OXFORD

Briefings in Bioinformatics, 18(2), 2017, 333–347

doi: 10.1093/bib/bbw012
Advance Access Publication Date: 4 March 2016
Paper

SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning

Zengrui Wu, Feixiong Cheng, Jie Li, Weihua Li, Guixia Liu and Yun Tang

Feixiong Cheng is presented at Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Harvard Medical School, 450 Brookline Avenue, Boston, MA 02215, USA, and Center for Complex Networks Research, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, MA 02115, USA.

Corresponding author. Feixiong Cheng, State Key Laboratory of Biotherapy/Collaborative Innovation Center for Biotherapy, West China Hospital, West China Medical School, Sichuan University, Chengdu, Sichuan 610041, China. Tel.: +86-21-6425-1052; Fax: +86-21-6425-3651. E-mail: fxcheng1985@gmail.com; Yun Tang, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. Tel.: +86-21-6425-1052; Fax: +86-21-6425-3651. E-mail: ytang234@ecust.edu.cn

Abstract

Computational prediction of drug–target interactions (DTIs) and drug repositioning provides a low-cost and high-efficiency approach for drug discovery and development. The traditional social network-derived methods based on the naïve DTI topology information cannot predict potential targets for new chemical entities or failed drugs in clinical trials. There are currently millions of commercially available molecules with biologically relevant representations in chemical databases. It is urgent to develop novel computational approaches to predict targets for new chemical entities and failed drugs on a large scale. In this study, we developed a useful tool, namely substructure–drug–target network-based inference (SDTNBI), to prioritize potential targets for old drugs, failed drugs and new chemical entities. SDTNBI incorporates network and chemoinformatics to bridge the gap between new chemical entities and known DTI network. High performance was yielded in 10-fold and leave-one-out cross validations using four benchmark data sets, covering G protein-coupled receptors, kinases, ion channels and nuclear receptors. Furthermore, the highest areas under the receiver operating characteristic curve were 0.797 and 0.863 for two external validation sets, respectively. Finally, we identified thousands of new potential DTIs via implementing SDTNBI on a global network. As a proof-of-principle, we showcased the use of SDTNBI to identify novel anticancer indications for nonsteroidal anti-inflammatory drugs by inhibiting AKR1C3, CA9 or CA12. In summary, SDTNBI is a powerful network-based approach that predicts potential targets for new chemical entities on a large scale and will provide a new tool for DTI prediction and drug repositioning. The program and predicted DTIs are available on request.

Zengrui Wu is a PhD student at East China University of Science and Technology, China, developing novel network-based methods and computational tools for the prediction of drug–target interactions and drug repositioning.
Feixiong Cheng is a research fellow in Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute at Harvard Medical School, and Center for Complex Networks Research at Northeastern University, USA. His current research interests include network medicine, systems biology and systems pharmacology.
Jie Li is a PhD student at East China University of Science and Technology, China, working mostly on systems pharmacology and microRNA pharmacogenomics studies.
Weihua Li is an associate professor at East China University of Science and Technology, China. His recent interests include molecular modeling and simulations of disease-related proteins.
Guixia Liu is a professor at East China University of Science and Technology, China. Her recent research activity focuses on computer-aided drug design.
Yun Tang is a professor at East China University of Science and Technology and director of Laboratory of Molecular Modeling and Design at the Shanghai Key Laboratory of New Drug Design, East China University of Science and Technology, China. His recent research interests scan systems pharmacology, computational toxicology and computational medicinal chemistry. He has authored over 120 publications.
Submitted: 30 October 2015; **Received (in revised form):** 6 January 2016

Key words: drug–target interaction; drug repositioning; network-based inference; chemoinformatics; failed drug; chemical substructure

Introduction

Over the past decades, traditional drug discovery paradigm has achieved a measure of success. But since the new century, this paradigm based on the hypothesis of 'one gene, one drug, one disease' has been facing many tricky challenges, such as high clinical attrition rate [1, 2]. Because of the rapid development of systems biology, a new paradigm named network pharmacology has put forward some fresh ideas and concepts to solve the aforementioned challenges [2]. Under this novel perspective, a single drug might act on multiple targets. Hence, drugs and targets can be organized as a complex network through numerous drug–target interactions (DTIs). Identification of new DTIs will help investigators to understand the therapeutic profiles or side effects of drugs, and find new uses for old drugs, namely drug repositioning [3–6]. However, to identify potential DTIs with low cost and high efficiency is still a big challenge. Therefore, it is urgently needed to develop novel approaches for DTI prediction and drug repositioning during drug discovery and development.

Comparing with traditional experimental assays, computational approaches have made it possible for us to quickly and inexpensively identify potential DTIs and repurpose existing drugs [6, 7]. These approaches include molecular docking [8, 9], machine learning [10–12], similarity-based methods [13, 14], etc. In recent years, another important series of computational approaches, namely network-based methods, was proposed for DTI prediction and drug repositioning. For instance, our group presented three network-based methods derived from recommendation algorithms for social networks, including network-based inference (NBI), drug-based similarity inference and target-based similarity inference. Several newly predicted interactions via NBI for five old drugs on estrogen receptors or dipeptidyl peptidase IV were validated by *in vitro* assays [15]. Furthermore, we improved the NBI method by assigning weighted values to edges or nodes, namely edge-weighted NBI (EWNBI) and node-weighted NBI (NWNBI), respectively [16]. The systematic evaluation revealed that the two weighted NBI methods marginally outperform the original NBI. Alaimo *et al.* [17] introduced a network-based method called domain tuned-hybrid (DT-Hybrid) derived from the NBI method. High performance was yielded for the DT-Hybrid by considering the additional knowledge about drug similarity and target similarity. Chen *et al.* [18] reported a method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH), which was an improvement of the random walk method by integrating three types of networks.

Despite their successful applications in DTI prediction and drug repositioning for known drugs, the aforementioned methods have an enormous limitation in the application domain: they cannot be used to predict potential targets for new chemical entities. These traditional network-based methods only used the naïve topology information in the existing DTI network [16–18]. Thus, they cannot interlink known DTI network with molecules without known target, such as new synthesized chemical structures and drugs failed in clinical trials. However, there are currently over 68 million commercially available molecules, including a large portion of described natural products in the ZINC database [19]. In addition, thousands of drugs with good pharmacokinetic properties and low toxicity are failed in clinical phases II and III because of poor clinical utility on the specific targets [20, 21]. The US National Center for Advancing Translational Sciences is spending US$20 million to focus initially on repurposing 58 failed drugs [22]. Therefore, it becomes urgent and important to develop novel computational approaches, such as network-based approaches, to predict potential targets for new chemical entities and failed drugs on a large scale.

In this study, we proposed an integrated network and chemoinformatics tool, named substructure–drug–target network-based inference (SDTNBI), for large-scale DTI prediction and drug repositioning. SDTNBI uses the chemical substructure, which is a defined series of features that can be shared by chemical structures, to bridge the gap between known drugs and new chemical entities. Previous studies have suggested that chemical substructures play crucial roles for computational evaluation of drug pharmacokinetics and DTI prediction [10, 23–25]. Specifically, SDTNBI integrates known DTI network, drug–substructure linkages and new chemical entity–substructure linkages to infer new targets for old drugs, failed drugs and new chemical entities in a way of resource diffusion (Figure 1). Systematic evaluation based on 10-fold cross validation, leave-one-out cross validation and external validation showed high accuracy and robustness of SDTNBI. Furthermore, we built a global network and computationally identified thousands of new potential DTIs via SDTNBI. As a proof-of-principle, we showcased the use of SDTNBI to identify novel anticancer indications for nonsteroidal anti-inflammatory drugs (NSAIDs) by inhibiting AKR1C3, CA9, CA12 or CDK2. Put together, SDTNBI will provide a new alternative tool for DTI prediction and drug repositioning on a large scale during drug discovery and development.

Materials and methods

Data sets

Two DTI networks containing known chemical–protein interactions for G protein-coupled receptors (GPCRs) and kinase superfamily (Kinases) were collected from the ChEMBL database (accessed in May 2010) [26]. Two external validation sets corresponding to the two DTI networks were collected from the DrugBank database [27]. These DTI networks and external validation sets of GPCRs and Kinases were prepared based on our previous study [16]. In addition, we further compiled two DTI networks covering ion channels (ICs) and nuclear receptors (NRs) to cover more target families by integrating bioactivity data from the ChEMBL database (version 19) [26] and BindingDB database (downloaded in 2014) [28]. Chemical structures were converted into canonical SMILES format. Five physicochemical properties: molecular weight (MW), log P, the number of hydrogen bond donors (HBDs) and the number of hydrogen bond acceptors (HBAs), were calculated by using the OpenBabel toolkit (version 2.3.2) [29]. Only those data items that met the following five criteria were retained: (i) K_i, K_d, IC_{50} or $EC_{50} \leq 10$ μM; (ii) the target is a human protein; (iii) the target can be represented in a unique UniProt accession number; (iv) the compound can be successfully represented in canonical SMILES format; (v) the compound is drug-like, namely MW $\leq$ 500 Dalton, log P $\leq$ 5, the

