

TargetNet: a web service for predicting potential drug–target interaction profiling via multi-target SAR models

Zhi-Jiang Yao^{1,2} · Jie Dong¹ · Yu-Jing Che³ · Min-Feng Zhu³ · Ming Wen² · Ning-Ning Wang¹ · Shan Wang² · Ai-Ping Lu⁴ · Dong-Sheng Cao^{1,4}

Received: 19 January 2016 / Accepted: 6 May 2016 / Published online: 11 May 2016
© Springer International Publishing Switzerland 2016

Abstract Drug–target interactions (DTIs) are central to current drug discovery processes and public health fields. Analyzing the DTI profiling of the drugs helps to infer drug indications, adverse drug reactions, drug–drug interactions, and drug mode of actions. Therefore, it is of high importance to reliably and fast predict DTI profiling of the drugs on a genome-scale level. Here, we develop the TargetNet server, which can make real-time DTI predictions based only on molecular structures, following the spirit of multi-target SAR methodology. Naïve Bayes models together with various molecular fingerprints were employed to construct prediction models. Ensemble learning from these fingerprints was also provided to improve the prediction ability. When the user submits a molecule, the server will predict the activity of the user’s molecule across 623

human proteins by the established high quality SAR model, thus generating a DTI profiling that can be used as a feature vector of chemicals for wide applications. The 623 SAR models related to 623 human proteins were strictly evaluated and validated by several model validation strategies, resulting in the AUC scores of 75–100 %. We applied the generated DTI profiling to successfully predict potential targets, toxicity classification, drug–drug interactions, and drug mode of action, which sufficiently demonstrated the wide application value of the potential DTI profiling. The TargetNet webserver is designed based on the Django framework in Python, and is freely accessible at <http://targetnet.scbdd.com>.

Keywords Web server · SAR models · Drug–target interaction · Multi-target SAR · Naïve Bayes

Zhi-Jiang Yao, Jie Dong and Yu-Jing Che have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9915-2) contains supplementary material, which is available to authorized users.

✉ Dong-Sheng Cao
oriental-cds@163.com

¹ School of Pharmaceutical Sciences, Central South University, Changsha 410013, People’s Republic of China

² College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, People’s Republic of China

³ School of Mathematics and Statistics, Central South University, Changsha 410083, People’s Republic of China

⁴ Institute of Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, SAR, People’s Republic of China

Introduction

Drug–target interactions (DTIs) are central to current drug discovery processes and public health fields [1, 2]. In drug discovery process, one of the challenges is to identify the potential targets for drug-like compounds. Once the target is successfully identified, several receptor-based drug design methods could be easily used to optimize the structures of compounds, aiming at improving the biological activities of these compounds. A lot of efforts have been invested for studying various targets in both academic institutions and pharmaceutical industries. However, it is time-consuming and expensive to determine whether a chemical and a target are to interact with each other in a cellular network purely by means of experimental techniques. Although some computational methods were developed in this regard based on the knowledge of the 3D

(dimensional) structure of protein, unfortunately their usage are quite limited because the 3D structures for most targets such as many GPCRs are still unknown. Furthermore, analyzing the DTI profiling of the drugs helps to infer drug indications [3], adverse drug reactions [4–6], drug–drug interactions [7, 8], and drug mode of actions [9, 10], etc. Therefore, it is of high importance to reliably and fast predict DTI profiling of the chemicals on a genome-scale level.

Currently, two computational approaches are generally used for studying drug–target relations. (1) The inverse- or reverse-docking approach predicts the interactome of chemicals toward a representative collection of proteins based on various molecular docking programs [11, 12]. For example, Li et al. [11] developed a web server called TarFisDock to identify drug targets from 698 prepared potential targets in advance. Kharkar et al. [12] reviewed the state-of-the-art and future prospects of the reverse docking for drug repositioning and drug rescue. Minho et al. provided large-scale reverse docking profiles by expanding the scope of target space to a set of all protein structures currently available, and developed several new applications such as predicting the druggability of protein targets and protein function prediction based on docking profile similarity [13]. However, a serious problem for docking is that it cannot be applied to proteins whose 3-D structures are unknown. Additionally, the single DTI prediction by docking programs may need to cost seconds even several minutes. Thus, the docking of a chemical toward multiple proteins needs to cost several hours, which seriously limits its wide applications. (2) Various chemogenomics methods simultaneously consider chemical information and protein information to infer chemical-protein associations [14–17]. For example, Nagamine et al. [18] built a statistical model for predicting DTIs based on 519 approved drugs and their associated 29 targets, by using amino acid sequences, two-dimensional chemical structures and mass-spectrometry data. He et al. [19] established classification models for predicting DTI network using chemical functional groups and biological features. Yu et al. [20] made a systematic prediction of multiple DTIs from chemical, genomic and pharmacological data, by using support vector machine and random forest. Xiao et al. [21] developed a sequence-based classifier based on two-dimensional fingerprints of compounds and the pseudo amino acid composition of proteins to predict the interactions between GPCRs and drugs in cellular networking. However, these approaches usually have relatively low prediction accuracies when the number of proteins or the space of DTI data becomes very large [22]. Recently, a variety of statistical methods have been increasingly developed to predict DTIs by integrating multiple evidence sources [23–34]. Yamanishi et al. [25] proposed a bipartite graph learning method to predict true

interacting pairs from the integration of chemical and genomic spaces. Bleakley et al. [27] proposed a bipartite local model by transforming edge-prediction problems into binary classification problems. Xia et al. [29] used a semi-supervised learning method to predict DTIs from heterogeneous biological spaces. Jocab et al. [30] proposed a kernel-based learning framework that constructed the pairwise kernel to measure the similarity between drug–target pairs. However, the drawback of the pairwise kernel is that there will be a large number of samples to be classified (i.e., drugs multiplied by the number of targets) which poses remarkable computational complexity. To avoid this problem, more recently van Laarhoven et al. [31] developed a Gaussian interaction profile kernel for predicting DTIs. Mizutani et al. [34] related DTI network with drug side effects using sparse canonical correlation analysis.

We developed an open web service called TargetNet to net or predict the binding of multiple targets for any given molecule, following the spirit of multi-target SAR methodology. TargetNet simultaneously constructs a large number of SAR models based on current chemogenomics data to make future predictions. 623 Naïve Bayes models together with various molecular fingerprints were employed to construct prediction models for 623 proteins. Ensemble learning from these fingerprints was also provided to improve their prediction ability. When the user submits a molecule, the server will predict the activity of the user's molecule across 623 proteins by the established high quality SAR model for each protein, thus generating a DTI profiling that can be used as a feature vector for wide applications. The 623 SAR models related to 623 proteins were strictly evaluated and validated by several model validation strategies, resulting in the AUC scores of 75–100 %. We applied the generated DTI profiling to successfully predict potential targets, toxicity classification, drug–drug interactions, and drug mode of action, which sufficiently demonstrated the wide application value of the potential DTI profiling. We recommend DTI profiling to analyze and represent various complex molecular data under investigation. Further, we hope that the package will be helpful when exploring questions concerning target identification, candidate drug screening, drug effect evaluation, and poly-pharmacology or multi-target characterization of candidate chemicals [35].

