209 drugs have only one kind of indication, 313 drugs have two kinds of indications, while the rest possess more than two kinds of indications. Figure 1 shows the relationship between the number of drugs and the number of their corresponding indications. Like the cases of dealing with multilabel classification problems such as predicting multiple attributes of protein or compounds [16, 22, 23], the proposed method would provide the prediction results by ranking the candidate indications from the most likely one to the least one.

In addition, to evaluate the generalization of the proposed method, we employed an independent validation test dataset, denoted by DS_{te}, consisting of 32 drug compounds that were gathered from the recently published literature [1, 24, 25]. The drugs in the test dataset meet the following two criteria: (1) involving drug repositioning that has been experimentally confirmed; (2) being not included in DS_1. These 32 drug compounds and their original indication and reported indication are listed in Table 1.

Table 1: Detailed information of samples in DS_{te}.

Name	ID	Original indication	Reported indication
Statins	CID000446156	Myocardial infarction	Prostate cancer, leukemia
Metformin	CID000004091	Diabetes mellitus	Breast cancer, adenocarcinoma, prostate, colorectal cancer
Rapamycin	CID005284616	Immunosuppressant	Colorectal cancer, lymphoma, leukemia
Methotrexate	CID000126941	Acute leukemia	Osteosarcoma, breast cancer, Hodgkin lymphoma
Zoledronic acid	CID000068740	Antibone resorption	Multiple myeloma, prostate cancer, breast cancer
Wortmannin	CID000312145	Antifungal	Leukemia
Thiocolchicoside	CID000072067	Muscle relaxant	Leukemia, multiple myeloma
Noscapine	CID000275196	Antitussive, antimalarial, analgesic	Multiple cancer types
Galantamine	CID000009651	Polio, paralysis, anaesthesia	Alzheimer's disease
Ropinirole	CID000005095	Hypertension	Parkinson's disease, idiopathic restless leg syndrome
Tofisopam	CID000005502	Anxiety-related conditions	Irritable bowel syndrome
Finasteride	CID000057363	Benign prostatic hyperplasia	Hair loss
Mifepristone	CID000055245	Pregnancy termination	Psychotic major depression
Minoxidil	CID000004201	Hypertension	Hair loss
Paclitaxel	CID000036314	Cancer	Restenosis
Phentolamine	CID000005775	Hypertension	Impaired night vision
Sildenafil	CID000005212	Angina	Male erectile dysfunction
Tadalafil	CID000110635	Cardiovascular disease, inflammation	Male erectile dysfunction
Topiramate	CID005284627	Epilepsy	Obesity
Zidovudine	CID000035370	Cancer	HIV/AIDS
Allopurinol	CID000002094	Tumor lysis syndrome	Gout
Amphotericin	CID005280965	Fungal infections	Leishmaniasis
Colchicine	CID000006167	Gout	Recurrent pericarditis
Retinoic acid	CID000444795	Acne	Acute prophylaxis
Bimatoprost	CID005311027	Glaucoma	Promoting eyelash growth
Ceftriaxone	CID005479530	Bacterial infections	Amyotrophic lateral sclerosis
Colesevelam	CID000160051	Hyperlipidemia	Type 2 diabetes mellitus
Disulfiram	CID000003117	Alcoholism	Melanoma
Naproxen	CID000156391	Inflammation, pain	Anti-Alzheimer's disease
Minocycline	CID054675783	Acne	Ovarian cancer, glioma
Dapoxetine	CID000071353	Analgesia, depression	Premature ejaculation
Bromocriptine	CID000031101	Parkinson's disease	Diabetes mellitus

2.1.2. Chemical Interactions. Some recent studies indicate that interactive compounds are more likely to share common functions than noninteractive ones [16, 26]. The functions of a drug compound can in part determine which diseases it can treat. In view of this, it may be feasible to utilize the information of interactive compounds to predict diseases that a query drug can treat. The information of interactive compounds was downloaded from STITCH (chemical_chemical.links.detailed.v3.1.tsv.gz, http://stitch.embl.de/) [18], a well-known database containing the interaction information of chemicals and proteins. In detail, chemicals are associated with other chemicals and proteins by evidence derived from experiments, databases, and the literature (http://stitch.embl.de/) in STITCH. In the obtained file, each interaction contains two compounds and five scores that indicate the likelihood of the interaction in five different ways. In detail, the score titled "Similarity" was the Tanimoto 2D chemical similarity score [27, 28] calculated by the open-source Chemistry Development Kit [29]; the score titled "Experimental" was obtained by chemical's activities from MeSH pharmacological actions and NCI60 screens; the score titled "Database" was obtained according to chemical reactions contained in pathway databases; the score titled "Textmining" was obtained based on a cooccurrence scheme and a natural language processing (NLP) approach [30, 31]; while the score titled "Combined_score" integrates all

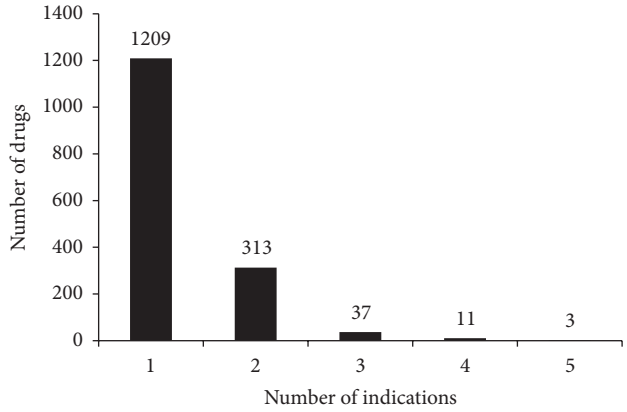

Figure 1: A plot of the number of drugs in DS_2 versus the number of indications.

the aforementioned items. For detailed description, readers can refer to Kuhn et al.'s paper [18]. Accordingly, "Combined_score" was used to quantify the interactivity of two compounds: two compounds with the "Combined_score" greater than zero are deemed as interactive compounds. Furthermore, each interaction is labeled by this score, also termed as confidence score in this study, to indicate the likelihood of its occurrence; that is, an interaction with higher confidence score means that the corresponding compounds can interact with each other with higher probability. For two drug compounds d_1 and d_2, the confidence score of the interaction between them is denoted by $w^i(d_1, d_2)$. In particular, if the interaction between two compounds is not reported in STITCH, its confidence score was set to zero.