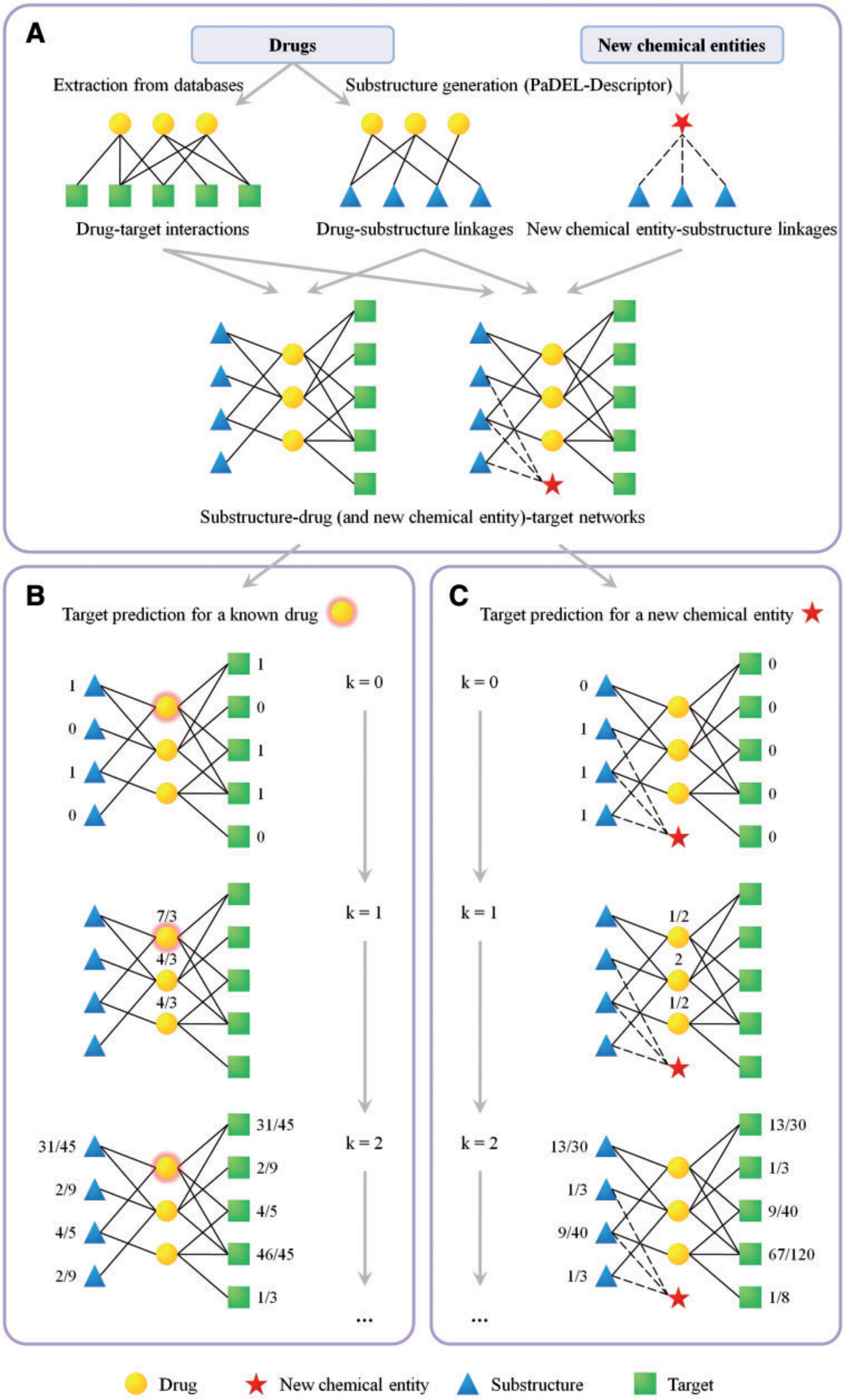

A
Drugs
New chemical entities
Extraction from databases
Substructure generation (PaDEL-Descriptor)
Drug-target interactions
Drug-substructure linkages
New chemical entity-substructure linkages
Substructure-drug (and new chemical entity)-target networks
B
Target prediction for a known drug
k = 0
k = 1
k = 2
...
1
0
1
0
1
0
1
1
0
7/3
4/3
4/3
31/45
2/9
4/5
2/9
31/45
2/9
4/5
46/45
1/3
C
Target prediction for a new chemical entity
k = 0
k = 1
k = 2
...
0
1
1
1
0
0
0
0
0
1/2
2
1/2
13/30
1/3
9/40
1/3
13/30
1/3
9/40
67/120
1/8
Drug
New chemical entity
Substructure
Target

Figure 1. Schematic diagram of SDTNBI. (**A**) The process of substructure–drug (and new chemical entity)–target network construction, (**B**) an example of the process of predicting targets for a known drug represented as a circle with glow effect, (**C**) an example of the process of predicting targets for a new chemical entity represented as a star. Circle: drug, star: new chemical entity, triangle: chemical substructure, rectangle: protein target, solid line: drug–substructure linkage and drug–target interaction, dash line: new chemical entity–substructure linkage, k: the number of resource-spreading processes.

number of HBDs ≤ 5 and the number of HBAs ≤ 10. Finally, two DTI networks of ICs and NRs were built by removing duplicates from different data sources.

Furthermore, in order to construct a global network covering genome-wide targets, a comprehensive DTI network was built. All drugs in this data set were downloaded from the DrugBank database (version 4.1) [27]. Only those molecules whose structure was explicitly given were used. Molecules with conflicting structures were removed. Bioactivity data of these selected drugs were extracted from the aforementioned prepared databases using the same filtering criteria, except drug-like properties.

Chemical substructure dictionary

In this study, we systematically evaluated the performance of seven types of fingerprints generated by the PaDEL-Descriptor software (version 2.18) [30], including CDK Fingerprint (CDK), CDK Extended Fingerprint (CDKExt), CDK Graph Only Fingerprint (Graph), MACCS Fingerprint (MACCS), PubChem Fingerprint (PubChem), Substructure Fingerprint (FP4) and Klekota-Roth Fingerprint (KR). These fingerprints were used to represent chemical substructures for each molecule. In addition, salt ions were removed from molecules, and nitro groups were standardized in this step.

Description of SDTNBI

Network construction

As described in Figure 1A, known DTI bipartite networks and drug–substructure linkages were obtained after data preparation and chemical substructure generation. They were subsequently integrated to construct a substructure–drug–target network. Denoting that $D = \{D_1, D_2, \ldots, D_{N_D}\}$ is a set of N_D known drugs, $S = \{S_1, S_2, \ldots, S_{N_S}\}$ is a set of N_S chemical substructures and $T = \{T_1, T_2, \ldots, T_{N_T}\}$ is a set of N_T targets. The substructure–drug–target network can be represented as a tripartite graph G(V, E), where $V = D \cup S \cup T$ is the set of its vertices, and E is the set of its edges containing DTIs and drug–substructure linkages.

Such substructure–drug–target network can be used to prioritize potential DTIs for known drugs appearing in the existing network via calculation processes explained in next subsection, but not enough for target prediction of new chemical entities. To solve this issue, we built linkages between the new chemical entities and the known substructure–drug–target network through substructures shared by the new chemical entities and known drugs in the existing DTI network. Let $C = \{C_1, C_2, \ldots, C_{N_C}\}$ be a set of N_C new chemical entities. This substructure–drug (and new chemical entity)–target network, which is an extension of graph G, can be represented as a new graph G′(V′, E′), where $V' = V \cup C$ is the set of its vertices, and E′ is the set of its edges containing DTIs, drug–substructure linkages and newly added compound–substructure linkages.

Prioritizing targets for known drugs

Potential targets can be recommended to known drugs by using resource diffusion processes in the substructure–drug–target network, namely aforementioned graph G. For each drug D_i in the network, it has initial resources located in both its targets and its substructures. Initially, each substructure and each target of D_i equally spread their resources to neighbor drugs. Subsequently, each of those drugs equally spreads its resources to neighbor nodes. Thus, D_i will obtain final resources located in several targets, suggesting that D_i may have potential interactions with these targets. The amount of the resources of D_i located in target T_j can be seen as the score of the interaction between D_i and T_j, where higher score implies higher possibility that D_i can interact with T_j. Such two steps of resource reallocation can be continued in this way. It is noteworthy that, only when the number of resource-spreading processes (symbolized as k) is even, initial resources can be reallocated to the vertices standing for targets rather than vertices standing for drugs. Figure 1B showed an example of prioritizing potential targets for a known drug (called D_e here) in a small substructure–drug–target network. In the initial state ($k = 0$), scores located in the neighbor nodes of D_e were its initial resources. After two resource-spreading processes ($k = 2$), for each target node T_j, the score located in T_j was the score of $\boldsymbol{D_e}$–$\boldsymbol{T_j}$ interaction.

Mathematically, the graph G can be represented by an adjacency matrix A. This is an $(N_D + N_S + N_T)$ order square matrix, defined as:

$$A = \begin{bmatrix} O & M_{DS} & M_{DT} \\ M_{DS}^T & O & O \\ M_{DT}^T & O & O \end{bmatrix} \tag{1}$$

where M_{DS} is a $N_D \times N_S$ matrix and M_{DT} is a $N_D \times N_T$ matrix. They can be respectively defined as:

$$M_{DS}(i,j) = \begin{cases} 1 & \text{if } D_i \text{ is linked with } S_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$M_{DT}(i,j) = \begin{cases} 1 & \text{if } D_i \text{ is linked with } T_j \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Let W be an $(N_D + N_S + N_T)$ order square matrix, defined as:

$$W(i,j) = \frac{A(i,j)}{\sum_{l=1}^{N_D+N_S+N_T} A(i,l)}. \tag{4}$$

The final resource matrix F can be calculated via following equation:

$$F = A \times W^k, \tag{5}$$

where W^k is the transfer matrix and k is the number of resource-spreading processes. The value of $F(i, N_D + N_S + j)$ $(0 < i \le N_D, 0 < j \le N_T)$ is the score of D_i–T_j interaction.

Predicting targets for new chemical entities

In the substructure–drug (and new chemical entity)–target network, namely aforementioned graph G′, new chemical entities can be seen as special drugs that have no known targets. For each new chemical entity C_i in this network, its initial resources only located in its substructures. Each substructure equally spreads its resources to neighbor drugs, and then each of those drugs equally spreads its resources to neighbor nodes. Thus, these targets obtained final resources, which can be regarded as the potential targets of C_i, where higher score implies higher possibility that C_i can act on T_j. Such two steps can also be repeated in this way. Figure 1C showed an example of predicting potential targets for a new chemical entity (called C_e here) in a small substructure–drug (and new chemical entity)–target network. In the initial state ($k = 0$), scores located in the neighbor nodes of C_e were its initial resources. After two resource-spreading processes ($k = 2$), for each target node T_j, the score located in T_j is the score of C_e–T_j interaction.

Mathematically, the adjacency matrix of graph G′ can be represented as an $(N_C + N_D + N_S + N_T)$ order square matrix:

$$A' = \begin{bmatrix} O & O & M_{CS} & O \\ O & O & M_{DS} & M_{DT} \\ M_{CS}^T & M_{DS}^T & O & O \\ O & M_{DT}^T & O & O \end{bmatrix}, \tag{6}$$

where M_{CS} is a $N_C \times N_S$ matrix, defined as:

$$M_{CS}(i,j) = \begin{cases} 1 & \textit{if } C_i \textit{ is linked with } S_j \\ 0 & \textit{otherwise} \end{cases}. \tag{7}$$

Let B be an $(N_C + N_D + N_S + N_T)$ order square matrix, defined as:

$$B = \begin{bmatrix} O & O & O & O \\ O & O & M_{DS} & M_{DT} \\ O & M_{DS}^T & O & O \\ O & M_{DT}^T & O & O \end{bmatrix}. \tag{8}$$

Let W’ be an $(N_C + N_D + N_S + N_T)$ order square matrix, defined as:

$$W'(i,j) = \begin{cases} \dfrac{B(i,j)}{\sum_{l=1}^{N_C+N_D+N_S+N_T} B(i,l)} & \textit{if } \sum_{l=1}^{N_C+N_D+N_S+N_T} B(i,l) \neq 0 \\ 0 & \textit{otherwise} \end{cases}. \tag{9}$$

The final resource matrix F’ can be computed as below:

$$F' = A' \times W'^k, \tag{10}$$

where W'^k is the transfer matrix and k is the number of resource-spreading processes. The value of $F'(i, N_C + N_D + N_S + j)$ $(0 < i \le N_C, 0 < j \le N_T)$ is the score of C_i–T_j interaction.