Methods

Preparation of the library drugs and targets

We used BindingDB database as our training datasets. BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions

of proteins considered to be candidate drug–targets with ligands that are small, drug-like molecules [36]. Activity data were filtered to keep only activity end-point points that had half-maximum inhibitory concentration (IC₅₀), half-maximum effective concentration (EC₅₀) or K_i values. Herein, to ensure that enough number of molecules could be used in model building, we previously selected those targets with larger than 200 biological activity data. Following this procedure, 109,061 compounds associated with 623 target proteins remained with 115,257 activity end-points, which were used for model building. The proteins were divided to five classes including enzymes (276) containing kinases (85), ion channels (9), receptors (255), transporter (14) and others (69). The crosslink information related to these targets could be found in the Targets section in the TargetNet website. The list of associated proteins is included in Supplementary Information S1.

Preparation of the positive and negative set

For those compounds with more than one activity values, we took the mean value of their activity values as the final activity value. A compound was considered active when the mean activity value was below 10 μM. All compounds higher than 10 μM are considered inactive. Following this split, maybe some human proteins have very little number of negative samples. To balance the number between positive samples and negative samples for each human protein, we randomly selected certain number of compounds from other human proteins to generate the negative samples for these human proteins. That is to say, the negative samples we used consist of two parts: truly inactive samples and randomly selected unknown interactions. The number of these selected negative samples together with inactive samples should be basically equal to the number of the active samples for these human proteins. These prepared positive set and negative set were used for the subsequent model building. The SMILES formats of the compounds involved in the positive set and negative set for each human protein could be downloaded from the TargetNet website.

Molecular representation

We used molecular substructure fingerprints to describe the information of molecular structure instead of commonly molecular descriptors such as topological, constitutional, geometrical, quantum chemical properties. Substructure fingerprints directly encode molecular structure in a series of binary bits that represent the presence or absence of particular substructures in the molecular [37]. It has the potential to keep the overall complexity of molecules, although it divides the whole molecule into lots of fragments. And it does not need reasonable three-dimensional

conformation of molecules and thereby does not lead to error accumulation for the description of molecular structures. In addition, it gives a direct relationship between molecular structure and property [38]. In the study, several commonly used molecular fingerprints are used to construct the substructure dictionaries, including FP2, Daylight-like, MACCS, Estate, ECFP2, ECPF4 and ECFP6. The FP2 fingerprint is a path-based fingerprint which indexes small molecule fragments based on linear segments of up to 7 atoms. Each remaining fragment is assigned a hash number which is used to set a bit in a 1024 bit vector. The Daylight-like fingerprints are hashed fingerprints encoding each atom type, all augmented atoms and all paths of length 2–7 atoms, giving a total string of 1024 bits. The MACCS fingerprint uses a dictionary of MDL keys, which contains a set of 166 mostly common substructure features. There are referred to as the MDL public MACCS keys. There is a one-to-one correspondence between each SMARTS pattern and bit in the MACCS fingerprint. For each SMARTS pattern, if its corresponding substructure is present in the given molecule, the corresponding bit in the fingerprint is set to 1; conversely, it is set to 0 if the substructure is absent in the molecule. Electrotopological State (E-state) fingerprints represent the presence/absence of 79 E-state substructures [39]. The ECFP2, ECPF4 and ECFP6 fingerprints are in the family called Morgan fingerprints by setting the diameter of the atom environment to 2, 4 and 6, which is known as circular fingerprints. The fingerprints are calculated by the PyDPI which is a Python package developed for calculating various molecular descriptors and fingerprints [40].

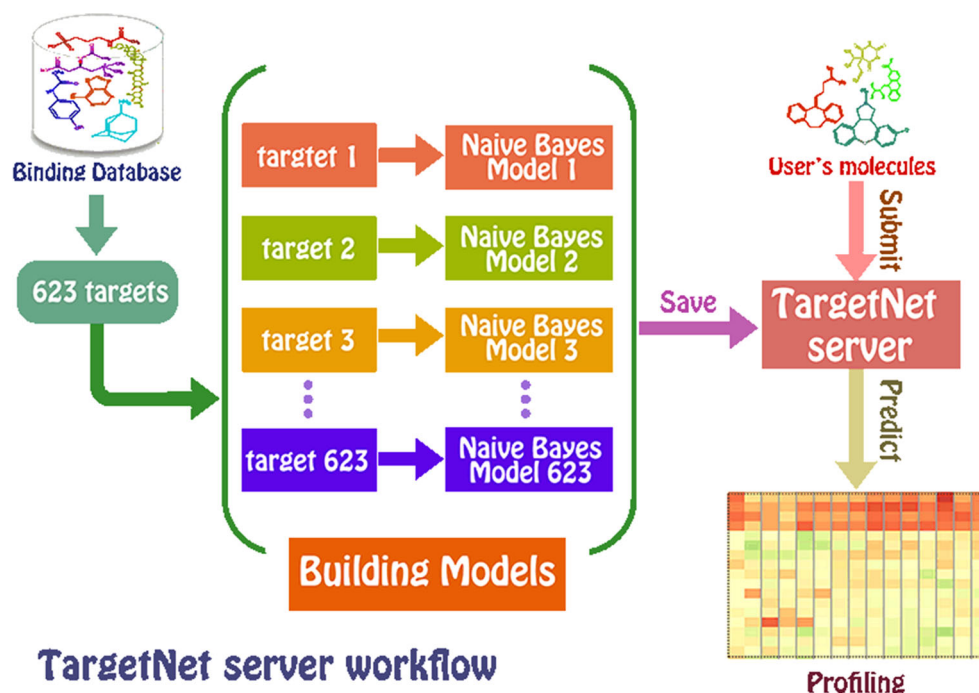
Naive Bayesian classifiers

A series of high confidence SAR models were built using BindingDB (see Fig. 1). Naive Bayes models were built with different fingerprint representations for 623 proteins. The Naive Bayes method for predicting DTI profiling was chosen as it provided both good performance for noisy data sets and a high speed of calculation [41–43]. Bayes' theorem describes the probability of the event A based on the condition B that might be related to the event. It is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naive Bayes classifiers are a class of probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions. The probability model of classification is a conditional model $P(C|F_1, \dots, F_n)$ over a dependent class variable C with a number of classes, conditional on feature from F_1 to F_n . Herein, C presents the target active class of a molecule: active class (+) or