2.1.3. Chemical Representation and Similarities. The similarity of two compounds in structure is a classic measurement of the relationship of two compounds. Many representation systems have been established to represent compounds. One of the most well-known systems is SMILES (Simplified Molecular Input Line Entry System) [32], a line notation for representing molecules and reactions using ASCII strings. In this study, we also used this system to represent each drug compound. Furthermore, several fingerprints have been established to calculate the similarity of two chemicals based on their SMILES strings up to now [33–35]. Since different fingerprints may induce different similarity scores of two given chemicals, thereby providing different results [36] for some problems of classification and prediction, we tried fingerprints FP2 [33], MACCS [34], ECFP (ECFP_2, ECFP_4, ECFP_6) [35], and FCFP (FCFP_2, FCFP_4, FCFP_6) [35] in this study to calculate the similarity score of chemicals and attempted to select the best one for the prediction of drug indications. For two drug compounds d_1 and d_2, the similarity scores based on different fingerprints, calculated by Open Babel [33] or RDKit [37], were all denoted by $w^s(d_1, d_2)$, where superscript s indicated which type of fingerprint was used to calculate similarity scores.

2.2. Prediction Method. It has been confirmed that interactive compounds are more likely to share similar functions than non-interactive ones [16, 23]. On the other hand, it is known that compounds with similar structures often share common functions [20]. Because drug indications can be viewed as drug functions, it is appropriate to use known drug indications to predict drugs with unknown indications.

Supposing that there are n drugs in the training set S', say $d_1, d_2, \ldots, d_n$, we need to predict the indications of a query drug d_q based on chemical interactions and chemical similarities as follows.

2.2.1. Prediction Based on Chemical Interactions. As described above, interactive compounds often share similar functions [16, 23], thereby having similar indications with higher probability. For a query drug compound d_q and indication D_j, the score that d_q possesses D_j was determined by the k drug compounds with tag D_j in the training set S', say $d_{i_1}, d_{i_2}, \ldots, d_{i_k}$, such that the confidence scores of the interactions between them and d_q are the first k maximum scores, and was calculated by

$$R^i\left(d_q \Longrightarrow D_j\right) = \sum_{l=1}^{k} w^i\left(d_q, d_{i_l}\right), \quad j = 1, 2, \ldots, 56, \tag{2}$$

where k is a predefined positive integer. It is necessary to point out that (2) is identical to the method in Chen et al.'s study [16] (refer to (6) in Chen et al.'s study [16]) when $k = 1$, while it is same as the method in [38] (refer to (3) in Chen et al.'s study [38]) when k is set to n, where n is the size of the training set.

Obviously, the larger the score $R^i(d_q \Rightarrow D_j)$ is, the more likely that the query drug d_q can treat disease D_j. When $R^i(d_q \Rightarrow D_j) = 0$ for some j, it means that the likelihood that the query drug having the indication D_j is zero. Because it is a multilabel classification problem where a drug may possess more than one indication, our method provided a series of candidate indications for any query drug, ranging from the most likely one to the least likely one. For example, if the results of (2) were

$$R^i\left(d_q \Longrightarrow D_2\right) \geq R^i\left(d_q \Longrightarrow D_6\right) \geq R^i\left(d_q \Longrightarrow D_{46}\right) \cdots \geq R^i\left(d_q \Longrightarrow D_{23}\right) > 0, \tag{3}$$

it can be inferred that the most likely indication of the query drug is D_2, followed by D_6, D_{46}, and so forth. Furthermore, D_2 is called the 1st order prediction, D_6 the 2nd order prediction, and so forth.

Note that the outcomes of (2) might be trivial as follows:

$$R^i\left(d_q \Longrightarrow D_j\right) = 0 \quad \forall j = 1, 2, \ldots, 56. \tag{4}$$

Under such circumstance, there were no interactive compounds of d_q in the training set and no meaningful result can be obtained by this method. We then use the following method based on chemical similarities in structures for further prediction.

2.2.2. Prediction Based on Chemical Similarities. Likewise, because compounds with similar structures often share common functions [20], chemical similarities were applied to predict drug indications if chemical interactions give no meaningful result. For a query drug d_q and indication D_j, k drug compounds with tag D_j in the training set S', still say $d_{i_1}, d_{i_2}, \ldots, d_{i_k}$, were selected such that the similarity scores between these drug compounds and d_q are the first k maximum scores. Now, we calculated the score that d_q can treat indication D_j as follows:

$$R^s\left(d_q \Longrightarrow D_j\right) = \sum_{l=1}^{k} w^s\left(d_q, d_{i_l}\right), \quad j = 1, 2, \ldots, 56, \tag{5}$$

where $w^s(d_q, d_{i_l})$ was the chemical similarity of d_q and d_{i_l} which may be based on FP2, MACCS, ECFP (ECFP_2, ECFP_4, ECFP_6), or FCFP (FCFP_2, FCFP_4, FCFP_6). The rest procedures were same as those of the method based on chemical interactions. Also, given a query drug, the method will provide a series of candidate indications.

2.2.3. Prediction by Integrating Chemical Interactions and Similarities. By integrating chemical interactions and chemical similarities, the indications of a given drug compound d_q were predicted as follows:

(i) the method based on chemical interactions (cf. (2)) was first applied to predict the indications;

(ii) if the outcomes of (2) are trivial as indicated by (4), the method based on chemical similarities (cf. (5)) was then used to make further prediction.

2.3. Cross-Validation and Accuracy Measurement

2.3.1. Cross-Validation Method. In statistical prediction, subsampling test, jackknife test, and independent test are often used to examine the performance of the constructed classifiers [39]. Among these three methods, jackknife test is deemed to be the least arbitrary and can always provide a unique result for a given dataset and a given prediction model because both the training samples and the test samples are fixed [16]. Therefore, it has been widely used by investigators to evaluate the performance of their classifiers [16, 38, 40–49]. Accordingly, it was also used in this study to optimize parameters in methods based on chemical interactions and chemical similarities and compare the performance of different methods.

Subsampling test [50], also named k-fold cross-validation, is another widely used cross-validation method. In this method, the dataset is equally and randomly divided into k parts. Samples in each part are used as testing samples in turn and samples in the rest $k-1$ parts train the prediction method. Thus, each sample is tested exactly once. Compared to jackknife test, k-fold cross-validation costs less computing time and provides similar predicted results. It has also been used in many studies [19, 51–55]. Accordingly, it was used here to examine the proposed method where k was set to 5, that is, 5-fold cross-validation. In addition, we also used independent test to evaluate the proposed method because an independent validation test dataset DS_{te} was constructed as mentioned in Section 2.1.1.

2.3.2. Accuracy Measurement. As described in Section 2.2, the query drug was assigned a series of candidate indications, ranging from the most likely one to the least one. To evaluate the correctness of the candidate indication, the ith order prediction accuracy was calculated by

$$\mathrm{ACC}_i = \frac{\mathrm{PD}_i}{N}, \quad i = 1, 2, \ldots, 56, \tag{6}$$

where N denoted the total number of samples, while PD_i denoted the number of samples whose ith order prediction is correct. For example, when $i = 1$, that is, the 1st order prediction accuracy, the 1st order prediction of each investigated sample was collected and PD_1 was the number of these predictions which were correct, thereby obtaining the 1st order prediction accuracy according to (6). It is obvious that ACC_i is the ratio of correct ith order predicted samples to all samples. If a prediction method yields high ACC_i with small i and low ACC_i with large i, it is deemed as an effective prediction. Since it is difficult to infer the number of indications for certain drug, investigators always pay more attention to the 1st order prediction than others. On the other hand, the 1st order prediction of certain drug indicated its most likely indication. In view of this, the first order prediction accuracy is the most important indicator of the performance of the method.