Benchmark evaluation

The performance of SDTNBI was evaluated by 10-fold cross validation, leave-one-out cross validation and external validation.

Ten-fold cross validation

The 10-fold cross validation was widely used in evaluating network-based methods [5, 15–17]. In a 10-fold cross validation, for the substructure–drug–target network to be evaluated, DTIs were randomly divided into 10 parts. One part was used as the test set in turn. The remnant network, which contained other nine parts and all drug–substructure linkages, was used as the training set. Thus, 10 pairs of training set and test set were obtained. Each pair can be used to calculate a group of evaluation indicators. To reduce randomness of results, the 10-fold cross validation was repeated by 10 times so that 100 groups of evaluation indicators can be used to systematically evaluate the model.

Leave-one-out cross validation

The leave-one-out validation was used to further evaluate the performance of SDTNBI. For the substructure–drug–target network to be evaluated, all DTIs for one drug were extracted to be the test set in turn. The remnant network that contained the DTIs of other drugs and the drug–substructure linkages of this and other drugs was used as the training set. In this way, pairs of training set and test set were generated for the subsequent evaluation indicator calculation. The number of pairs is equal to the number of drugs participated in evaluation.

External validation

The actual generalization ability of predicting potential targets for new chemical entities was also evaluated by several external validation sets. Compounds in the external validation set with known targets were seen as new chemical entities in prediction. After recommended potential targets for those compounds, evaluation indicators were calculated by comparing newly predicted compound–target interactions with known compound–target interactions.

Evaluation indicator calculation

For each pair of training set and test set, nodes that lost all its linkages in the training set were removed from both the training set and the test set. After computational DTI prediction on the training set, all possible DTIs were obtained. For each drug D_i participating in the evaluation, a sorted list of its newly predicted targets was generated. By comparing these predicted target lists with the test set, several evaluation indicators were calculated to show the accuracy and robustness of the models we built, including precision (P), recall (R), precision enhancement (e_P) and recall enhancement (e_R). The details of these measurements can be found in our previous study [16], briefly described as below:

$$P(L) = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{X_i(L)}{L} \tag{11}$$

$$R(L) = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{X_i(L)}{X_i} \tag{12}$$

$$e_P(L) = P(L) \cdot \frac{M \cdot N}{X} \tag{13}$$

$$e_R(L) = R(L) \cdot \frac{N}{L}, \tag{14}$$

where M and N are the number of drugs and targets participated in evaluation (M is always equal to 1 in leave-one-out cross validation), X is the total number of missing DTIs (i.e. DTIs which were divided into test set) of M drugs, X_i is the number of missing DTIs of drug D_i and $X_i(L)$ is the number of true-positive predictions (i.e. missing DTIs which were correctly recovered)

Table 1. Overview of DTIs in different data sets

Data set	Target	N_D	N_T	N_{DT}	*Sparsity (%)*
Networks	GPCRs	4741	97	17,111	3.72
	Kinases	2827	206	13,647	2.34
	ICs	7929	97	8944	1.16
	NRs	5218	35	7366	4.03
	Global	1844	1032	10,185	0.54
External sets	GPCRs	92	46	271	6.40
	Kinases	188	28	202	3.84

Note. N_D = the number of drugs; N_T = the number of targets; N_{DT} = the number of DTIs; Sparsity = the ratio between N_{DT} and the number of all possible DTIs.

ranked in the top L places of the predicted target list of D_i. Moreover, for each pair of training set and test set, a receiver operating characteristic (ROC) curve was generated by computing a series of true-positive rates and false-positive rates under different L (L = 1, 2, ..., N), and a precision-recall curve was also drawn by calculating an array of P(L) and R(L) values under different L (L = 1, 2, ..., N). The areas under these ROC curves were calculated to further show the performance.

Results

Statistics of the benchmark data sets

Five DTI networks and two external validation sets were used in this study (Table 1). They were consisted by (i) GPCRs: 17 111 interactions connecting 4741 molecules and 97 GPCRs, (ii) Kinases: 13 647 interactions connecting 2827 molecules and 206 kinases, (iii) ICs: 8944 interactions connecting 7929 molecules and 97 ICs, (iv) NRs: 7366 interactions connecting 5218 molecules and 35 NRs and (v) the global network covering genome-wide targets (Global): 10 185 interactions connecting 1844 drugs and 1032 target proteins. Specifically, in the construction process of the global network, 6800 drugs with structure information were collected from the DrugBank database. After converting their structures into the canonical SMILES format, we found that 232 of them were conflicting, namely two or more DrugBank IDs shared one canonical SMILES. As mentioned in above section, before extracting bioactivity data from the prepared databases, they were all removed in order to keep the quality of the DTI network. Two external validation sets contained 271 interactions connecting 92 drugs and 46 GPCRs, and 202 interactions connecting 188 drugs and 28 kinases, respectively (Table 1). All drugs in the external validation sets were not included in corresponding DTI networks. Meanwhile, seven types of fingerprints were used to systematically describe chemical substructures and were generated for each drug in these data sets. The number of drugs and drug–substructure linkages are given in Supplementary Table S1. Substructure–drug (and new chemical entity)–target networks were further constructed for prediction and method validation.

Performance evaluation with cross validation

To evaluate the performance of our models, 10-fold cross validation was carried out under different conditions, including different number of resource-spreading processes (k) and different fingerprints. After 10-fold cross validation was repeated 10 times, for each model, its model performance was evaluated by measuring the averages and standard deviations of several selected evaluation indicators, including area under the receiver operating characteristic curve (AUC), and P, R, e_P, e_R under selected L. The relationship between AUC values and k values in 10-fold cross validation is given in Supplementary Table S2. For all network models built via SDTNBI, the AUC value reached its peak when k = 2, and decreased with the increase in k value (Figure 2). The best model performance was independent of different fingerprint types. Hence, 2 was selected as the optimal value of parameter k. Using k = 2, the highest AUC values: 0.966 ± 0.002, 0.958 ± 0.003, 0.971 ± 0.002, 0.932 ± 0.005 and 0.949 ± 0.004 were yielded for the models of GPCRs-FP4, Kinases-KR, ICs-KR, NRs-KR and Global-FP4, respectively. The details of other evaluation indicators in 10-fold cross validation when k = 2 are given in Table 2. Moreover, the precision-recall curves in 10-fold cross validation when k = 2 are given in Supplementary Figure S1. From these evaluation indicators, we found that GPCRs-FP4, Kinases-KR, ICs-KR, NRs-KR and Global-FP4 also showed better performance than other models.

To further evaluate the performance of global network models, leave-one-out cross validation was used under different k values and different fingerprints. The relationship between AUC values and k values in leave-one-out cross validation is given in Supplementary Table S3. We found that the model performance maximized when k = 2, and reduced with the increase in k value. This is consistent with the 10-fold cross validation. Under the optimal condition k = 2, the two highest AUC values of Global were 0.910 ± 0.150 and 0.891 ± 0.150 for the models of Global-KR and Global-FP4, respectively. The details of other evaluation indicators when k = 2 are shown in Table 3. We found that KR and FP4 fingerprints outperform other types of fingerprints, which is also consistent with the 10-fold cross validation.

Hence, based on systematic evaluation, the above results suggested that the performance of KR and FP4 fingerprints, especially KR, was better than CDK, CDKExt and Graph fingerprints in SDTNBI. Considering that KR was defined by chemical substructures enriched for bioactivities [25], the possible explanation is that more specific substructure fragments defined in KR and FP4 could better describe molecules with different bioactivities. In addition, Supplementary Table S1 indicated that the drug–substructure networks with lower sparsity were generated by FP4 or KR. Moreover, the outstanding performance of FP4 or KR also showed in our previous quantitative structure–activity relationship studies [23, 31]. In the future, we will implement and evaluate more types of high-quality fingerprints, such as the Extended Connectivity Fingerprints [32] and three-dimensional fingerprints [33].

Evaluation of the model generalization ability

We further evaluated the model generalization ability of SDTNBI using two external validations sets under different parameter k and different fingerprint types. Consistent with the 10-fold cross validation, for each type of fingerprint, the AUC value reached maximum when k = 2, and decreased with the increase of k value (Supplementary Table S4). Using k = 2, the highest AUC values, 0.797 and 0.863, were yielded for the models of GPCRs-KR and Kinases-KR, respectively. In addition to AUC values, these two models also showed the best performance based on other evaluation indicators when k = 2 (Table 4). For instance, the recall values are 0.573 and 0.657 for the top 20 predictions (L = 20) of GPCRs-KR and Kinases-KR, respectively, suggesting a potential 60% success rate in the future experimental validation. Moreover, from the precision-recall curves (Figure 3), we found that the performance of KR and FP4 fingerprints was better than other fingerprints for the models of GPCRs, and KR fingerprint slightly outperformed other fingerprints for the models of Kinases. Compared with the precision-recall curves in the 10-fold cross validation (Supplementary Figure S1), the curves in external validation did not strictly follow the rule of anticorrelation. The possible reason might be that the number of interactions in the two external validation sets was small (Table 1). Nevertheless, GPCRs-KR and Kinases-KR further showed high predictive performance on two external validation sets.