Fig. 1 The algorithmic workflow of the TargetNet web server



inactive class (-). F_1 to F_n represent the calculated values for the feature values (molecular fingerprints). While Bayes's theorem is used, the equation is decomposed as

$$P(+|F_1, \dots, F_n) = \frac{P(F_1, \dots, F_n|+)P(+)}{P(F_1, \dots, F_n)}$$

In plain English, using Bayesian probability terminology, the above equation can be represented as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

where $P(F_1, \dots, F_n|+)$ is the conditional probability of a particular compound being classified as target active; $P(+)$ is the prior probability extracted from a set of compounds in the training set; $P(F_1, \dots, F_n)$ is the marginal probability of the calculated descriptors that will occur in the training set. Then, we raised an assumption that each feature, F_i is independent from every other feature F_j . The mathematical procedure to train a Naive Bayesian classifier was described previously. An advantage of the Naive Bayesian classification is that a small amount of training data is needed to estimate the parameters necessary for classification. And features can contain lots of zeros while the models can get good estimates. Moreover, Naive Bayes classification can process lots of data, learn fast, and be tolerant of random noise [42, 43]. The Naive Bayes classifiers were developed in scikit-learn which is a python package for machine learning.

Performance evaluation

For each model, we applied fivefold cross validation and external validation to evaluate the prediction performance of models. For fivefold cross validation, the data set is split into five roughly equal-sized parts firstly, and then we fit the model to four parts of the data and calculate the error rate of the other part. The process is repeated five times so that every part can be predicted as a validation set. To observe the stability of models, we repeated the cross validation program ten times to report standard deviations of each statistics. For the external validation, the data were split into two parts for the validation step: compounds were clustered and assigned a cluster number. Clusters with an odd number were assigned to the test set, and the clusters with an even number were assigned to the training set. Models were built with the training set, and the test set was scored. Finally, a model was built with all data and scored against itself—the training set and whole set should provide similar validation statistics. Statistics on the performance of the models were reported, including commonly used ones in classification schemes: the sensitivity (SE), specificity (SP), the accuracy (ACC), Matthews correlation coefficient (MCC) and F1-score values.

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

The above statistics were calculated for both test sets and cross-validation sets. Herein, to obtain the best model performance, we compared 7 types of molecular fingerprints when establishing the prediction models. To obtain the better prediction ability, we also ensemble all fingerprint models to obtain the average output (see Fig. 2). In addition, the receiver operating characteristic (ROC) curve was plotted. The ROC analysis provides an overall score and does not need to specify a cut-off for distinguishing active from inactive compounds. The area under the receiver operating characteristic curve (AUC) provides an indication of the ability of the model to prioritize active compounds over inactive compounds. The ROC curve is the plot of the true positive versus the false positive rate.

Results

Model evaluation

For each protein, we constructed and evaluated one SAR model based on several model validation criteria. To obtain high performance SAR models, we compared the prediction ability of seven molecular fingerprints. Figure 3 shows the box plot of AUC scores of 623 target models based on seven molecular fingerprints. Clearly, our constructed SAR

models for 623 proteins based on different molecular fingerprints all obtained reasonable prediction performance on the whole. Three ECFP fingerprints based on different diameters seem to give the better prediction. Among three, the Naïve Bayes model using ECFP4 yields the best prediction. The models based on ECFP4 fingerprint obtained the prediction accuracies of 70–99.8 %, the AUC scores of 75.9–100 %, the MCC scores of 0.42–1.00, F scores of 68.7–100 %, respectively. To further compare the prediction ability of seven molecular fingerprints and observe the prediction performance of 623 targets, we also count the target number in which the AUC scores is higher than some threshold. Figure 4 shows the plot of target number versus AUC score for seven molecular fingerprints. Clearly, the ECFP4 fingerprints obtained the best prediction again. For example, 97.3 % of the models based on ECFP4 fingerprints obtained the AUC scores higher than 90 %. From these prediction statistics, we can see that these SAR models are reliable and robust, and could be used for predicting new molecules. Furthermore, all 623 SAR models and used datasets relating to 623 proteins could be downloaded from the TargetNet website. They can freely be applied to various study problems conducted by the users. For more detailed information, the user could refer to the Documentation section in the website.

Application 1: predicting potential target proteins for the given molecule

As an example, we submitted the drug bromocriptine to TargetNet for a prediction test. The server predicts that bromocriptine might interact with a new protein D(1A) dopamine receptor (Uniprot ID: Q95136). Bromocriptine

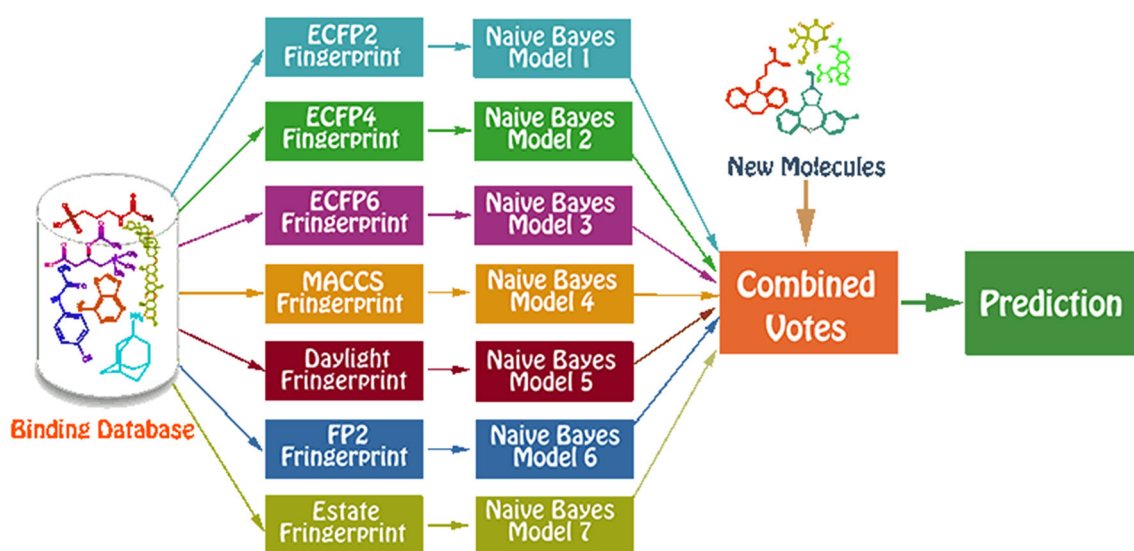


Fig. 2 The algorithmic workflow of ensemble prediction in the TargetNet web server