On the other hand, in pattern recognition and information retrieval, recall and precision are often used to evaluate the performance of the method. For multilabel classification problem, recall and precision of the first t order predictions can be calculated by the following formulae:

$$\begin{aligned} \mathrm{Recall}_t &= \frac{1}{N}\sum_{j=1}^{N}\frac{P_j^t}{N^j}, \\ \mathrm{Precision}_t &= \frac{1}{N}\sum_{j=1}^{N}\frac{P_j^t}{t}, \end{aligned} \tag{7}$$

where N^j represented the number of known indications of the jth sample in the dataset and P_j^t represented the number of correct predictions of the jth sample in the dataset among its first t order predictions. Obviously, $\mathrm{ACC}_1 = \mathrm{Precision}_1$. Since different drug compounds have different numbers of known indications, we set the parameter t in (7) to the smallest integer that is no less than the average number of known indications in the dataset, which can be computed by

$$\mathrm{Average} = \frac{\sum_{j=1}^{N} N^j}{N}; \tag{8}$$

that is, $t = \lceil \mathrm{Average} \rceil$. Obviously, larger Recall_t and $\mathrm{Precision}_t$ imply better prediction performance of the method.

Table 2: Best performance of the method based on chemical similarities for different types of fingerprint and values of k.

Type of fingerprint	Highest 1st order prediction accuracy (%)	k
ECFP_2	48.70	3
ECFP_4	49.39	2
ECFP_6	49.11	5
FCFP_2	42.87	2,3
FCFP_4	48.07	3
FCFP_6	48.99	3
FP2	43.91	3
MACCS	43.39	2,3

3. Results and Discussion

3.1. Optimization of the Methods Based on Chemical Similarities and Chemical Interactions. As mentioned in Section 2.1.3, eight types of fingerprints, including ECFP (ECFP_2, ECFP_4, ECFP_6), FCFP (FCFP_2, FCFP_4, FCFP_6), FP2, and MACCS, were used to calculate the similarity score of two chemicals. To build a more effective prediction method, it is necessary to compare the performance of the method based on chemical similarities on DS_1, where chemical similarities were calculated based on different types of fingerprints and k was set to $1, 2, \ldots, 15, 1732$. The performance of these methods evaluated by jackknife test was available as Supplementary Material IV. It can be observed that when the similarity scores were based on same type of fingerprint, the 1st order prediction accuracies followed an increasing trend before reaching the highest accuracy and then followed a descending trend. Table 2 lists the highest 1st order prediction accuracies for different types of fingerprint and the values of k with which these accuracies can be obtained. It is easy to see that using ECFP_4 and setting $k = 2$ provided the highest 1st order prediction accuracy. Thus, we used this type of fingerprint and set $k = 2$ to build the method based on chemical similarities. In addition, since the proposed method integrated the method based on chemical similarities, the similar drug compounds under fingerprint ECFP_4 should be excluded in order to strictly examine our method. In view of this, the similarity scores based on fingerprint ECFP_4 were used to refine the dataset DS_1 by setting the threshold 0.7, thereby obtaining the dataset DS_2.

In the dataset DS_2, there were 896 drug compounds that have the information of chemical interactions. These drugs comprised the dataset $DS^{(i)}$. The classification model based on chemical interactions (cf. (2)) was conducted on $DS^{(i)}$. To select an optimal parameter k, it was evaluated by jackknife test and k was set to $1, 2, \ldots, 15, 895$. The prediction accuracies thus obtained are available in Supplementary Material V, from which we can observe that the 1st order prediction accuracies followed an increasing trend with the increasing of k when $k < 5$, while the accuracies descended with the increase of k when $k > 5$ (see Table 3 for details). Since the parameter k means the number of interactions that were used to calculate the score that the query drug possesses a certain indication, the score cannot reflect the true likelihood that the query drug has an indication when k is small, while with the increase of k, more and more interactions with low confidence scores are added, which may be noises to the prediction, thereby influencing the predicted results. The highest 1st order prediction accuracy of 58.48% was obtained when k was set to 5. Thus, we set $k = 5$ for the method based on chemical interactions.

Table 3: The 1st order prediction accuracies with different k obtained by the method based on chemical interactions on $DS^{(i)}$ evaluated by jackknife test.

Value of k	The 1st order prediction accuracy
1	47.77%
2	55.92%
3	57.59%
4	58.26%
5	58.48%
6	58.37%
7	58.15%
8	58.04%
9	58.04%
10	58.04%
11	57.81%
12	57.81%
13	57.70%
14	57.70%
15	57.70%
895	57.59%

3.2. Performance of the Proposed Method on DS_2. For clarity, the dataset DS_2 is separated into two subsets, $DS^{(i)}$ and $DS^{(s)}$, where $DS^{(i)}$ consisted of 896 drug compounds that have the information of chemical interactions, while $DS^{(s)}$ contained the rest 677 drug compounds that have no such information. Then the method based on chemical interactions with $k = 5$ was applied to process $DS^{(i)}$, while the method based on chemical similarities with fingerprint ECFP_4 and $k = 2$ was used to process $DS^{(s)}$. The predicted results thus obtained are given as follows.

3.2.1. Performance of the Method Based on Chemical Interactions on $DS^{(i)}$. Using the 896 drugs in $DS^{(i)}$, the classification model based on chemical interactions (cf. (2)) with $k = 5$ was constructed and evaluated by 5-fold cross-validation. To widely examine the method, it was executed five times on $DS^{(i)}$. The predicted results thus obtained are available in Supplementary Material VI. Table 4 lists the first 20 prediction accuracies for each time. It can be seen that the 1st order prediction accuracies were between 55% and 58% and the mean value of these accuracies was 57.00%. For each time, the prediction accuracies generally followed a descending trend with the increase of the order number, indicating that the candidate indications of the samples in $DS^{(i)}$ were sorted

Table 4: The first 20 prediction accuracies obtained by the method based on chemical interactions on $DS^{(i)}$ evaluated by 5-fold cross-validation for 5 times.