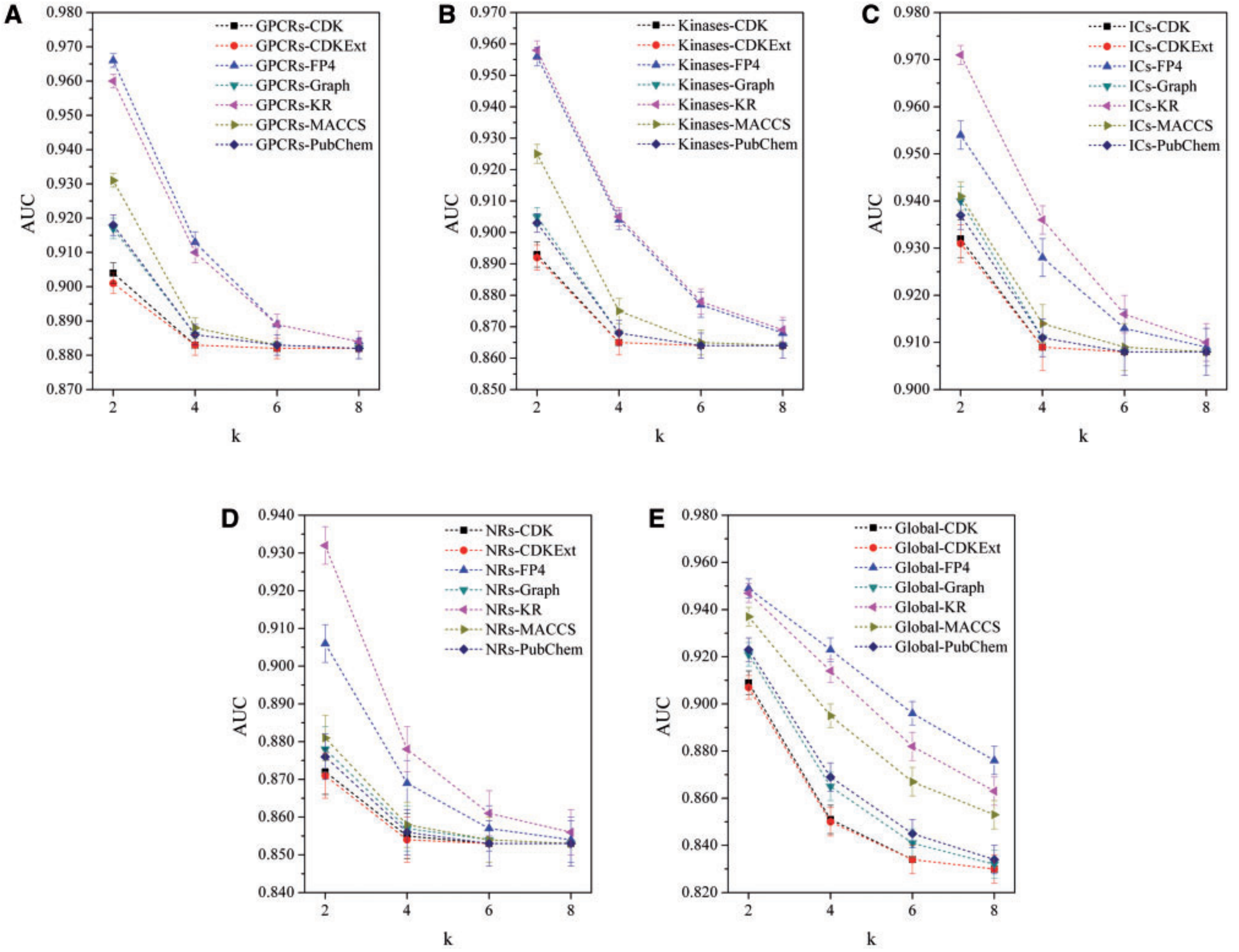


Figure 2. The AUC-k curves for the models of (**A**) GPCRs, (**B**) Kinases, (**C**) ICs, (**D**) NRs and (**E**) Global in the 10-fold cross validation.

Comparison with previous methods

Based on our knowledge, SDTNBI is the first network-based approach that can predict potential targets for new chemical entities on a large scale. Hence, we cannot directly compare the performance of SDTNBI with several previously published methods, such as NBI [15], DT-Hybrid [17], NRWRH [18] and ***CSNAP*** [34]. In a recent study, Cheng *et al.* [16] have predicted potential targets for new compounds by integrating the traditional network-based prediction and drug similarity information. The same data sets and external validation sets covering GPCRs and Kinases were used in SDTNBI and the previous study [16]. It is reasonable to compare the performance of SDTNBI with that drug similarity-based method. The drug similarity-based method yielded the highest AUC values, 0.769 and 0.828, for the data sets of GPCRs and Kinases, respectively [16]. Table 4 showed that SDTNBI has the highest AUC values, 0.797 and 0.863, for GPCRs and Kinases, respectively, outperforming the drug similarity-based method.

In addition, we compared SDTNBI with machine learning-based methods. In 2012, Cheng *et al.* [10] developed two machine learning-based methods, multi-target quantitative structure–activity relationship (mt-QSAR) and computational chemogenomics, for the prediction of chemical–protein interactions. Herein, an external validation set containing 173 interactions connecting 162 compounds and 15 targets were collected from the previous work to assess the generalization ability of our models. Fifteen targets in this data set were included in our DTI network for Kinases. In addition, 162 compounds in this data set were absent in our original DTI network for Kinases. Hence, it is reasonable to compare SDTNBI with the two previously published machine learning-based methods using this external validation set. The evaluation indicators and ROC curves in this external validation set were provided in Supplementary Table S5 and Supplementary Figure S2, respectively. The AUC values from 0.922 to 0.927 were yielded for the models of Kinases built via SDTNBI across seven different types of fingerprints (Supplementary Table S5), outperforming the AUC values, 0.865 and 0.604, for the models of kinases built via mt-QSAR and computational chemogenomics [10]. Collectively, aforementioned analyses suggested that SDTNBI outperformed the previously developed drug similarity-based method [16] and two machine learning-based methods (i.e. mt-QSAR and computational chemogenomics) [10].

Case studies: discovery of potential anticancer targets for NSAIDs

NSAIDs are a big class of drugs that can be used for antipyretic, analgesic and anti-inflammatory effects. Recently, NSAIDs have shown potential anticancer indications [35]. However, the exact molecular mechanisms of antitumor effects of NSAIDs were still

Table 2. The performance of models in the 10-fold cross validation when k = 2

Target	FP	P (L = 20)	R (L = 20)	e_P (L = 20)	e_R (L = 20)	AUC
GPCRs	CDK	0.051 ± 0.001	0.880 ± 0.007	4.16 ± 0.06	4.18 ± 0.06	0.904 ± 0.003
	CDKExt	0.051 ± 0.001	0.878 ± 0.007	4.15 ± 0.06	4.17 ± 0.06	0.901 ± 0.003
	FP4	0.057 ± 0.001	0.980 ± 0.003	4.65 ± 0.06	4.66 ± 0.06	0.966 ± 0.002
	Graph	0.052 ± 0.001	0.902 ± 0.007	4.27 ± 0.06	4.29 ± 0.06	0.917 ± 0.003
	KR	0.057 ± 0.001	0.976 ± 0.004	4.63 ± 0.06	4.64 ± 0.07	0.960 ± 0.002
	MACCS	0.055 ± 0.001	0.945 ± 0.006	4.47 ± 0.07	4.49 ± 0.07	0.931 ± 0.002
	PubChem	0.052 ± 0.001	0.903 ± 0.007	4.27 ± 0.06	4.29 ± 0.06	0.918 ± 0.003
Kinases	CDK	0.040 ± 0.001	0.642 ± 0.013	6.38 ± 0.13	6.53 ± 0.13	0.893 ± 0.004
	CDKExt	0.040 ± 0.001	0.636 ± 0.013	6.32 ± 0.13	6.47 ± 0.14	0.892 ± 0.004
	FP4	0.055 ± 0.001	0.868 ± 0.010	8.75 ± 0.12	8.83 ± 0.13	0.956 ± 0.003
	Graph	0.042 ± 0.001	0.672 ± 0.012	6.73 ± 0.13	6.83 ± 0.13	0.905 ± 0.003
	KR	0.056 ± 0.001	0.877 ± 0.009	8.82 ± 0.11	8.92 ± 0.11	0.958 ± 0.003
	MACCS	0.047 ± 0.001	0.737 ± 0.013	7.40 ± 0.14	7.50 ± 0.14	0.925 ± 0.003
	PubChem	0.042 ± 0.001	0.671 ± 0.012	6.72 ± 0.13	6.83 ± 0.13	0.903 ± 0.003
ICs	CDK	0.047 ± 0.001	0.921 ± 0.009	4.38 ± 0.07	4.39 ± 0.07	0.932 ± 0.004
	CDKExt	0.046 ± 0.001	0.916 ± 0.009	4.36 ± 0.07	4.37 ± 0.07	0.931 ± 0.004
	FP4	0.049 ± 0.000	0.959 ± 0.007	4.57 ± 0.07	4.58 ± 0.07	0.954 ± 0.003
	Graph	0.047 ± 0.001	0.934 ± 0.009	4.45 ± 0.07	4.46 ± 0.07	0.940 ± 0.003
	KR	0.050 ± 0.000	0.985 ± 0.004	4.70 ± 0.07	4.70 ± 0.07	0.971 ± 0.002
	MACCS	0.048 ± 0.000	0.940 ± 0.008	4.48 ± 0.07	4.48 ± 0.07	0.941 ± 0.003
	PubChem	0.047 ± 0.001	0.930 ± 0.009	4.43 ± 0.07	4.44 ± 0.07	0.937 ± 0.003
NRs	CDK	0.050 ± 0.001	0.972 ± 0.006	1.70 ± 0.01	1.70 ± 0.01	0.872 ± 0.006
	CDKExt	0.050 ± 0.001	0.971 ± 0.006	1.70 ± 0.01	1.70 ± 0.01	0.871 ± 0.006
	FP4	0.051 ± 0.001	0.983 ± 0.006	1.72 ± 0.01	1.72 ± 0.01	0.906 ± 0.005
	Graph	0.051 ± 0.001	0.981 ± 0.005	1.72 ± 0.01	1.72 ± 0.01	0.878 ± 0.006
	KR	0.051 ± 0.000	0.994 ± 0.003	1.74 ± 0.01	1.74 ± 0.01	0.932 ± 0.005
	MACCS	0.051 ± 0.001	0.977 ± 0.006	1.71 ± 0.01	1.71 ± 0.01	0.881 ± 0.006
	PubChem	0.050 ± 0.001	0.975 ± 0.006	1.71 ± 0.01	1.71 ± 0.01	0.876 ± 0.006
Global	CDK	0.041 ± 0.002	0.385 ± 0.018	20.53 ± 0.70	19.39 ± 0.88	0.909 ± 0.005
	CDKExt	0.041 ± 0.002	0.377 ± 0.017	20.25 ± 0.70	19.01 ± 0.85	0.907 ± 0.005
	FP4	0.061 ± 0.002	0.665 ± 0.018	30.11 ± 0.74	33.50 ± 0.90	0.949 ± 0.004
	Graph	0.046 ± 0.002	0.438 ± 0.019	22.71 ± 0.69	22.07 ± 0.96	0.921 ± 0.005
	KR	0.059 ± 0.002	0.640 ± 0.018	29.27 ± 0.71	32.26 ± 0.92	0.947 ± 0.004
	MACCS	0.053 ± 0.002	0.539 ± 0.019	26.33 ± 0.67	27.13 ± 0.94	0.937 ± 0.004
	PubChem	0.046 ± 0.002	0.444 ± 0.019	23.01 ± 0.70	22.39 ± 0.95	0.923 ± 0.005

Note. FP = the fingerprint type used in generating drug–substructure linkages; P = precision; R = recall; $e_{P\,=}$ precision enhancement; $e_{R\,=}$ recall enhancement.