Fig. 3 Box plot of AUC scores of 623 target models based on seven molecular fingerprints

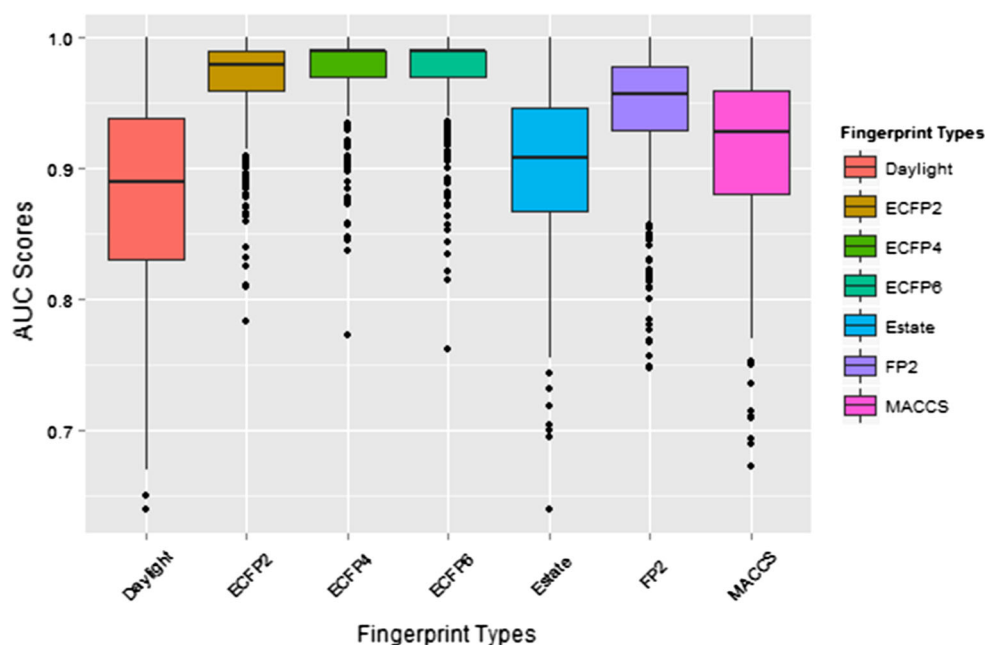
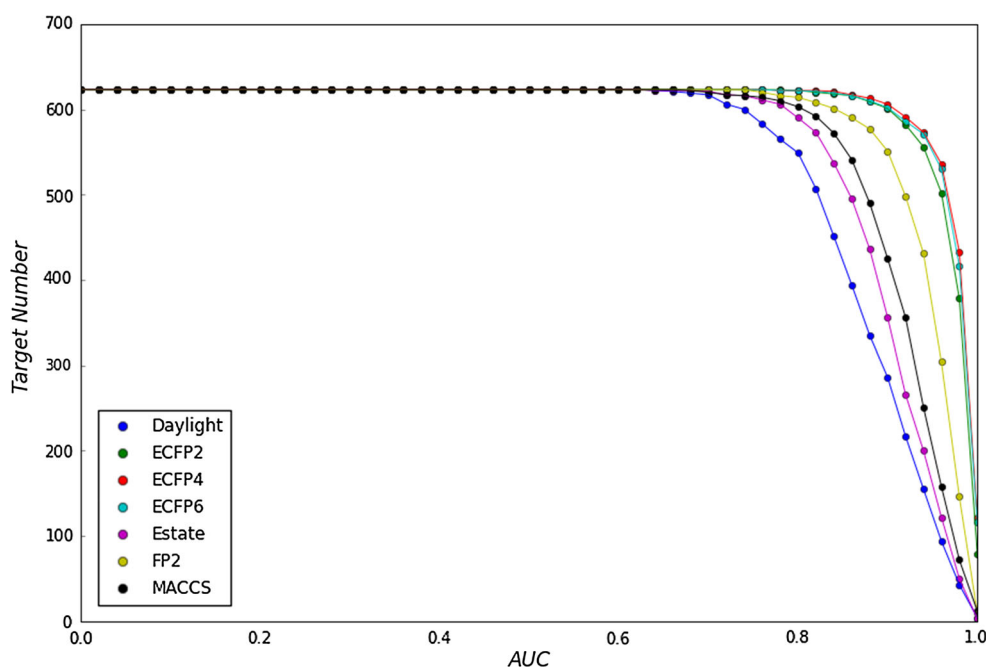


Fig. 4 The plot of target number versus AUC score for seven molecular fingerprints



mesylate is a semisynthetic ergot alkaloid derivative with potent dopaminergic activity. It is indicated for the management of signs and symptoms of Parkinsonian Syndrome. After checking the related literature, we obtained the binding affinity between bromocriptine and D(1A) dopamine receptor ($K_i = 1.444 \mu\text{M}$). Moreover, we also found that most of the approved targets for bromocriptine are predicted in the top 30 associations. This case study demonstrates that our server could predict potential targets to a certain extent.

Application 2: *in silico* toxicity prediction by DTI profiling

We used the DTI profiling as the feature vector to perform the toxicity prediction. Three data sets from Distributed Structure-Searchable Toxicity database network were employed as the benchmark data to evaluate the performance of DTI profiling [38] (Supplementary Information S2). Herein, we used Random Forest (RF) to construct the classification models. These models obtained AUC scores

of 0.82, 0.78, and 0.88, respectively (see Fig. 5a). These prediction results are comparable to even better than those from FP2 fingerprint (see Table 1), which indicating the predictivity of the DTI profiling representation. It should be noted that despite these two representations yield the similar accuracy, as observed in the EPAFHM data, their sensitivity and specificity are very different, which indicated that the information included in two representations should be different. To further compare two molecular representations, we counted the accurately predicted chemicals from two representations, and then compared their differences (see Fig. 5b). One can clearly see that the accurately predicted molecules are not totally same for DTI profiling and FP2 fingerprint. That is to say, although they obtained the similar prediction accuracy, the accurately predicted molecules for each representation are still different from each other. This seems to indicate that different molecular representations preferred different molecules accurately predicted. Maybe the combination of two molecular representations should continue to improve the prediction performance by considering their complementarity. The application study demonstrates that the DTI profiling generated by TargetNet could be used as a new molecular representation, instead of the traditional chemical structural representation, for various studies, such as structure–activity relationship (SAR), absorption, distribution, metabolism, elimination and toxicity (ADMET) prediction, virtual screening, and so on.