Order	First time (%)	Second time (%)	Third time (%)	Fourth time (%)	Fifth time (%)	Mean (%)	Standard deviation (%)
1	56.37	55.95	57.31	57.47	57.92	57.00	0.82
2	21.98	24.01	22.03	22.17	22.35	22.51	0.85
3	8.91	7.25	8.90	6.90	6.84	7.76	1.06
4	5.98	5.32	4.22	5.77	5.25	5.31	0.68
5	3.16	4.19	4.11	4.41	4.56	4.09	0.55
6	2.59	2.49	2.40	2.04	1.94	2.29	0.29
7	1.47	2.38	2.51	2.49	2.51	2.27	0.45
8	1.69	1.47	1.26	1.36	1.60	1.47	0.18
9	2.37	1.13	1.48	1.36	1.25	1.52	0.49
10	0.68	1.02	1.48	1.02	0.91	1.02	0.29
11	1.24	1.25	0.80	1.24	0.91	1.09	0.22
12	1.01	1.02	1.37	1.13	1.37	1.18	0.18
13	1.35	1.25	1.03	1.24	1.14	1.20	0.12
14	0.90	0.45	0.57	0.68	0.57	0.63	0.17
15	0.56	0.57	0.91	0.79	0.68	0.70	0.15
16	0.68	0.79	0.46	0.23	0.57	0.54	0.22
17	1.13	0.79	0.68	1.24	0.46	0.86	0.32
18	0.90	0.79	0.23	1.13	0.57	0.72	0.34
19	1.13	0.57	0.91	1.02	0.68	0.86	0.23
20	0.56	1.36	1.26	0.68	1.14	1.00	0.36

Table 5: The Recalls and Precisions of the first two predictions obtained by three methods on $DS^{(i)}$, $DS^{(s)}$, and DS_2, respectively.

	$DS^{(i)}$		$DS^{(s)}$		DS_2	
Order of time	Recall ($t = 2$) (%)	Precision ($t = 2$) (%)	Recall ($t = 2$) (%)	Precision ($t = 2$) (%)	Recall ($t = 2$) (%)	Precision ($t = 2$) (%)
1st	61.55	39.18	48.95	28.79	56.06	34.65
2nd	62.37	39.98	47.81	28.26	55.99	34.84
3rd	62.42	39.67	47.24	27.76	55.69	34.39
4th	62.45	39.82	49.68	29.32	56.86	35.22
5th	62.68	40.14	49.39	29.09	56.80	35.25
Mean	62.29	39.76	48.62	28.65	56.28	34.87

quite well. In addition, the standard deviations of the five prediction accuracies with the same order were almost lower than 1%, indicating that this method was quite stable on $DS^{(i)}$. The average number of indications that samples in $DS^{(i)}$ can treat was 1.31; that is, Average = 1.31. Thus, the first two predictions of each sample in $DS^{(i)}$ were considered. After calculating (7) with $t = 2$, we obtained 5 Recalls and 5 Precisions, listed in columns 2 and 3 of Table 5. The mean values of Recalls and Precisions were 62.29% and 39.76%, suggesting that the method based on chemical interactions is quite effective to the prediction of drug indications.

3.2.2. Performance of the Method Based on Chemical Similarities on $DS^{(s)}$. For the 677 drugs in $DS^{(s)}$ that have no information of chemical interactions, the method based on chemical similarities (cf. (5)) with fingerprint ECFP_4 and $k = 2$ was used to make prediction and evaluated by 5-fold cross-validation. Also, this method was executed 5 times. The predicted results thus obtained are also available in Supplementary Material VI (the first 20 prediction accuracies for each time are listed in Table 6), from which we can see that five 1st order prediction accuracies were between 43% and 46%. The mean value of these accuracies was 44.45%. Similarly, the prediction accuracies always followed a descending trend with the increase of prediction order for each time, indicating that the method based on chemical similarities also arranged the candidate indications of the samples in $DS^{(s)}$ quite well. It can also be observed from Supplementary Material VI that the standard deviations of the five prediction accuracies with the same order were all lower than 1%, indicating that this method was quite stable on $DS^{(s)}$. The average number of indications that drugs in $DS^{(s)}$ can treat was 1.22. Thus, we still considered the first two predictions for each sample in $DS^{(s)}$ which produced

Table 6: The first 20 prediction accuracies obtained by the method based on chemical similarities on $DS^{(s)}$ evaluated by 5-fold cross-validation for 5 times.

Order	First time (%)	Second time (%)	Third time (%)	Fourth time (%)	Fifth time (%)	Mean (%)	Standard deviation (%)
1	44.17	43.62	43.90	45.86	44.68	44.45	0.88
2	13.41	12.90	11.62	12.77	13.51	12.84	0.75
3	6.85	5.94	8.18	6.39	5.89	6.65	0.94
4	5.54	6.67	4.73	5.22	6.90	5.81	0.93
5	4.52	4.06	5.45	3.92	4.45	4.48	0.60
6	2.19	3.33	3.87	4.64	3.02	3.41	0.92
7	3.64	3.33	2.30	2.90	2.73	2.98	0.53
8	1.90	1.74	3.16	1.31	3.16	2.25	0.86
9	2.77	2.75	2.30	2.32	1.58	2.34	0.48
10	2.48	3.48	1.43	1.74	1.44	2.11	0.87
11	2.04	1.45	1.72	1.16	2.44	1.76	0.50
12	2.19	2.61	2.01	2.76	1.44	2.20	0.52
13	2.33	1.74	2.58	2.03	1.29	2.00	0.50
14	2.19	2.17	2.73	1.31	2.30	2.14	0.52
15	0.44	1.88	1.15	1.60	1.58	1.33	0.56
16	1.02	1.16	0.72	1.16	1.15	1.04	0.19
17	0.87	0.87	1.29	1.02	1.15	1.04	0.18
18	0.87	1.16	1.29	1.02	0.72	1.01	0.23
19	1.60	1.01	1.87	1.31	1.01	1.36	0.37
20	1.60	0.72	0.72	1.02	1.29	1.07	0.38

5 Recalls and 5 Precisions by (7) with $t = 2$. These values are listed in columns 4 and 5 of Table 5, from which we can observe that the mean values of Recalls and Precisions were 48.62% and 28.65%, respectively. These results indicate that the method based on chemical similarities is also effective in the prediction of drug indications.

3.2.3. Performance of the Integrated Method on DS_2. The integrated method combined the predicted results mentioned in Sections 3.2.1 and 3.2.2. The predicted results for each of 5 times were also available in Supplementary Material VI, while Table 7 lists the first 20 prediction accuracies obtained by the integrated method for each time. It can be seen that the five 1st order prediction accuracies were between 50% and 53% and the mean value of these accuracies was 51.48%. Furthermore, the standard deviations of the five prediction accuracies with the same order were all lower than 1%, suggesting that the integrated method was quite stable on DS_2. The average number of indications of samples in DS_2 was 1.27 (2,005/1,573), meaning that the average correct rate would be 1.27/56 = 2.27% if one predicts them by random guess. It is much lower than the five 1st order prediction accuracies obtained by the integrated method. In view of the average number, we consider the first two predictions for each sample in DS_2. The outcomes of (7) with $t = 2$ yield 5 Recalls and 5 Precisions, which are listed in columns 6 and 7 of Table 5. The mean value of Recall and Precision was 56.28% and 34.87%, respectively.