Table 3. The performance of models in leave-one-out cross validation when k = 2

Target	FP	P (L = 20)	R (L = 20)	e_P (L = 20)	e_R (L = 20)	AUC
Global	CDK	0.058 ± 0.138	0.209 ± 0.339	10.78 ± 17.51	10.78 ± 17.51	0.839 ± 0.188
	CDKExt	0.056 ± 0.136	0.202 ± 0.332	10.41 ± 17.15	10.41 ± 17.15	0.836 ± 0.188
	FP4	0.072 ± 0.133	0.367 ± 0.414	18.94 ± 21.34	18.94 ± 21.34	0.891 ± 0.150
	Graph	0.065 ± 0.147	0.245 ± 0.368	12.64 ± 19.00	12.64 ± 19.00	0.853 ± 0.180
	KR	0.099 ± 0.167	0.493 ± 0.438	25.42 ± 22.60	25.42 ± 22.60	0.910 ± 0.150
	MACCS	0.068 ± 0.148	0.264 ± 0.377	13.60 ± 19.45	13.60 ± 19.45	0.865 ± 0.167
	PubChem	0.064 ± 0.143	0.247 ± 0.366	12.73 ± 18.90	12.73 ± 18.90	0.856 ± 0.176

Note. FP = the fingerprint type used in generating drug–substructure linkages; P = precision; R = recall; $e_{P\,=}$ precision enhancement; e_R = recall enhancement.

unknown. Herein, to illustrate the practical applications of SDTNBI, we investigated the potential molecular mechanisms of the antitumor effects of NSAIDs, by integrating the known and newly predicted DTIs for NSAIDs via SDTNBI and related gene–disease associations.

In this study, NSAIDs were collected from our previous study [36]. Then, potential new targets ranked in the top five with the highest predicted scores were yielded for each NSAID by using the global model Global-FP4 (k = 2), which showed the best performance in the 10-fold cross validation. The gene symbols of all known and predicted targets were extracted from the UniProt database [37]. The corresponding gene–disease associations were downloaded from the Comparative Toxicogenomics Database [38], and only those curated items covering cancer were retained. The cancers from those gene–cancer associations were subsequently filtered and merged by referring to the MeSH (http://www.nlm.nih.gov/mesh/) tree structure categories. Finally, 211 DTIs connecting 21 NSAIDs and 55 targets, and 100 gene–disease associations connecting 28 genes and 29 cancers were obtained (Figure 4). A drug–gene–disease subnetwork for these NSAIDs was then constructed using the CytoScape software (version 3.1.1) [39].

Table 4. The performance of models in external validation when k = 2

Target	FP	P (L = 20)	R (L = 20)	e_P (L = 20)	e_R (L = 20)	AUC
GPCRs	CDK	0.070	0.446	2.29	2.16	0.753
	CDKExt	0.068	0.437	2.24	2.12	0.751
	FP4	0.083	0.566	2.74	2.74	0.784
	Graph	0.073	0.477	2.42	2.31	0.761
	KR	0.090	0.573	2.95	2.78	0.797
	MACCS	0.073	0.478	2.42	2.32	0.758
	PubChem	0.074	0.478	2.45	2.32	0.759
Kinases	CDK	0.030	0.579	5.71	5.96	0.852
	CDKExt	0.030	0.579	5.71	5.96	0.852
	FP4	0.031	0.600	5.91	6.18	0.847
	Graph	0.030	0.587	5.81	6.04	0.852
	KR	0.034	0.657	6.53	6.77	0.863
	MACCS	0.031	0.595	5.91	6.13	0.852
	PubChem	0.031	0.599	5.86	6.17	0.852

Note. FP = the type of fingerprints used in generating drug–substructure associations; P = precision; R = recall; e_P = precision enhancement; e_R = recall enhancement.

As shown in Figure 4, cyclooxygenase-2 (COX-2) encoded by *PTGS2*, a well-known primary target for NSAIDs, has high degree in the drug–gene–disease network. Previous studies have suggested that COX-2 plays crucial roles in cancer [40], such as colorectal cancer [41–43], pancreatic cancer [44] and breast cancer [45]. In addition, inhibition of COX-2 by NSAIDs has potential anticancer indications for colorectal cancer [46] and breast cancer [47]. In this study, we chose several other potential targets that are less reported to have new anticancer indications to explore new potential mechanisms of the anticancer indications for NSAIDs via SDTNBI and network analyses. For instance, AKR1C3, a member of aldo/ketoreductase superfamily, was predicted as a novel potential antitumor target for several NSAIDs. Several previous studies reported that AKR1C3 plays crucial roles in various diseases, including prostate cancer [48–50]. Previous preclinical studies demonstrated that multiple NSAIDs, such as Diclofenac [51], Flurbiprofen [52], Ibuprofen [53], Indomethacin [54] and Meclofenamic acid [55], have the potential antiprostate cancer indications. Moreover, NSAIDs have been well characterized as potent AKR1C3 inhibitors based on previously pharmacological experiments [56, 57] and co-crystal structure data [58, 59]. Thus, the predicted interactions such as Diclofenac-AKR1C3 and Ibuprofen-AKR1C3 via SDTNBI were consistent with published evidence. We also found that carbonic anhydrases are potential anticancer targets for several NSAIDs. Several carbonic anhydrase isoforms, including CA9 [60, 61] and CA12 [62, 63], are associated with breast cancer. Previous studies have reported that NSAIDs, including Acetaminophen [64] and Celecoxib [47], have potential antibreast cancer effects. Biological assays have showed that NSAIDs, including Acetaminophen, Celecoxib and Valdecoxib, are potent CA9 or CA12 inhibitors [65–67]. Collectively, inhibiting AKR1C3, CA9 or CA12 by NSAIDs may provide new strategy for cancer chemoprevention.

In addition, we also computationally identified new potential anticancer indications for several NSAIDs, such as Carprofen, Etodolac and Rofecoxib, via SDTNBI. Specifically, we computationally identified that CDK2 may be targeted by Carprofen, Etodolac and Rofecoxib (Figure 4). Previous studies demonstrated that CDK2 plays crucial roles mediating tumorigenesis and tumor progression in several cancer types [68], such as melanoma [69] and prostate cancer [70]. For example, a recent study showed that Carprofen induces cell apoptosis by targeting the p38 MAPK pathway in prostate cancer cells [71]. Cheng *et al.* [72] found that Etodolac inhibited CDK2, CDK4 and CDC2 expression and further repressed tumor growth in human hepatocellular carcinoma cell lines. Tanaka *et al.* [73] found that oral doses of Rofecoxib showed a potential value for the treatment of non-small cell lung cancer. In summary, the potential anticancer indications for Carprofen, Etodolac and Rofecoxib predicted via SDTNBI are consistent with the previous literature evidences. Further study will be needed to provide experimental validations, which we hope will be prompted by the findings herein. Finally, all newly predicted targets ranked in the top 20 for 1844 drugs via the aforementioned best global model are available on request for experimental investigation in the future.

Tool development

We developed a toolkit called NetInfer via C ++for predicting new potential targets for known drugs, failed drugs and new chemical entities on a large scale. Both SDTNBI and previously developed NBI were implemented in this toolkit. NetInfer is light weight and does not need the support of any third-party math libraries such as linear algebra libraries. To accelerate the calculation and decrease the cost of memory space, different data structures were designed for the sparse and dense matrices. It provides functions of prediction, cross validation and external validation. In a uniform platform, researchers can input their in-house data into our toolkit and then obtain predictive lists or evaluation indicators.

Input file format

To input their own data into NetInfer for DTI prediction, users should convert the data into a specific format. Each network stores in a text file, which contains numerous lines representing edges in the network and five columns separated by 'Tab' character. For a edge between node A and node B, the five columns denote (i) the type of node A, (ii) the identifier of node A, (iii) the type of node B, (iv) the identifier of node B and (v) the weighted value of edge A-B, respectively. Currently, NetInfer only supports unweighted networks so that the weighted values are always '1'.

To prioritize targets for known drugs, users have to prepare an input file of known DTI network. For example, the raw data can be downloaded from relative databases such as DrugBank [27], BindingDB [28] and ChEMBL [26]. Subsequently, users convert the data into cleaned DTI pairs without duplicates, and then save them into a text file with aforementioned format. The detailed protocols are provided in Figure 5. Each line represents a DTI in this text file, where five columns denote (i) the type of drug node, (ii) the drug identifier, (iii) the type of target node, (iv) the target identifier and (v) the weighted value of edge. We recommended users to use two simple strings 'DRUG' and 'TARGET' as the type of drug node and target node, respectively. In current version, NetInfer supports different types of drug identifiers. For example, the drug identifiers can be predefined by existing databases, such as DrugBank ID, ChEMBL ID and PubChem Compound ID, or defined by users. In this study, to avoid the duplicates of drug molecules from different data sources, we used a unique in-house hash code generated from canonical SMILES via a hashing algorithm as the drug identifier. Similarly, different types of target identifiers, such as UniProt accession numbers or gene symbols, are supported by NetInfer.

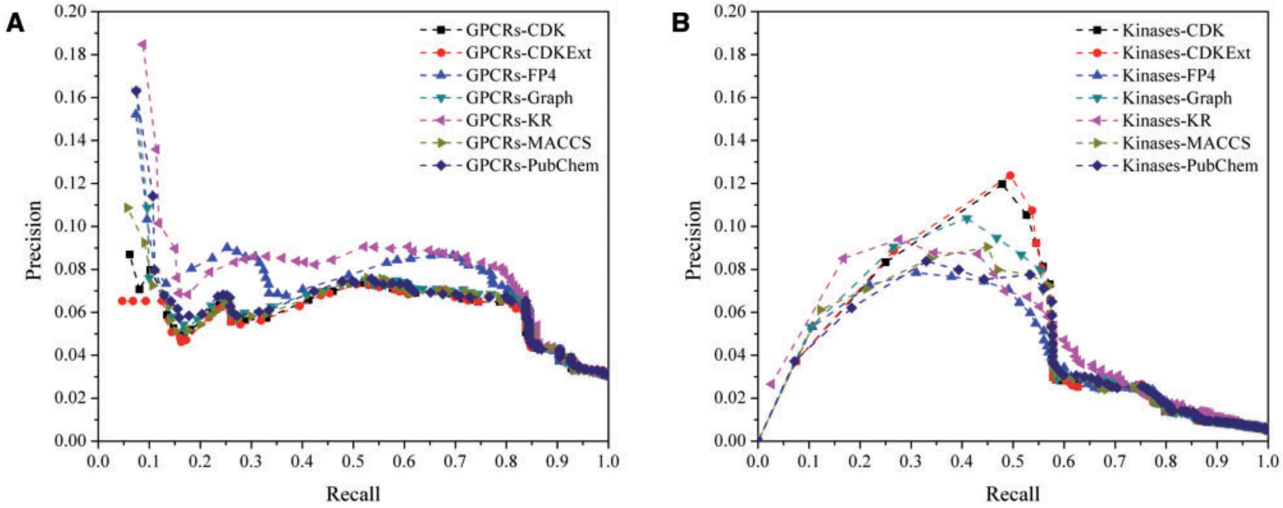


Figure 3. The precision-recall curves of (**A**) GPCRs and (**B**) Kinases in external validation when k = 2.