Application 3: *in silico* DDI prediction by DTI profiling

Drug–drug interactions (DDIs) may cause serious side-effects that draw great attention from both academia and

industry [44, 45]. Since some DDIs are mediated by unexpected drug–protein interactions, it is reasonable to analyze the DTI profiling of the drugs to predict their DDIs. Herein, we used RF combined with the DTI profiling by TargeNet to construct the DDI prediction model. The used DDI data consist of 1125 drugs and 6743 drug–drug interactions, and could be found in Supplementary Information S3. We applied RF to construct the classification model. In this study, all real drug–drug interaction pairs (i.e., 6743 DTIs) are used as the positive samples. For negative examples we select random, non-interacting pairs from these drug molecules. They are constructed as follows: (1) separate the pairs in the above positive samples into single drugs; (2) re-couple these singles into pairs in a way that none of them occurs in the corresponding positive dataset. To overcome the bias caused by unbalanced problems, we randomly picked the negative pairs formed above until they reached the number one time as many as the positive pairs. We evaluated three levels of model performance: the DDI associations with two drugs included in the training set (Level 1), the DDI associations with only one drug included in the training set (Level 2), and the DDI associations with two drugs not included in the training set (Level 3). To well control the DDI associations, we employed leave-one-out cross validation to evaluate the model. Table 2 lists the prediction statistics based on the DTI profiling for three validation levels. For Level 1, the model obtained the best prediction performance, and gave the accuracy of 90.8 %, the sensitivity of 92.8 %, and the specificity of 88.8 %, respectively. For Level 2, the model gave the accuracy of 71.6 %, the sensitivity of 74.5 %, and the specificity of 68.6 %, respectively. For Level 3, the model gave the accuracy of 68.9 %, the sensitivity of 62.9 %, and the specificity of 75.1 %, respectively. For

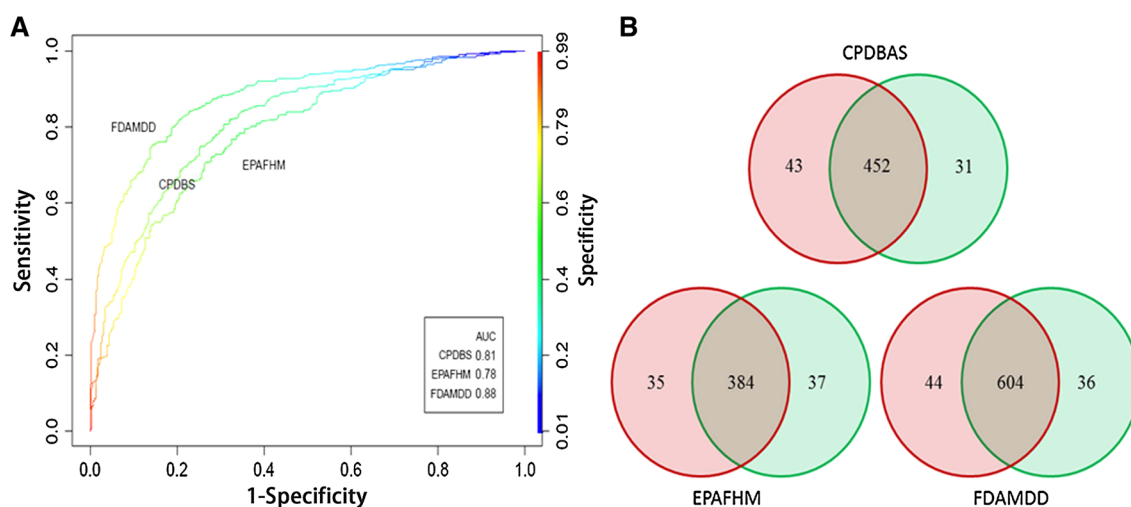


Fig. 5 The prediction results of three toxicity datasets. **a** The ROC curves of three toxicity datasets based on fivefold cross validation. **b** Venn diagram for toxicity predictions using DTI profiling and FP2 fingerprint, respectively

Table 1 The prediction performance of three toxicity datasets based on DTI profiling and FP2 fingerprint

Methods	Datasets	ACC	SE	SP	F1	MCC	AUC
DTI profiling	CPDBAS	0.746 ± 0.003	0.775 ± 0.002	0.715 ± 0.003	0.608 ± 0.002	0.491 ± 0.003	0.815 ± 0.003
	EPAFHM	0.713 ± 0.004	0.728 ± 0.002	0.698 ± 0.003	0.593 ± 0.002	0.426 ± 0.003	0.782 ± 0.003
	FDAMDD	0.807 ± 0.002	0.819 ± 0.002	0.792 ± 0.002	0.621 ± 0.003	0.611 ± 0.002	0.877 ± 0.003
FP2 fingerprint	CPDBAS	0.725 ± 0.004	0.754 ± 0.004	0.693 ± 0.003	0.601 ± 0.003	0.448 ± 0.004	0.818 ± 0.004
	EPAFHM	0.715 ± 0.004	0.686 ± 0.003	0.744 ± 0.004	0.579 ± 0.003	0.431 ± 0.003	0.782 ± 0.003
	FDAMDD	0.795 ± 0.002	0.814 ± 0.002	0.770 ± 0.001	0.619 ± 0.002	0.585 ± 0.002	0.881 ± 0.002

ACC accuracy, SE sensitivity, SP specificity, MCC Matthews correlation coefficient, AUC the area under ROC curves

Table 2 Prediction statistics of DDI data based on different validation levels

Validation	ACC	SE	SP	F1	MCC	AUC
Level 1	0.918 ± 0.001	0.928 ± 0.002	0.888 ± 0.001	0.650 ± 0.003	0.817 ± 0.002	0.969 ± 0.001
Level 2	0.716 ± 0.001	0.745 ± 0.001	0.686 ± 0.002	0.598 ± 0.002	0.432 ± 0.003	0.776 ± 0.002
Level 3	0.689 ± 0.005	0.629 ± 0.003	0.751 ± 0.005	0.557 ± 0.007	0.382 ± 0.005	0.761 ± 0.004

ACC accuracy, SE sensitivity, SP specificity, MCC Matthew's correlation coefficient, AUC the area under ROC curves

Level 1, two drugs associated with the predicted interactions are also included in the training set, and therefore its prediction is relatively easy. For Level 2, one of two drugs associated with the predicted interactions is in the training set, and therefore its prediction is harder than that from Level 1. Clearly, the predictions from Level 3 are the most difficult since two drugs associated with the predicted interactions are not in the training set. This conclusion is similar to that from Part et al. [46, 47], who discussed the flaw and importance of the evaluation schemes of pair-input computational predictions on the protein–protein interaction data sets. The model validation from three levels demonstrated that the DTI profiling by TargetNet could be used as the effective representation to evaluate the DDIs in the clinical trial and drug discovery process.