In addition, to sufficiently indicate the effectiveness of the integrated method, we collected the first two predictions for each sample in DS_2 and calculated the prediction accuracy for each category D_i, which was computed by

$$\mathrm{SN}^i = \frac{\mathrm{TP}^i}{C^i}, \quad i = 1, 2, \ldots, 56, \tag{9}$$

where C^i denoted the number of drug compounds labeled by D_i, that is, $C^i = |S_i|$, and TP^i denoted the number of drug compounds whose 1st order prediction or 2nd order prediction was D_i. These accuracies were listed in Supplementary Material VII. It can be seen that the mean values of accuracies of 12 categories were higher than 60%, where 2 of them (D_{11}, D_{56}) were higher than 80%. It is known that the category of large size can easily receive high prediction accuracy, while the category of small size can easily receive low prediction accuracy. However, this case should be avoided for an effective prediction method. To evaluate our method in this aspect, that is, investigating the linear correlation between the prediction accuracy of each category and the size of each category, we employed Pearson product-moment correlation coefficient which is a widely used measure of the linear correlation between two variables and can be computed by

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}, \tag{10}$$

where $\overline{x}$ is the mean value of $x_1, x_2, \ldots, x_n$ and $\overline{y}$ is the mean value of $y_1, y_2, \ldots, y_n$. Here, we set x_i to be the mean value of five SN^i, that is, values in the last column of Supplementary Material VII, and set y_i to be the number of drug compounds

Table 7: The first 20 prediction accuracies obtained by the integrated method on DS_2 evaluated by 5-fold cross-validation for 5 times.

Order	First time (%)	Second time (%)	Third time (%)	Fourth time (%)	Fifth time (%)	Mean (%)	Standard deviation (%)
1	51.05	50.54	51.37	52.38	52.07	51.48	0.75
2	18.25	19.14	17.42	18.05	18.44	18.26	0.62
3	8.01	6.68	8.58	6.68	6.42	7.27	0.96
4	5.79	5.91	4.45	5.53	5.98	5.53	0.63
5	3.75	4.13	4.70	4.20	4.51	4.26	0.37
6	2.42	2.86	3.05	3.18	2.42	2.78	0.36
7	2.42	2.80	2.42	2.67	2.61	2.58	0.17
8	1.78	1.59	2.10	1.34	2.29	1.82	0.38
9	2.54	1.84	1.84	1.78	1.40	1.88	0.41
10	1.46	2.10	1.46	1.34	1.14	1.50	0.36
11	1.59	1.34	1.21	1.21	1.59	1.39	0.19
12	1.53	1.72	1.65	1.84	1.40	1.63	0.17
13	1.78	1.46	1.72	1.59	1.21	1.55	0.23
14	1.46	1.21	1.53	0.95	1.34	1.30	0.23
15	0.51	1.14	1.02	1.14	1.08	0.98	0.27
16	0.83	0.95	0.57	0.64	0.83	0.76	0.16
17	1.02	0.83	0.95	1.14	0.76	0.94	0.15
18	0.89	0.95	0.70	1.08	0.64	0.85	0.18
19	1.34	0.76	1.34	1.14	0.83	1.08	0.27
20	1.02	1.08	1.02	0.83	1.21	1.03	0.14

labeled by D_i divided by 2,005, that is, $y_i = |S_i|/2005$, where 2,005 was the sum of the number of drug compounds in each category. By (10), the obtained rate was 0.53, yielding that the linear correlation of these two variables was not significant. For example, the categories D_{56} and D_{11} obtained the highest two prediction accuracies (cf. Supplementary Material VII); however, their sizes were only 7 and 14 (cf. Supplementary Material II) which were very small. All of these results indicate that the integrated method performed quite well for the prediction of drug indications.

3.3. Comparison of Different Methods. At a first glance at the Supplementary Material VI, the method based on chemical interactions with $k = 5$ seems to outperform the method based on chemical similarities with fingerprint ECFP_4 and $k = 2$. However, these predicted results were derived from two different datasets. To make a comparison using the same dataset, we executed the method based on chemical similarities with fingerprint ECFP_4 and $k = 2$ on $DS^{(i)}$, in which each sample can be predicted by the method based on chemical interactions. It was also evaluated by jackknife test. Listed in columns 2 and 3 of Supplementary Material VIII are the prediction accuracies obtained by the methods for the prediction of indications that samples in $DS^{(i)}$ can treat. The 1st order prediction accuracy by the method based on chemical interactions was 58.48%, while it was 42.52% by the method based on chemical similarities. To compare the performance of the methods more thoroughly, we calculated Recall and Precision for the first t order predictions and plot two curves with Recalls as their X-axis and Precisions as their Y-axis. Figure 2 shows the two curves, from which we can see that the Recall and Precision obtained by the method based on chemical interactions are always higher than those obtained by the method based on chemical similarities. All of these indicate that the method based on chemical interactions is superior to the method based on chemical similarities for the prediction of drug indications. Thus, we arranged the method based on chemical interactions as the first choice while the method based on chemical similarities as a backup. The arrangement in this study conforms to the results in Chen et al.'s study [16]. The main reason is that the confidence score of an interaction between two compounds, which was used in the method based on chemical interactions, contains different kinds of information of compounds, such as their activities, structures, reactions, and so forth [18], while the method based on chemical similarities only used the information of compound structures.

The integrated method proposed in this study sequentially used the confidence scores of interactions between chemicals and similarity scores of chemicals. Another simple integrated scheme, termed as the method based on integrated scores, is to combine these scores in advance and then make prediction. Given a query drug d_q, the score that d_q can treat indication D_j was computed by

$$R^{\text{integrated}}\left(d_q \Longrightarrow D_j\right) = \frac{R^i\left(d_q \Rightarrow D_j\right) + R^s\left(d_q \Rightarrow D_j\right)}{2}, \quad j = 1, 2, \ldots, 56, \tag{11}$$

where s is ECFP_4 and the parameters k in $R^i(d_q \Rightarrow D_j)$ and $R^s(d_q \Rightarrow D_j)$ were 5 and 2, respectively. The following procedure was same as those of the method based on chemical interactions and chemical similarities.

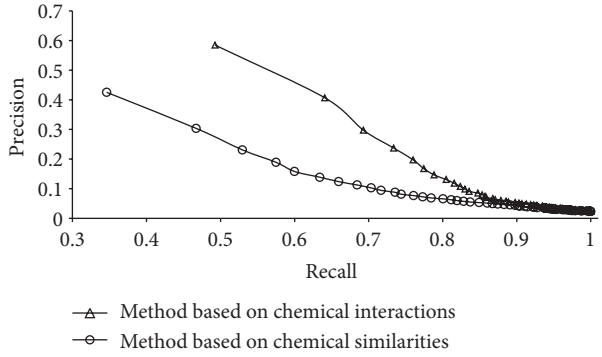

Figure 2: Two curves with Recalls as their X-axis and Precisions as their Y-axis. Recalls and precisions were obtained by method based on chemical interactions with $k = 5$ and method based on chemical similarities with fingerprint ECFP_4 and $k = 2$.