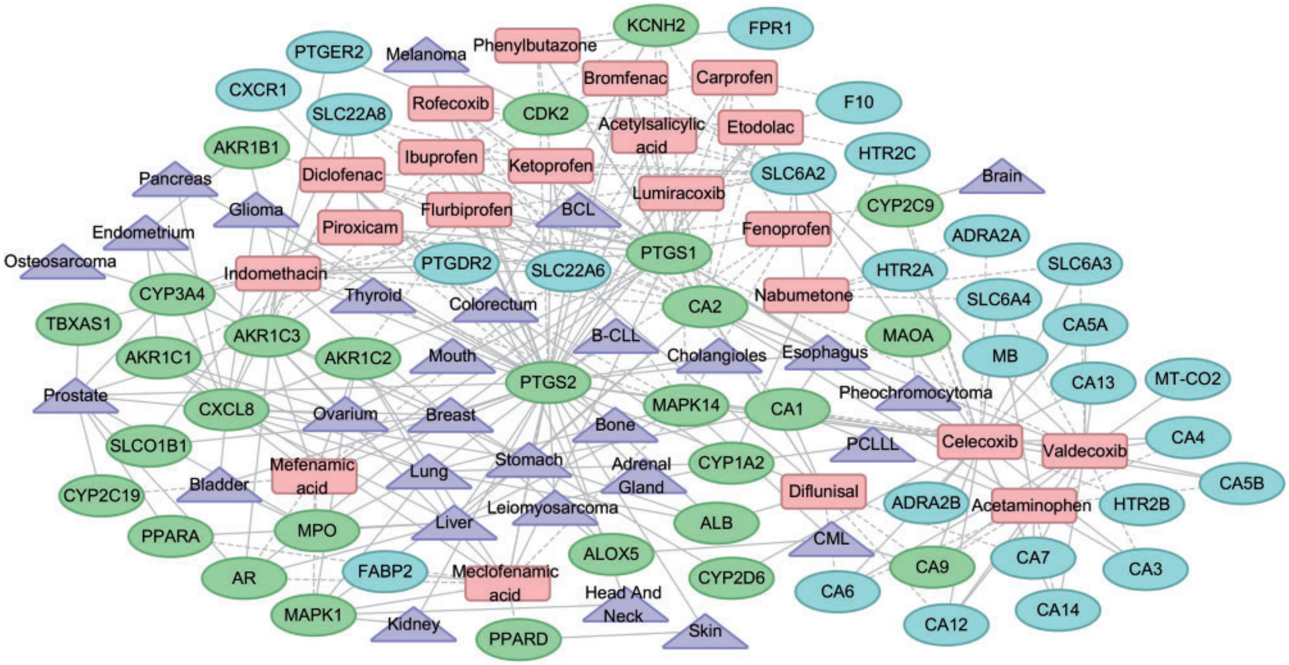


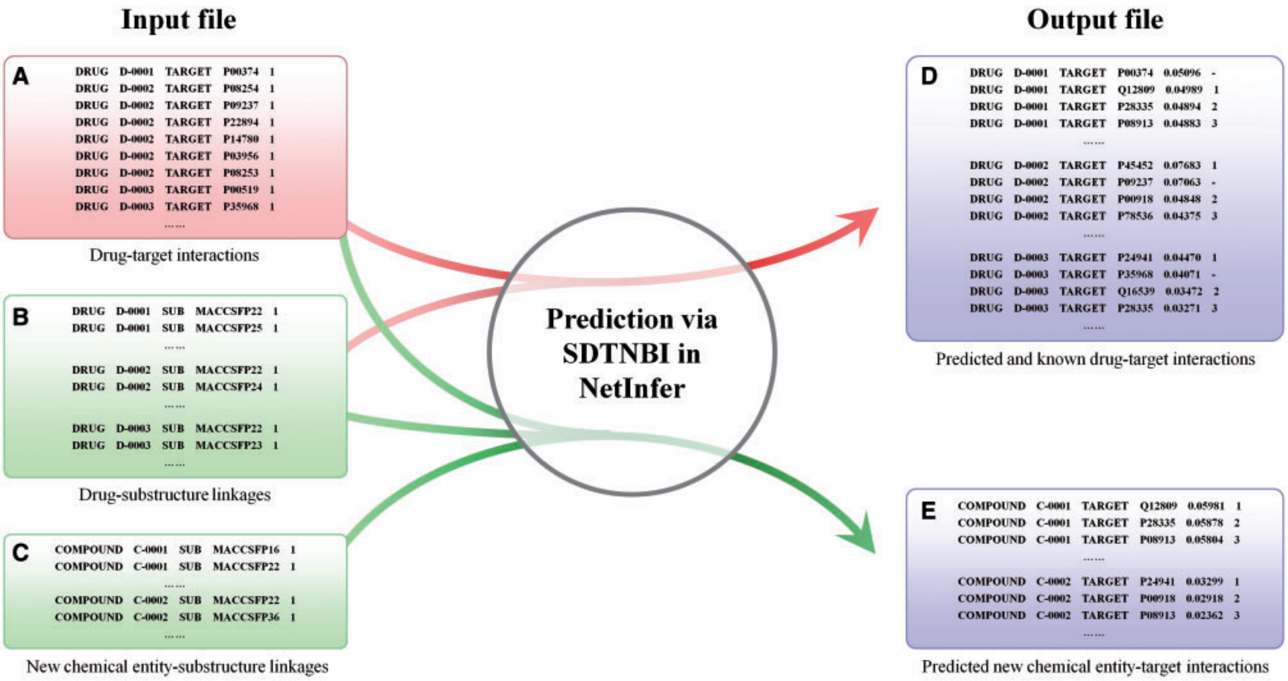
Figure 4. The drug–gene–disease network of NSAIDs. Red rectangle: NSAID, green circle: gene associated with cancer, blue circle: gene unassociated with cancer, purple triangle: cancer, solid line: known drug–gene or gene–disease association, dash line: predicted drug–gene association.

For instance, as shown in Figure 5A, the first line means that there is a drug 'D-0001' that can act on a target 'P00374' in a DTI network, and the lines from 2 to 7 means that the drug 'D-0002' has known interactions with six targets, including 'P08254', 'P09237', etc.

In addition to a known DTI network, an input file of known drug–substructure network should also be prepared. As shown in Figure 5B, each line represents a drug–substructure linkage in this text file, where five columns denote (i) the type of drug node (e.g. 'DRUG'), (ii) the drug identifier, (iii) the type of substructure node (e.g. 'SUB'), (iv) the substructure identifier and (v) the weighted value of edge. We recommended users to use 'DRUG' and 'SUB' as the type of drug node and substructure node, respectively. The string used as the type of drug node as well as drug identifiers should be consistent with the former input file. SDTNBI in NetInfer will use these two input files to construct a substructure–drug–target network (Figure 1A), and then predict potential targets for known drugs.

Preparing input files for new chemical entities

To predict potential targets for new chemical entities, users have to prepare another text file of the new chemical entity–substructure network, in addition to the input files of DTI network and drug–substructure network. In this input file, each line represents a new chemical entity–substructure linkage, where five columns denote (i) the type of new chemical entity node, (ii) the identifier of new chemical entity, (iii) the type of substructure node, (iv) the substructure identifier and (v) the weighted value of edge. We recommended users to use 'COMPOUND' and 'SUB' as the type of new chemical entity node



```
DRUG   D-0001   TARGET   P00374   1
DRUG   D-0002   TARGET   P08254   1
DRUG   D-0002   TARGET   P09237   1
DRUG   D-0002   TARGET   P22894   1
DRUG   D-0002   TARGET   P14780   1
DRUG   D-0002   TARGET   P03956   1
DRUG   D-0002   TARGET   P08253   1
DRUG   D-0003   TARGET   P00519   1
DRUG   D-0003   TARGET   P35968   1
……
```

Drug-target interactions

```
DRUG   D-0001   SUB   MACCSFP22   1
DRUG   D-0001   SUB   MACCSFP25   1
……
DRUG   D-0002   SUB   MACCSFP22   1
DRUG   D-0002   SUB   MACCSFP24   1
……
DRUG   D-0003   SUB   MACCSFP22   1
DRUG   D-0003   SUB   MACCSFP23   1
……
```

Drug-substructure linkages

```
COMPOUND   C-0001   SUB   MACCSFP16   1
COMPOUND   C-0001   SUB   MACCSFP22   1
……
COMPOUND   C-0002   SUB   MACCSFP22   1
COMPOUND   C-0002   SUB   MACCSFP36   1
……
```

New chemical entity-substructure linkages

```
DRUG   D-0001   TARGET   P00374   0.05096   -
DRUG   D-0001   TARGET   Q12809   0.04989   1
DRUG   D-0001   TARGET   P28335   0.04894   2
DRUG   D-0001   TARGET   P08913   0.04883   3
……
DRUG   D-0002   TARGET   P45452   0.07683   1
DRUG   D-0002   TARGET   P09237   0.07063   -
DRUG   D-0002   TARGET   P00918   0.04848   2
DRUG   D-0002   TARGET   P78536   0.04375   3
……
DRUG   D-0003   TARGET   P24941   0.04470   1
DRUG   D-0003   TARGET   P35968   0.04071   -
DRUG   D-0003   TARGET   Q16539   0.03472   2
DRUG   D-0003   TARGET   P28335   0.03271   3
……
```

Predicted and known drug-target interactions

```
COMPOUND   C-0001   TARGET   Q12809   0.05981   1
COMPOUND   C-0001   TARGET   P28335   0.05878   2
COMPOUND   C-0001   TARGET   P08913   0.05804   3
……
COMPOUND   C-0002   TARGET   P24941   0.03299   1
COMPOUND   C-0002   TARGET   P00918   0.02918   2
COMPOUND   C-0002   TARGET   P08913   0.02362   3
……
```

Predicted new chemical entity-target interactions

Figure 5. The input file format and output file format of SDTNBI in NetInfer. (**A**) An example input file of known DTIs, (**B**) an example input file of drug–substructure linkages, (**C**) an example input file of new chemical entity–substructure linkages, (**D**) an example output file of predicted and known DTIs, (**E**) an example output file of predicted new chemical entity–target interactions, red arrow: the process of predicting potential targets for known drugs with two input files, green arrow: the process of predicting potential targets for new chemical entities with three input files.

and substructure node, respectively. The string used as the type of substructure node should be consistent with the prepared input file of drug–substructure network. Moreover, the fingerprint type for generating new chemical entity–substructure network must be same as which was used in drug–substructure network. For instance, as shown in Figure 5C, the input file shows that the new chemical entity 'C-0001' harbors several substructures, such as 'MACCSFP16', 'MACCSFP22', etc.