Application 4: identify network of drug mode of action by DTI profiling

Drug mode of action (MOA) of novel compounds has been predicted using chemical structural features, phenotypic features, or genome-wide expression profiles [9, 48]. As we all know, the molecular targets of drugs are directly related to their MOAs. Here, we draw a “drug network” based on the DTI profiling calculated by TargetNet. We first calculated the DTI profiling from 909 drugs, and then calculated the Pearson correlation coefficients between any two drugs, and finally visualized the drug–drug network according to the anatomical therapeutic chemical classification (ATC) system. Figure 6 shows the network where each node corresponds to a drug, and two nodes are connected by an edge, if the corresponding similarity coefficient is larger

than a predefined threshold (the similarity threshold of 0.97 for this network). The width of an edge is proportional to the similarity between the drugs connected by the edge. The network consists of 494 drugs and their corresponding 835 associations (see Supplementary Information S4 for these associations). In the network, drugs with similar MOAs are connected or lie in the same community. Firstly, we checked if the first level of ATC from two connected drugs is the same. From Fig. 6, one can clearly see that 327 edges from 835 edges indicate the same ATC level. This seems to indicate that the first level in ATC could reflect drug MOAs to a certain extent. For example, two drugs flunisolide and fluciclonide have the similar DTI profiling with a similarity coefficient of 0.987. We found that the principle mechanism of action of two drugs is to activate the glucocorticoid receptors. However, the flunisolide is a corticosteroid often prescribed as treatment for allergic rhinitis, and has been approved as aeroBid, nasalide, and nasarel, while fluciclonide is a topical glucocorticoid used in the treatment of eczema. Likewise, two drugs medrysone and methylprednisolone have also the very similar DTI profiling with the similarity coefficient of 0.993. Medrysone is used as the treatment of allergic conjunctivitis, vernal conjunctivitis, episcleritis, and epinephrine sensitivity, while methylprednisolone is used as adjunctive therapy for short-term administration in rheumatoid arthritis. According to our literature search, we found that they are thought to act by the induction of phospholipase A2 inhibitory proteins, collectively called lipocortins. It is postulated that these proteins control the biosynthesis of potent mediators of inflammation such as prostaglandins and leukotrienes by inhibiting the release of their common

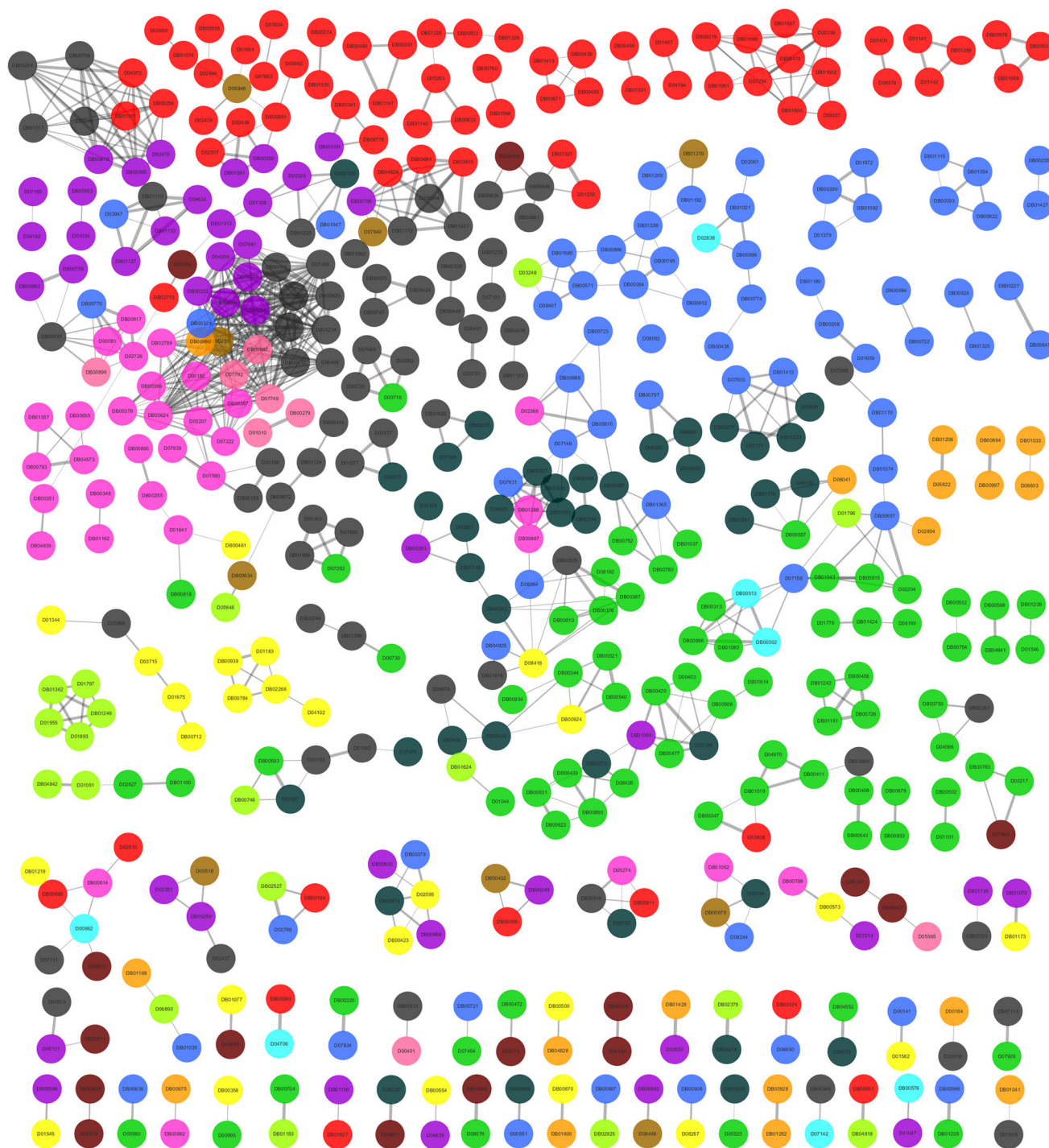


Fig. 6 Drug network obtained by selecting a threshold of 0.97 for the similarity. Each node represents a compound. Two nodes are linked by an edge if their similarity is higher than the predefined threshold. The first level in ATC is indicated by *different colors*

precursor, arachidonic acid. This shows the predictive ability of the network since similarity measures are computed using only the DTI profiling without knowledge of the MOAs of the drugs. The application study demonstrates that the DTI profiling by TargetNet could be used for identifying the drug mode of action (MOA) to some extent.