The original motive of employing this method is to make comparison with the proposed method. However, since $R^i(d_q \Rightarrow D_j) = 0$ $(j = 1, 2, \ldots, 56)$ for each sample in $\mathrm{DS}^{(s)}$, that is, the predicted results obtained by the method based on chemical similarities and the method based on integrated scores on $\mathrm{DS}^{(s)}$ were same, the method based on integrated scores was conducted on $\mathrm{DS}^{(i)}$ evaluated by jackknife test. The obtained prediction accuracies were listed in column 4 of Supplementary Material VIII, from which we can see that the 1st order prediction accuracy was 58.82%. It was almost same as that of the method based on chemical interactions with $k = 5$, while it was much higher than that of the method based on chemical similarities with fingerprint ECEP_4 and $k = 2$. It can be easily inferred that this integrated scheme and the method based on chemical interactions were almost at the same level. Since the confidence score of two chemicals, used in the method based on chemical interactions, contains the information of their similarity information [18], that is, the score calculated by (5) and added to (11) was redundant, it is reasonable that the performance of these two methods was almost the same. It can be further inferred that the performance of the method based on integrated scores and that of the proposed method were also at the same level, because the predicted results obtained by the method based on chemical similarities and the method based on integrated scores on $\mathrm{DS}^{(s)}$ were the same.

3.4. Performance of the Integrated Method on DS_{te}. The integrated method combined the method based on chemical interactions with $k = 5$ and the method based on chemical similarities with fingerprint ECEP_4 and $k = 2$. To test the generalization of this method, it was conducted on $\mathrm{DS_{te}}$ to predict indications of drug compounds in it. To calculate the prediction accuracy, the original indications and reported indications of each sample in $\mathrm{DS_{te}}$ were combined together as the known indications, thereby yielding the 1st prediction accuracy of 50.00%, which is almost identical to the 1st prediction accuracy obtained by the method on $\mathrm{DS_2}$. Furthermore, the 2nd prediction accuracy was 21.88%. All of these suggest that the proposed method has a good generalization.

3.5. Illustration of the Predictive Results. Since 5-fold cross-validation is unstable, that is, different partitions may produce different predictions for a given sample, the analysis of the results evaluated by 5-fold cross-validation is not very reliable. On the other hand, jackknife test can avoid this case. In view of this, the integrated method was again conducted on $\mathrm{DS_2}$, evaluated by jackknife test. The obtained prediction accuracies for the methods based on chemical interactions and chemical similarities and integrated method were available as Supplementary Material IX. The 1st order prediction accuracies of the method based on chemical interactions on $\mathrm{DS}^{(i)}$, the method based on chemical similarities on $\mathrm{DS}^{(s)}$, and the integrated method on $\mathrm{DS_2}$ were 58.48%, 47.27%, and 53.66%, respectively, which were a little higher than the corresponding methods on the datasets evaluated by 5-fold cross-validation. In addition, the Recalls of the first two predictions for three methods were 64.08%, 51.38%, and 58.61%, respectively, while the Precisions were 40.68%, 30.21%, and 36.17% for three methods, respectively. In the following paragraphs of this section, further discussions were described based on predictions of each sample in $\mathrm{DS_2}$ and $\mathrm{DS_{te}}$.

Interestingly, some examples in $\mathrm{DS_2}$ showed that the new clinical indications were predicted in the first 2 order predictive diseases based on chemical similarities. From the jackknife test of the dataset $\mathrm{DS_2}$ which contains 1,573 drug compounds, we analyzed several examples that new indications were accurately predicted which were not included in the original datasets. We presented the results as follows: thalidomide (CID000005426), whose original indication is antiemetic in pregnancy [56] and new indication is multiple myeloma (acted as TNF-α inhibitor) [57], is predicted to treat diseases such as antineoplastic (1st order prediction, new clinical indication) and antibacterial (2nd order prediction); leflunomide (CID000003899), whose original indication is rheumatoid arthritis (targeted at DHODH) [58] and new indication is prostate cancer (targeted at PDGEF, EGFR, FGFR and NF-κB) [59], is predicted to treat disease such as antineoplastic (1st order prediction, new clinical indication) and antiinflammatory (2nd order prediction); chlorpromazine (CID000002726), whose original indication is antiemetic (antihistamine) [60] and new indication is nonsedating tranquillizer (dopamine receptor blockade) [61], is predicted to treat disease such as Anxiolytic (1st order prediction, new clinical indication) and antipsychotic (2nd order prediction).

The indications of samples in $\mathrm{DS_{te}}$ were also predicted by our method. As described in Section 3.4, the 1st order prediction accuracy was 16/32 = 50.00% and the 2nd order prediction accuracy was 7/32 = 21.88%. Meanwhile, 20 out of 32 drugs were correctly predicted for the first two orders, where 15 out of 32 drugs were predicted correctly in aspect of original indications and 8 out of 32 drugs were predicted correctly in aspect of repositioned indication, although 3 out of the 8 drugs were predicted correctly responding to

Table 8: 8 instances to illuminate accurate prediction of new indications in validation test dataset.

Name	ID	1st order prediction	2nd order prediction	Original indication	New indication
Rapamycin	CID005284616	Antineoplastic[a]	Antiinflammatory[c]	Immunosuppressant (acted as mTOR inhibitor) [67]	Colorectal cancer, lymphoma, leukemia [68, 69]
Zoledronic	CID000068740	Antineoplastic[a]	Antiinflammatory[c]	Antibone resorption (acted as osteoclast inhibitor) [70]	Multiple myeloma, Prostate cancer, breast cancer [71, 72]
Wortmannin	CID000312145	Antidiabetic[b]	Antineoplastic[a]	Antifungal [25]	Leukemia [73]
Galantamine	CID000009651	Anti-Alzheimer's disease[a]	Antihypertensive[c]	Polio (acted as acetylcholinesterase inhibitor) [1]	Alzheimer's disease [1]
Ropinirole	CID000005095	Antipsychotic[c]	Antiparkinsonian[a]	Antihypertension (acted as dopamine-2 agonist) [1]	Parkinson's disease [1]
Zidovudine	CID000035370	Antiviral[b]	Antineoplastic[a]	Anticancer [1]	Anti-HIV [1]
Allopurinol	CID000002094	Uricosuric[a]	Antineoplastic[b]	Tumor lysis syndrome [74]	Gout [75]
Colesevelam	CID000160051	Antihypolipidemic[b]	Antidiabetic[a]	Antihyperlipidemia [64]	Type 2 diabetes mellitus [65, 66]

a: correctly predicted in new indications;
b: correctly predicted in original indications;
c: incorrectly predicted in original indications.

the original indication. The description of 8 instances with accurate prediction of new indication in validation test set was shown in Table 8.