The input file of the new chemical entity–substructure network can be prepared as following steps. First, two columns separated by 'Tab' character, respectively, containing the SMILES strings and identifiers of new chemical entities to be predicted, are saved in a SMILES file. Subsequently, users can generate substructures for these molecules with their selected fingerprints. For instance, this SMILES file can be converted into a comma separated value (CSV) format file with compound–substructure information by the PaDEL-Descriptor software [30]. And then, an in-house Python script can be used to convert the CSV file into a text file that can be inputted into NetInfer. An example describes as below:

```
import csv
f = open(’OUTPUT.txt’, ’w’)
for i, x in enumerate(csv.reader(open(’INPUT.csv’, ’r’))):
  if i == 0:
    y = x
  else:
    for j in range(1, len(x)):
       if x[j] == ’1’:
          f.write(’\t’.join([’COMPOUND’,   x[0],   ’SUB’,   y[j],
’1’]) + ’\n’)
f.close()
```

where 'INPUT.csv' is the CSV format file to be converted, and 'OUTPUT.txt' is the text file to be generated. Users can replace the filename according to practical situation. It is noteworthy that this script can also be used to prepare the aforementioned input file of drug–substructure network if replacing the 'COMPOUND' by 'DRUG'. Finally, SDTNBI in NetInfer will use three input files, namely this input file of new chemical entity–substructure network and aforementioned two input files of known DTIs and drug–substructure network, to construct a substructure–drug (and new chemical entity)–target network (Figure 1A), and then predict potential targets for new chemical entities.

Target prediction via SDTNBI

After preparing input files, users can execute SDTNBI in NetInfer to prioritize potential targets for known drugs, failed drugs or new chemical entities via command line. In the command line, the number of resource-spreading processes (k) and the number of targets to be predicted for each drug or compound are given by users. Furthermore, according to actual projects, users can provide two input files for prioritizing targets for known drugs, or three input files for predicting targets for failed drugs or new chemical entities. The detailed command lines for prediction, cross validation and external validation were all presented at our Web site (http://lmmd.ecust.edu.cn/methods/sdtnbi/).

Output file format

The predicted results of NetInfer are saved in a text file. Each line also represents a DTI in this output file, where six columns separated by 'Tab' character denote (i) the type of node A, (ii) the identifier of node A, (iii) the type of node B, (iv) the identifier of node B, (v) the score of A-B and (vi) the ranking position of B at the sorted list of all predicted objects of A. It is noteworthy that known DTIs will also be outputted into this file if they exist, whereas their ranking position will be represented as character '-', instead of an integer number. For instance, as shown in Figure 5D, the known drug 'D-0001' with a known interaction with 'P00364' was predicted to have potential targets 'Q12809', 'P28335', etc. Similarly, as shown in Figure 5E, the new chemical entity 'C-0002' without known targets was predicted to have potential targets 'P24941', 'P00918', etc. Users can use these known and predicted DTIs to build networks for analysis, or choose newly predicted interactions with high scores for further experimental validation.

Discussion

In this study, we developed a novel integrated network and chemoinformatics tool, named SDTNBI, to predict potential targets not only for old drugs but also for failed drugs and new chemical entities on a large scale. Although several existing network-based prediction methods such as NBI [15], EWNBI and NWNBI [16] have achieved high accuracy and robustness in DTI prediction, they could not prioritize potential targets for new chemical entities without known targets in DTI network, such as newly synthesized structures and failed drugs in phases II and III. To overcome this defect, SDTNBI imports chemical substructures to link those new chemical entities with known DTI network. It can predict hundreds of potential targets for thousands of drugs or new chemical entities at the same time by systematically searching a global substructure–drug–target tripartite network. Based on systematic evaluation, we found that SDTNBI has the best performance when the number of resource-spreading processes (k) is 2, and the performance decreased with the increasing of k. The possible reason of this phenomenon might be that the resources will be located more dispersedly in the network as the increasing of k. This dispersion might lead to worse method performance. Meanwhile, seven commonly used fingerprints, which are all freely available, were systematically tested. Under the optimal condition, SDTNBI yields high performance in the 10-fold cross validation, leave-one-out cross validation and external validation. Furthermore, we developed a useful toolkit and predicted thousands of new potential DTIs for Food and Drug Administration-approved or clinical investigational drugs, which are available for future validation for academic users such as experimental scientists. For example, case studies for identifying anticancer mechanisms of action of NSAIDs have shown practical applications of SDTNBI. Hence, SDTNBI would provide a highly efficient and low-cost tool for DTI prediction and drug repositioning in drug discovery.

There are several significant contributions and advantages of SDTNBI compared with previously published methods, such as network-based methods [15, 16], machine learning-based methods [10–12] and molecular docking [8, 9, 36]. First, the most key contribution of SDTNBI is that it can be used to predict potential targets for new chemical entities, whereas traditional network-based methods cannot. Machine learning-based methods and models with high accuracy have been developed for predicting targets to new chemical entities. For example, Yamanishi *et al.* [11] built bipartite graph learning models by integrating chemical protein sequence similarity information, and then further built the supervised bipartite graph learning model by integrating chemical similarity, protein sequence similarity information and pharmacological data [12]. However, traditional machine learning-based models have several potential pitfalls. In generally, most machine learning-based models are built via constructing positive against negative samples. High quality and large-scale diverse molecules with wide target coverage and gold negative data sets are crucial for machine learning-based model development [7]. However, it is always difficult to construct a negative data set with enough size and gold standard because there are few experimentally validated inactive data published in literatures. Most negative data sets in the machine learning-based models are randomly constructed based on the principle of 'one-versus-the-rest', such as our previously developed mt-QSAR and computational chemogenomics methods [10], resulting in negative samples much more than positive samples. The accuracy of models was often reduced with these low-quality negative samples, even though they might be selected to keep the balance between positive and negative samples. In this study, SDTNBI inherited the advantages of both network-based and chemoinformatics-based approaches. It builds network models based on only positive DTIs to predict potential targets for both known drugs and new chemical entities on a large scale. As described in the method comparison, SDTNBI outperformed the previously developed drug similarity-based method [16], mt-QSAR and computational chemogenomics methods [10]. Moreover, like other network-based methods, SDTNBI does not rely on three-dimensional structures of protein targets and small molecules. Hence, it can predict more varieties of targets for user-given molecules compared with molecular docking approaches, such as TarFisDock [8] and ***DRAR-CPI*** [9]. Finally, SDTNBI can also be used in other networks, such as drug–disease network [6], drug–side effect network [74], drug–microRNA network [5] and drug–single-nucleotide polymorphisms network by systematically incorporating multiple-scale biomedical data, such as drug–gene signatures from LINCSCLOUD (http://www.lincscloud.org), structured data generated from electronic health records [75] and drug pharmacogenomics data from PharmGKB [76].

However, there are still several possible limitations of SDTNBI. First, it cannot predict potential DTIs for targets absent in the existing DTI network because of no reachable paths among those targets and the known network. Second, if there is a molecule with a special chemical structure, which shared none or only a few substructures with known drugs in the DTI network, SDTNBI cannot accurately predict targets for it. Third, for each drug in the substructure–drug–target tripartite network, it has two types of neighbor nodes standing for substructures and targets, respectively. But we did not differentiate them in resource-spreading processes. Equal amount of initial resources was allocated to each of its neighbor nodes, no matter it being a substructure or a target. Meanwhile, equal weighted values were set for each of its edges, no matter it being a drug–substructure linkage or a DTI. This equal treatment might cause potential unbalances and then decrease the performance of SDTNBI. Moreover, the influence of hub nodes was not investigated yet. We are actively developing improved methods to solve these potential limitations, such as node-weighted or edge-weighted network-based approaches. Nevertheless, SDTNBI would provide powerful tools for DTI prediction and drug repositioning for drug discovery and personalized medicine in the future.

Key Points

- We proposed a novel tool, namely substructure–drug–target network-based inference (SDTNBI), to prioritize potential targets for old drugs, failed drugs and new chemical entities on a large scale.
- SDTNBI shows high performance on both cross validation and external validation, which demonstrates its potential applications in drug discovery.
- SDTNBI was implemented in a useful toolkit and hence would be used for large-scale DTI prediction and drug repositioning.

Funding

This work was supported by the National Natural Science Foundation of China (grant no. 81373328, 81373329 and 81573020), the 863 Project (grant no. 2012AA020308), the Fundamental Research Funds for the Central Universities (grant no. WY1113007) and the 111 Project (grant no. B07023).