TargetNet web service

To share our results with pharmacologists and chemists, we finally constructed a web-based prediction server: TargetNet. The TargetNet webserver is freely accessible at <http://targetnet.scbdd.com>. It is running upon Linux/

Apache/Django platform and supported by background Python language, which enables multiple accesses simultaneously. Figure 7 shows the server workflow showcasing TargetNet prediction. The TargetNet can be accessed by selecting ‘Webserver’ link. Users are required to submit a molecular structure or a molecular file with SMILES format. An example drug molecule and file is provided for a quick test. When a user molecule is submitted, the prediction model for each protein will be called, and thus the prediction scores of this drug toward all proteins in the database are calculated via 623 established SAR models with selected performance statistics. The user can also select different performance statistics (e.g., AUC, accuracy, MCC and F score) to determine the number and prediction ability of models. Depending on the number of the submitted molecules, the process time ranges from seconds up to several minutes. Generally speaking, the process time for one molecule calculation is about 5–10 s. Users can also track the real-time calculation process online. For convenience, the user is allowed to draw a drug molecule via JME editor. Examples with standard input formats are also provided to guide the users.

Currently, the TargetNet web service has been applied by more than 500 visits from 39 different countries registered since October 20, 2015. In the 3 months, our TargetNet webserver runs well, and there is no problem of failure for the needs from users. For more details, the editor

could refer to the live statistics in the right corner of our TargetNet homepage.

The user will be able to view the following outputs:

1. DTI probabilities of user’s molecule with 623 proteins in library. The result table includes “Details”, “Uni-prot_ID”, “Protein” and “Probability”. The user can also look over the detailed information for each human protein as needed and conveniently type in a keyword to look for a certain item in the results through the ‘Search’ button. The final prediction results can be downloaded as different formats.
2. The Lipinski’s rule of five for user’s molecule together with the molecular structure.

Discussion

We compared TargetNet with two current methods that can yield the DTI profiling. The first is the inverse- or reverse-docking approach, which predicts the interactome of drugs toward a representative collection of target proteins based on various molecular docking programs. TargetNet has two advantages over inverse- or reverse-docking: (1) the calculation speed of TargetNet is faster than that from inverse- or reverse-docking. As we all know, the single drug–target interaction prediction by docking programs may cost seconds even several minutes. Thus, the docking of a drug toward

The screenshot displays the TargetNet webserver interface. At the top, there is a navigation menu with links for Home, Webserver, Targets, Documentation, Download, Help, and Contact. Below the menu, there is a search bar and a main form for inputting a molecule and selecting performance statistics. The form includes fields for SMILES, performance statistics (AUC, etc.), a cutoff slider, and fingerprint type. A table of calculation results is shown on the right, with columns for Details, Uni-prot_ID, Protein, and Prob. The table lists several proteins and their corresponding probabilities. Annotations with red boxes and arrows highlight key features: '1' points to the 'Webserver' link in the navigation menu; '2' points to the 'SMILES' input field; '3' points to the 'Copy' button in the download options; 'The format of download' points to the download options (Copy, CSV, Excel, PDF, Print); 'Look for a certain item' points to the search bar; 'Link to the protein' points to the 'View' link in the table; and 'The probability of the interaction' points to the 'Prob' column in the table.

Details	Uni-prot_ID	Protein	Prob
View	P31749	RAC-alpha serine/threo...	0.0
View	P34998	Corticotropin-releasing factor receptor 1	0.0
View	P34995	Prostaglandin synthase 2	0.002
View	Q9BY41	Histone deacetylase 6	0.004
View	P08575	Receptor-type tyrosine-protein phosphatase C	1.0
View	P03372	Estrogen receptor	0.996
View	P08473	Nephrilysin	0.0

Fig. 7 The server workflow showcasing TargetNet prediction

multiple proteins needs to cost several hours, which seriously limits its wide applications. (2) The proteins used in the inverse- or reverse docking must have explicit 3D protein structures and the binding pockets, while the proteins in TargetNet can be arbitrary as long as they have enough interactive compounds confirmed experimentally. The second is the chemogenomics approach, which considers drug information and protein information to infer drug–target associations. The computational chemogenomics approaches, however, have relatively low prediction accuracy when the number of target proteins or the space of DTI data becomes large [22]. Herein, TargetNet only used the chemical structural information to differentiate active from non-active for the given protein, based on the SAR principle. Therefore, TargetNet is more flexible and applicable. Furthermore, we also compared our TargetNet with several popular web servers which are established to predict DTIs or potential drug targets by using different methodologies. For instance, DINIES is a web server for predicting unknown drug–target interaction networks from various types of biological data in the framework of supervised network inference [49]. However, the DINIES method requires the detailed side-effect information or/and protein amino acid sequence which is applicable only to marketed drugs for which side-effect information is available, while TargetNet only needs chemical structures, and therefore is more easy-to-use and flexible. Additionally, their method integrates multiple heterogeneous data to calculate the interaction, which is usually time-consuming, while TargetNet only needs about 5–10 s to cope with one molecule, and thereby its computational speed is faster than DINIES. CDRUG is a web server used for predicting anticancer activity from chemical structures of compounds encoded by the Daylight fingerprint [50]. TargetNet conducts certain 623 SAR models and calculated seven types of fingerprints, and has certain targets and more fingerprints than CDRUG. Compared to the CPI-Predictor proposed by Tand et al., the dataset used in two methods is very different although they are all based on SAR methodology. The CPI-Predictor mainly focuses on the GPCRs from GPCR SARfari database and kinases from kinase SARfari database, and then constructs their SAR models based only on the MACCS fingerprint. TargetNet mainly focuses on the Binding databases, and involves five classes of targets mentioned in the Methods section. Furthermore, TargetNet systematically compared seven types of molecular fingerprints or substructure fragments, and found that ECFP4 fingerprint is more predictive than the other fingerprints including MACCS. SwissTargetPrediction is a web server to infer the targets of bioactive small molecules based on the combination of 2D and 3D similarity values with known ligands. Compared to SwissTargetPrediction based on similarity, TargetNet applied SAR methodology to infer DTIs. However, it is worth noting that the performance of

TargetNet largely depends on the quality of each SAR model related to each protein. Those factors influencing the quality of SAR models will directly influence the prediction ability of TargetNet, and then influence the efficiency of DTI profiling, such as the size and diversity of datasets, model quality, molecular structural representations, etc. In the process of building SAR models, we have sufficiently considered several factors to obtain the high-quality SAR models. For example, the size of each dataset is limited to be not less than 200, and the diversity analysis of the dataset is also visualized. Furthermore, a series of model validations and evaluations are performed to ensure the reliability of models.