Further, some of our predictions are supported by *in vitro* assay results from different sources, which may provide mechanism-based interpretation of these potential novel indications. For example, for Quinacrine (CID000000237), the 2nd ranked indication is antiinflammatory. Several researches [62, 63] indicated that Quinacrine is an inhibitor of cytosolic phospholipase A2, which selectively hydrolyzes arachidonyl phospholipids in the sn-2 position releasing arachidonic acid. Together with the lysophospholipid activity, quinacrine is implicated in the initiation of the inflammatory response. The predicted indication of Colesevelam (CID00000160051) is antidiabetic (2nd indication). As we know, Colesevelam acts as bile acid sequestrants in the gastrointestinal tract upregulate bile acid synthesis (via cholesterol 7-alpha-hydroxylase) by means of utilizing cholesterol and reduced low-density lipoprotein cholesterol levels [64]. Although the exact mechanism of action for the glucose-lowering effect of Colesevelam is still unclear, it may exert the glycemic effect by altering the interaction of the bile acid pathways [65, 66]. From the above two cases, we may find that the prediction of our model may provide useful information for identifying new possible indications of some existing drugs.

These results demonstrated that our method can successfully identify some potential new indications for a drug, which supported the hypothesis that "similar drugs" are more likely to have the same therapeutic effects. In our method, interacted drugs were also considered "similar drugs."

4. Conclusions

In the study, we built an effective classifier to predict drug indications based on chemical interactions extracted from STITCH database and chemical structure similarity. The predictor based on chemical interactions outperformed the predictor based on chemical similarities. Therefore, we arranged chemical interaction before chemical similarity to build the predictor for each drug; that is, if the disease indications of a drug cannot be predicted by chemical interaction, then they are predicted by chemical similarity. As a result, the Recall rate and Precision of the first two predictions are 56.28% and 34.87%, respectively. As to the independent test set, the model yielded the accuracy of 50.00% for the 1st prediction and 21.88% for the 2nd prediction. And interestingly, some drug repositioning instances are correctly implicated by our method. A limitation of the method is that only 56 categories of drug indications are analyzed, which may be improved with the expansion of the drug indication data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Guohua Huang and Yin Lu contributed equally to this work.

Acknowledgments

This work was supported by Grants from National Basic Research Program of China (2011CB510101, 2011CB510102), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), the Grant of "The First-Class Discipline of Universities in Shanghai," National Science Foundation of China (31371335, 11371008, 91230201), Scientific Research Fund of Hunan Provincial Science and Technology Department (2014FJ3013), Hunan National Science Foundation

(Grant: 11JJ5001), and Scientific Research Fund of Hunan Provincial Education Department (Grant: 11C1125).

References

[1] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[2] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature Chemical Biology*, vol. 4, no. 11, pp. 682–690, 2008.

[3] S. Usdin, "Industry development: pipeline or flatline?" *BioCentury*, vol. 1, 2002.

[4] C. R. Chong and D. J. Sullivan Jr., "New uses for old drugs," *Nature*, vol. 448, no. 7154, pp. 645–646, 2007.

[5] J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning," *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 303–311, 2011.

[6] T. Noeske, B. C. Sasse, H. Stark, C. G. Parsons, T. Weil, and G. Schneider, "Predicting compound selectivity by self-organizing maps: cross-activities of metabotropic glutamate receptor antagonists," *ChemMedChem*, vol. 1, no. 10, pp. 1066–1068, 2006.

[7] F. Iorio, R. Bosotti, E. Scacheri et al., "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 33, pp. 14621–14626, 2010.

[8] M. J. Keiser, V. Setola, J. J. Irwin et al., "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.

[9] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[10] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002503, 2012.

[11] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, article 496, 2011.

[12] M. Re and G. Valentini, "Large scale ranking and repositioning of drugs with respect to DrugBank therapeutic categories," in *Bioinformatics Research and Applications*, vol. 7292 of *Lecture Notes in Computer Science*, pp. 225–236, 2012.

[13] E. Kotelnikova, A. Yuryev, I. Mazo, and N. Daraselia, "Computational approaches for drug repositioning and combination therapy design," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 3, pp. 593–606, 2010.

[14] J. Li, X. Zhu, and J. Y. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," *PLoS Computational Biology*, vol. 5, no. 7, Article ID e1000450, 2009.

[15] A. P. Chiang and A. J. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses," *Clinical Pharmacology and Therapeutics*, vol. 86, no. 5, pp. 507–510, 2009.

[16] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[17] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[18] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[19] L. L. Hu, C. Chen, T. Huang, Y. D. Cai, and K. C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS ONE*, vol. 6, no. 12, Article ID e29491, 2011.

[20] M. Dunkel, S. Günther, J. Ahmed, B. Wittig, and R. Preissner, "SuperPred: drug classification and target prediction," *Nucleic Acids Research*, vol. 36, pp. W55–59, 2008.

[21] *Comprehensive Medicinal Chemistry Database*, Accelrys, San Diego, Calif, USA, 2011.

[22] P. Du, T. Li, and X. Wang, "Recent progress in predicting protein sub-subcellular locations," *Expert Review of Proteomics*, vol. 8, no. 3, pp. 391–404, 2011.

[23] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, and K.-C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, no. 1, Article ID e14556, 2011.

[24] S. C. Gupta, B. Sung, S. Prasad, L. J. Webb, and B. B. Aggarwal, "Cancer drug discovery by repurposing: teaching new tricks to old dogs," *Trends in Pharmacological Sciences*, vol. 34, no. 9, pp. 508–517, 2013.

[25] B. M. Padhy and Y. K. Gupta, "Drug repositioning: re-investigating existing drugs for new therapeutic indications," *Journal of Postgraduate Medicine*, vol. 57, no. 2, pp. 153–160, 2011.

[26] L. Chen, J. Lu, T. Huang et al., "Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions," *PLoS ONE*, vol. 9, no. 9, Article ID e107767, 2014.

[27] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, "Do structurally similar molecules have similar biological activity?" *Journal of Medicinal Chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.

[28] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 6, pp. 983–996, 1998.

[29] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen, "Recent developments of the Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics," *Current Pharmaceutical Design*, vol. 12, no. 17, pp. 2111–2120, 2006.

[30] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 119–129, 2006.

[31] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork, "Extraction of regulatory gene/protein networks from Medline," *Bioinformatics*, vol. 22, no. 6, pp. 645–650, 2006.

[32] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, pp. 31–36, 1988.

[33] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: an open chemical toolbox," *Journal of Cheminformatics*, vol. 3, article 33, no. 10, 2011.

[34] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1273–1280, 2002.

[35] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[36] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi, "Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity," *Bioinformatics*, vol. 21, no. 1, pp. i359–i368, 2005.

[37] RDKit: Open-source cheminformatics, http://www.rdkit.org/.

[38] L. Chen, J. Lu, J. Zhang, K.-R. Feng, M.-Y. Zheng, and Y.-D. Cai, "Predicting chemical toxicity effects based on chemical-chemical interactions," *PLoS ONE*, vol. 8, no. 2, Article ID e56517, 2013.