References

1. Sams-Dodd F. Target-based drug discovery: is something wrong? *Drug Discov Today* 2005;**10**:139–47.
2. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 2008;**4**:682–90.
3. Cheng FX, Li WH, Wu ZR, *et al*. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J Chem Inf Model* 2013;**53**:753–62.
4. Cheng FX, Li WH, Zhou YD, *et al*. Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (PTDMs). *Mol Biosyst* 2013;**9**:1316–25.
5. Li J, Wu ZR, Cheng FX, *et al*. Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology. *Sci Rep* 2014;**4**:5576.
6. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**:303–11.
7. Ding H, Takigawa I, Mamitsuka H, *et al*. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform* 2014;**15**:734–47.
8. Li HL, Gao ZT, Kang L, *et al*. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* 2006;**34**:W219–24.
9. Luo H, Chen J, Shi LM, *et al*. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. *Nucleic Acids Res* 2011;**39**:W492–8.
10. Cheng FX, Zhou YD, Li J, *et al*. Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Mol Biosyst* 2012;**8**:2373–84.
11. 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.
12. Yamanishi Y, Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**:i246–54.
13. Keiser MJ, Roth BL, Armbruster BN, *et al*. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**:197–206.
14. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature* 2009;**462**:175–81.
15. Cheng FX, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comp Biol* 2012;**8**:e1002503.
16. Cheng FX, Zhou YD, Li WH, *et al*. Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One* 2012;**7**:e41064.
17. Alaimo S, Pulvirenti A, Giugno R, *et al*. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 2013;**29**:2004–8.
18. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;**8**:1970–8.
19. Lucas X, Gruning BA, Bleher S, *et al*. The purchasable chemical space: a detailed picture. *J Chem Inf Model* 2015;**55**:915–24.
20. Overington JP, Al-Lazikani B, Hopkins AL. Opinion - how many drug targets are there? *Nat Rev Drug Discov* 2006;**5**:993–6.
21. Hay M, Thomas DW, Craighead JL, *et al*. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014;**32**:40–51.
22. Mullard A. Drug repurposing programmes get lift off. *Nat Rev Drug Discov* 2012;**11**:1–2.
23. Shen J, Cheng FX, Xu Y, *et al*. Estimation of ADME properties with substructure pattern recognition. *J Chem Inf Model* 2010;**50**:1034–41.
24. Yamanishi Y, Pauwels E, Saigo H, *et al*. Extracting sets of chemical substructures and protein domains governing drug-target interactions. *J Chem Inf Model* 2011;**51**:1183–94.
25. Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics* 2008;**24**:2518–25.
26. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
27. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res* 2011;**39**:D1035–41.
28. Liu TQ, Lin YM, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.
29. O'Boyle NM, Banck M, James CA, *et al*. Open Babel: an open chemical toolbox. *J Cheminform* 2011;**3**:14.
30. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**:1466–74.
31. Cheng FX, Ikenaga Y, Zhou YD, *et al*. In silico assessment of chemical biodegradability. *J Chem Inf Model* 2012;**52**:655–69.
32. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
33. Yera ER, Cleves AE, Jain AN. Chemical structural novelty: on-targets and off-targets. *J Med Chem* 2011;**54**:6771–85.
34. Lo YC, Senese S, Li CM, *et al*. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comp Biol* 2015;**11**:e1004153.
35. Nan HM, Hutter CM, Lin Y, *et al*. Association of aspirin and NSAID use with risk of colorectal cancer according to genetic variants. *JAMA* 2015;**313**:1133–42.
36. Lu WQ, Cheng FX, Jiang J, *et al*. FXR antagonism of NSAIDs contributes to drug-induced liver injury identified by systems pharmacology approach. *Sci Rep* 2015;**5**:8114.

37. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:D115–19.
38. Davis AP, Murphy CG, Saraceni-Richards CA, *et al*. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res* 2009;**37**:D786–92.
39. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
40. Subbaramaiah K, Dannenberg AJ. Cyclooxygenase 2: a molecular target for cancer prevention and treatment. *Trends Pharmacol Sci* 2003;**24**:96–102.
41. Sano H, Kawahito Y, Wilder RL, *et al*. Expression of cyclooxygenase-1 and -2 in human colorectal cancer. *Cancer Res* 1995;**55**:3785–9.
42. Tsujii M, Kawano S, Dubois RN. Cyclooxygenase-2 expression in human colon cancer cells increases metastatic potential. *Proc Natl Acad Sci USA* 1997;**94**:3336–40.
43. Tsujii M, Kawano S, Tsuji S, *et al*. Cyclooxygenase regulates angiogenesis induced by colon cancer cells. *Cell* 1998;**93**: 705–16.
44. Tucker ON, Dannenberg AJ, Yang FK, *et al*. Cyclooxygenase-2 expression is up-regulated in human pancreatic cancer. *Cancer Res* 1999;**59**:987–90.
45. Hwang D, Scollard D, Byrne J, *et al*. Expression of cyclooxygenase-1 and cyclooxygenase-2 in human breast cancer. *J Natl Cancer Inst* 1998;**90**:455–60.
46. Gupta RA, DuBois RN. Colorectal cancer prevention and treatment by inhibition of cyclooxygenase-2. *Nat Rev Cancer* 2001;**1**:11–21.
47. Harris RE, Alshafie GA, Abou-Issa H, *et al*. Chemoprevention of breast cancer in rats by celecoxib, a cyclooxygenase 2 inhibitor. *Cancer Res* 2000;**60**:2101–3.
48. Dozmorov MG, Azzarello JT, Wren JD, *et al*. Elevated AKR1C3 expression promotes prostate cancer cell survival and prostate cell-mediated endothelial cell tube formation: implications for prostate cancer progressioan. *BMC Cancer* 2010;**10**:16.
49. Liu CF, Lou W, Zhu YZ, *et al*. Intracrine androgens and AKR1C3 activation confer resistance to enzalutamide in prostate cancer. *Cancer Res* 2015;**75**:1413–22.
50. Yepuru M, Wu ZZ, Kulkarni A, *et al*. Steroidogenic enzyme AKR1C3 is a novel androgen receptor-selective coactivator that promotes prostate cancer growth. *Clin Cancer Res* 2013;**19**:5613–25.
51. Inoue T, Anai S, Onishi S, *et al*. Inhibition of COX-2 expression by topical diclofenac enhanced radiation sensitivity via enhancement of TRAIL in human prostate adenocarcinoma xenograft model. *BMC Urol* 2013;**13**:9
52. Wechter WJ, Leipold DD, Murray ED, *et al*. E-7869 (R-flurbiprofen) inhibits progression of prostate cancer in the TRAMP mouse. *Cancer Res* 2000;**60**:2203–8.
53. John-Aryankalayil M, Palayoor ST, Cerna D, *et al*. NS-398, ibuprofen, and cyclooxygenase-2 RNA interference produce significantly different gene expression profiles in prostate cancer cells. *Mol Cancer Ther* 2009;**8**:261–73.
54. Liedtke AJ, Adeniji AO, Chen M, *et al*. Development of potent and selective indomethacin analogues for the inhibition of AKR1C3 (type 5 17 beta-hydroxysteroid dehydrogenase/prostaglandin F synthase) in castrate-resistant prostate cancer. *J Med Chem* 2013;**56**:2429–46.
55. Soriano-Hernandez AD, Galvan-Salazar HR, Montes-Galindo DA, *et al*. Antitumor effect of meclofenamic acid on human androgen-independent prostate cancer: a preclinical evaluation. *Int Urol Nephrol* 2012;**44**:471–7.
56. Byrns MC, Steckelbroeck S, Penning TM. An indomethacin analogue, N-(4-chlorobenzoyl)-melatonin, is a selective inhibitor of aldo-keto reductase 1C3 (type 2 3 alpha-HSD, type 5 17 beta-HSD, and prostaglandin F synthase), a potential target for the treatment of hormone dependent and hormone independent malignancies. *Biochem Pharmacol* 2008;**75**:484–93.
57. Gobec S, Brozic P, Rizner TL. Nonsteroidal anti-inflammatory drugs and their analogues as inhibitors of aldo-keto reductase AKR1C3: new lead compounds for the development of anticancer agents. *Bioorg Med Chem Lett* 2005;**15**:5170–5.
58. Flanagan JU, Yosaatmadja Y, Teague RM, *et al*. Crystal structures of three classes of non-steroidal anti-inflammatory drugs in complex with aldo-keto reductase 1C3. *PLoS One* 2012;**7**:16.
59. Lovering AL, Ride JP, Bunce CM, *et al*. Crystal structures of prostaglandin D-2 11-ketoreductase (AKR1C3) in complex with the nonsteroidal anti-inflammatory drugs flufenamic acid and indomethacin. *Cancer Res* 2004;**64**:1802–10.
60. Hussain SA, Ganesan R, Reynolds G, *et al*. Hypoxia-regulated carbonic anhydrase IX expression is associated with poor survival in patients with invasive breast cancer. *Br J Cancer* 2007;**96**:104–9.
61. Lou YM, McDonald PC, Oloumi A, *et al*. Targeting tumor hypoxia: suppression of breast tumor growth and metastasis by novel carbonic anhydrase IX inhibitors. *Cancer Res* 2011;**71**:3364–76.
62. Watson PH, Chia SK, Wykoff CC, *et al*. Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma. *Br J Cancer* 2003;**88**:1065–70.
63. Barnett DH, Sheng S, Charn TH, *et al*. Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. *Cancer Res* 2008;**68**:3505–15.
64. Takehara M, Hoshino T, Namba T, *et al*. Acetaminophen-induced differentiation of human breast cancer stem cells and inhibition of tumor xenograft growth in mice. *Biochem Pharmacol* 2011;**81**:1124–35.
65. Weber A, Casini A, Heine A, *et al*. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J Med Chem* 2004;**47**:550–7.
66. Di Fiore A, Pedone C, D'Ambrosio K, *et al*. Carbonic anhydrase inhibitors: valdecoxib binds to a different active site region of the human isoform II as compared to the structurally related cyclooxygenase II 'selective' inhibitor celecoxib. *Bioorg Med Chem Lett* 2006;**16**:437–42.
67. Innocenti A, Vullo D, Scozzafava A, *et al*. Carbonic anhydrase inhibitors: inhibition of mammalian isoforms I-XIV with a series of substituted phenols including paracetamol and salicylic acid. *Biorg Med Chem* 2008;**16**:7424–8.
68. Shapiro GI. Cyclin-dependent kinase pathways as targets for cancer treatment. *J Clin Oncol* 2006;**24**:1770–83.
69. Du JY, Widlund HR, Horstmann MA, *et al*. Critical role of CDK2 for melanoma growth linked to its melanocyte-specific transcriptional regulation by MITF. *Cancer Cell* 2004;**6**:565–76.
70. Lee E, Son JE, Byun S, *et al*. CDK2 and mTOR are direct molecular targets of isoangustone A in the suppression of human prostate cancer cell growth. *Toxicol Appl Pharmacol* 2013;**272**:12–20.
71. Khwaia FS, Quann EJ, Pattabiraman N, *et al*. Carprofen induction of p75(NTR)-dependent apoptosis via the p38 mitogen-activated protein kinase pathway in prostate cancer cells. *Mol Cancer Ther* 2008;**7**:3539–45.

72. Cheng JD, Imanishi H, Liu WD, *et al*. Involvement of cell cycle regulatory proteins and MAP kinase signaling pathway in growth inhibition and cell cycle arrest by a selective cyclooxygenase 2 inhibitor, etodolac, in human hepatocellular carcinoma cell lines. *Cancer Sci* 2004;**95**:666–73.
73. Tanaka T, Delong PA, Amin K, *et al*. Treatment of lung cancer using clinically relevant oral doses of the cyclooxygenase-2 inhibitor rofecoxib - potential value as adjuvant therapy after surgery. *Ann Surg* 2005;**241**:168–78.
74. Cheng FX, Li WH, Wang XC, *et al*. Adverse drug events: database construction and in silico prediction. *J Chem Inf Model* 2013;**53**:744–52.
75. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. *Nat Biotechnol* 2015;**33**:360–3.
76. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**:163–5.