Conclusions

- i. TargetNet server can predict DTI profiling for the user's drug across 623 proteins in the database, which is supported by the prediction statistics from cross validations, independent validations, and applications.
- ii. TargetNet can help to infer drug indications, adverse drug reactions, drug–drug interactions, and drug mode of actions, and will have wide applications in drug discovery process.
- iii. The DTI profiling by TargetNet could be considered as a new molecular representation for various drug discovery studies.

Acknowledgments We would like to thank the Django group for their great Django server. We would also like to thank Dr. Peter Ertl for his JME molecular editor, and we thank the developers of D3.js. We would also like to thank three anonymous referees and the editor for their constructive comments, which greatly helped improve upon the original version of the manuscript.

Funding This work has been financially supported by grants from the Project of Innovation-driven Plan in Central South University, the National Natural Science Foundation of China (Grants No. 81402853), the National key basic research program (Grants No. 2015CB910700), and the Postdoctoral Science Foundation of Central South University, the Chinese Postdoctoral Science Foundation (2014T70794, 2014M562142). The studies meet with the approval of the university's review board.

Compliance with ethical standards

Conflict of interest None.

References

1. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) *Nat Biotechnol* 25(10):1119
2. Nunez S, Venhorst J, Kruse CG (2011) *Drug Discov Today* 17(1):10

3. Gottlieb A, Stein GY, Ruppini E, Sharan R (2011) *Mol Syst Biol* 7(1):496
4. Luo H, Chen J, Shi L, Mikailov M, Zhu H, Wang K, He L, Yang L (2011) *Nucleic Acids Res* 39(Suppl 2):W492
5. Cao DS, Xiao N, Li YJ, Zeng WB, Liang YZ, Lu AP, Xu QS, Chen A (2015) CPT: pharmacometrics & systems. *Pharmacology* 4(9):498
6. Wienkers LC, Heath TG (2005) *Nat Rev Drug Discov* 4(10):825
7. Luo H, Zhang P, Huang H, Huang J, Kao E, Shi L, He L, Yang L (2014) *Nucleic Acids Res* 42(W1):W46
8. Tatonetti NP, Ye PP, Daneshjou R, Altman RB (2012) *Sci Transl Med* 4(125):125ra31
9. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A (2010) *Proc Natl Acad Sci* 107(33):14621
10. Iorio F, Tagliaferri R, Bernardo Dd (2009) *J Comput Biol* 16(2):241
11. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J (2006) *Nucleic Acids Res* 34(suppl 2):W219
12. Kharkar PS, Warriar S, Gaud RS (2014) *Fut Med Chem* 6(3):333
13. Lee M, Kim D (2012) *BMC Bioinformatics* 13(Suppl 17):S6
14. Cao D-S, Liang Y-Z, Deng Z, Hu Q-N, He M, Xu Q-S, Zhou G-H, Zhang L-X, Deng Z, Liu S (2013) *PLoS One* 8(4):e57680
15. Cao D-S, Liu S, Xu Q-S, Lu H-M, Huang J-H, Hu Q-N, Liang Y-Z (2012) *Anal Chim Acta* 752:1
16. Bredel M, Jacoby E (2004) *Nat Rev Genet* 5(4):262
17. Klabunde T (2007) *Br J Pharmacol* 152(1):5
18. Nagamine N, Sakakibara Y (2007) *Bioinformatics* 23(15):2004
19. He Z, Zhang J, Shi X-H, Hu L-L, Kong X, Cai Y-D, Chou K-C (2010) *PLoS One* 5(3):e9603
20. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, Li X, Zhou W, Wang W, Wang Y (2012) *PLoS One* 7(5):e37608
21. Xiao X, Min J-L, Wang P, Chou K-C (2013) *PLoS One* 8(8):e72234
22. Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y (2012) *Mol BioSyst* 8(9):2373
23. Cheng F, Zhou Y, Li W, Liu G, Tang Y (2012) *PLoS One* 7(7):e41064
24. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y (2012) *PLoS Comput Biol* 8(5):e1002503
25. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) *Bioinformatics* 24(13):i232
26. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) *Science* 321(5886):263
27. Bleakley K, Yamanishi Y (2009) *Bioinformatics* 25(18):2397
28. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijjer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL (2009) *Nature* 462(7270):175
29. Xia Z, Wu L-Y, Zhou X, Wong S (2010) *BMC Syst Biol* 4(Suppl 2):S6
30. Jacob L, Vert J-P (2008) *Bioinformatics* 24(19):2149
31. van Laarhoven T, Nabuurs SB, Marchiori E (2011) *Bioinformatics* 27(21):3036
32. Chen X, Liu M-X, Yan G-Y (2012) *Mol BioSyst* 8(7):1970
33. Mei J-P, Kwok C-K, Yang P, Li X-L, Zheng J (2013) *Bioinformatics* 29(2):238
34. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y (2012) *Bioinformatics* 28(18):i522
35. Csermely P, Agoston V, Pongor S (2005) *Trends Pharmacol Sci* 26(4):178
36. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) *Nucleic Acids Res* 35(suppl 1):D198
37. Scott DE, Coyne AG, Hudson SA, Abell C (2012) *Biochemistry* 51(25):4990
38. Cao DS, Yang YN, Zhao JC, Yan J, Liu S, Hu QN, Xu QS, Liang YZ (2012) *J Chemom* 26(1–2):7
39. Cao D-S, Xu Q-S, Hu Q-N, Liang Y-Z (2013) *Bioinformatics* 29(8):1092
40. Cao D-S, Liang Y-Z, Yan J, Tan G-S, Xu Q-S, Liu S (2013) *J Chem Inf Model* 53(11):3086
41. Bender A, Mussa HY, Glen RC, Reiling S (2004) *J Chem Inf Comput Sci* 44(1):170
42. Wang S, Li Y, Wang J, Chen L, Zhang L, Yu H, Hou T (2012) *Mol Pharm* 9(4):996
43. Watson P (2008) *J Chem Inf Model* 48(1):166
44. Zhang L, Zhang Y, Zhao P, Huang S-M (2009) *AAPS J* 11(2):300
45. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-w, Matheny ME, Xu H (2012) *J Am Med Inf Assoc* 19(E1):E28
46. Park Y, Marcotte EM (2012) *Nat Methods* 9(12):1134
47. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T (2015) *Brief Bioinform* 16(2):325
48. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN (2006) *Science* 313(5795):1929
49. Yamanishi Y, Kotera M, Moriya Y, Sawada R, Kanehisa M, Goto S (2014) *Nucleic Acids Res* 42(W1):W39
50. Li G-H, Huang J-F (2012) *Bioinformatics* 28(24):3334