[39] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[40] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, and N. Deng, "Predicting DNA- and RNA-binding proteins from sequences with kernel methods," *Journal of Theoretical Biology*, vol. 258, no. 2, pp. 289–293, 2009.

[41] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[42] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[43] Y.-S. Ding and T.-L. Zhang, "Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1887–1892, 2008.

[44] C. E. Jones, U. Baumann, and A. L. Brown, "Automated methods of predicting the function of biological sequences using GO and BLAST," *BMC Bioinformatics*, vol. 6, article 272, 2005.

[45] L. Chen, W.-M. Zeng, Y.-D. Cai, and T. Huang, "Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set," *Current Bioinformatics*, vol. 8, no. 2, pp. 200–207, 2013.

[46] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[47] L. Chen, B.-Q. Li, and K.-Y. Feng, "Predicting biological functions of protein complexes using graphic and functional features," *Current Bioinformatics*, vol. 8, no. 5, pp. 545–551, 2013.

[48] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[49] L. Chen, J. Lu, N. Zhang, T. Huang, and Y.-D. Cai, "A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes," *Molecular BioSystems*, vol. 10, no. 4, pp. 868–877, 2014.

[50] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI '95)*, vol. 2, pp. 1137–1143, San Mateo, Calif, USA, 1995.

[51] B.-Q. Li, K.-Y. Feng, L. Chen, T. Huang, and Y.-D. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS ONE*, vol. 7, no. 8, Article ID e43927, 2012.

[52] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[53] K.-J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino pairs," *Bioinformatics*, vol. 19, no. 13, pp. 1656–1663, 2003.

[54] M. Bhasin and G. P. S. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, no. 3, pp. 421–423, 2004.

[55] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, no. 1, pp. i38–i46, 2005.

[56] M. E. Franks, G. R. Macpherson, and W. D. Figg, "Thalidomide," *The Lancet*, vol. 363, no. 9423, pp. 1802–1811, 2004.

[57] T. A. Fehniger, J. C. Byrd, G. Marcucci et al., "Single-agent lenalidomide induces complete remission of acute myeloid leukemia in patients with isolated trisomy 13," *Blood*, vol. 113, no. 5, pp. 1002–1005, 2009.

[58] S. Teschner and V. Burst, "Leflunomide: a drug with a potential beyond rheumatology," *Immunotherapy*, vol. 2, no. 5, pp. 637–650, 2010.

[59] Y.-J. Ko, E. J. Small, F. Kabbinavar et al., "A multi-institutional phase II study of SU101, a platelet-derived growth factor receptor inhibitor, for patients with hormone-refractory prostate cancer," *Clinical Cancer Research*, vol. 7, no. 4, pp. 800–805, 2001.

[60] M. R. Knapp and H. K. Beecher, "Postanesthetic nausea, vomiting, and retching; evaluation of the antiemetic drugs dimenhydrinate (dramamine), chlorpromazine, and pentobarbital sodium," *Journal of the American Medical Informatics Association*, vol. 160, pp. 376–385, 1956.

[61] R. G. McCreadie, "Managing the first episode of schizophrenia: the role of new therapies," *European Neuropsychopharmacology*, vol. 6, supplement 2, pp. S2–S2, 1996.

[62] M. Hellstrand, E. Eriksson, and C. L. Nilsson, "Dopamine D2 receptor-induced COX-2-mediated production of prostaglandin E_2 in D_2-transfected Chinese hamster ovary cells without simultaneous administration of a Ca^{2+}-mobilizing agent," *Biochemical Pharmacology*, vol. 63, no. 12, pp. 2151–2158, 2002.

[63] W.-Y. Ong, X.-R. Lu, B. K.-C. Ong, L. A. Horrocks, A. A. Farooqui, and S.-K. Lim, "Quinacrine abolishes increases in cytoplasmic phospholipase A_2 mRNA levels in the rat hippocampus after kainate-induced neuronal injury," *Experimental Brain Research*, vol. 148, no. 4, pp. 521–524, 2003.

[64] A. Corsini, E. Windler, and M. Farnier, "Colesevelam hydrochloride: usefulness of a specifically engineered bile acid sequestrant for lowering LDL-cholesterol," *European Journal of Cardiovascular Prevention and Rehabilitation*, vol. 16, no. 1, pp. 1–9, 2009.

[65] P. Levy and P. S. Jellinger, "The potential role of colesevelam in the management of prediabetes and type 2 diabetes mellitus," *Postgraduate Medicine*, vol. 122, no. 3, pp. 1–8, 2010.

[66] B. Staels, "A review of bile acid sequestrants: potential mechanism(s) for glucose-lowering effects in type 2 diabetes mellitus," *Postgraduate medicine*, vol. 121, no. 3, pp. 25–30, 2009.

[67] Y. Alvarado, M. M. Mita, S. Vemulapalli, D. Mahalingam, and A. C. Mita, "Clinical activity of mammalian target of rapamycin inhibitors in solid tumors," *Targeted Oncology*, vol. 6, no. 2, pp. 69–94, 2011.

[68] C. Récher, O. Beyne-Rauzy, C. Demur et al., "Antileukemic activity of rapamycin in acute myeloid leukemia," *Blood*, vol. 105, no. 6, pp. 2527–2534, 2005.

[69] C. Sillaber, M. Mayerhofer, A. Böhm et al., "Evaluation of antileukaemic effects of rapamycin in patients with imatinib-resistant chronic myeloid leukaemia," *European Journal of Clinical Investigation*, vol. 38, no. 1, pp. 43–52, 2008.

[70] G. J. Morgan, F. E. Davies, W. M. Gregory et al., "First-line treatment with zoledronic acid as compared with clodronic acid in multiple myeloma (MRC Myeloma IX): a randomised controlled trial," *The Lancet*, vol. 376, no. 9757, pp. 1989–1999, 2010.

[71] M. Gnant, B. Mlineritsch, W. Schippinger et al., "Endocrine therapy plus zoledronic acid in premenopausal breast cancer," *The New England Journal of Medicine*, vol. 360, no. 7, pp. 679–691, 2009.

[72] G. Facchini, M. Caraglia, A. Morabito et al., "Metronomic administration of zoledronic acid and taxotere combination in castration resistant prostate cancer patients. Phase I ZANTE trial," *Cancer Biology and Therapy*, vol. 10, no. 6, pp. 543–548, 2010.

[73] P. Workman, P. A. Clarke, F. I. Raynaud, and R. L. M. van Montfort, "Drugging the PI3 kinome: from chemical tools to drugs in the clinic," *Cancer Research*, vol. 70, no. 6, pp. 2146–2157, 2010.

[74] S. Jeha, "Tumor lysis syndrome," *Seminars in Hematology*, vol. 38, supplement 10, no. 4, pp. 4–8, 2001.

[75] G. B. Elion, "The purine path to chemotherapy," *Science*, vol. 244, no. 4900, pp. 41–47, 1989.