

RESEARCH

Open Access



In-silico target prediction by ensemble chemogenomic model based on multi-scale information of chemical structures and protein sequences

Su-Qing Yang^{1,2†}, Liu-Xia Zhang^{3†}, You-Jin Ge², Jin-Wei Zhang⁴, Jian-Xin Hu², Cheng-Ying Shen², Ai-Ping Lu⁵, Ting-Jun Hou^{6*} and Dong-Sheng Cao^{1,5*}

Abstract

Identification and validation of bioactive small-molecule targets is a significant challenge in drug discovery. In recent years, various *in-silico* approaches have been proposed to expedite time- and resource-consuming experiments for target detection. Herein, we developed several chemogenomic models for target prediction based on multi-scale information of chemical structures and protein sequences. By combining the information of a compound with multiple protein targets together and putting these compound-target pairs into a well-established model, the scores to indicate whether there are interactions between compounds and targets can be derived, and thus a target prediction task can be completed by sorting the outputted scores. To improve the prediction performance, we constructed several chemogenomic models using multi-scale information of chemical structures and protein sequences, and the ensemble model with the best performance was used as our final model. The model was validated by various strategies and external datasets and the promising target prediction capability of the model, i.e., the fraction of known targets identified in the top-k (1 to 10) list of the potential target candidates suggested by the model, was confirmed. Compared with multiple state-of-art target prediction methods, our model showed equivalent or better predictive ability in terms of the top-k predictions. It is expected that our method can be utilized as a powerful computational tool to narrow down the potential targets for experimental testing.

Keywords Target prediction, Chemogenomic, XGBoost, Ensemble model

[†]Su-Qing Yang and Liu-Xia Zhang have contributed equally to this work

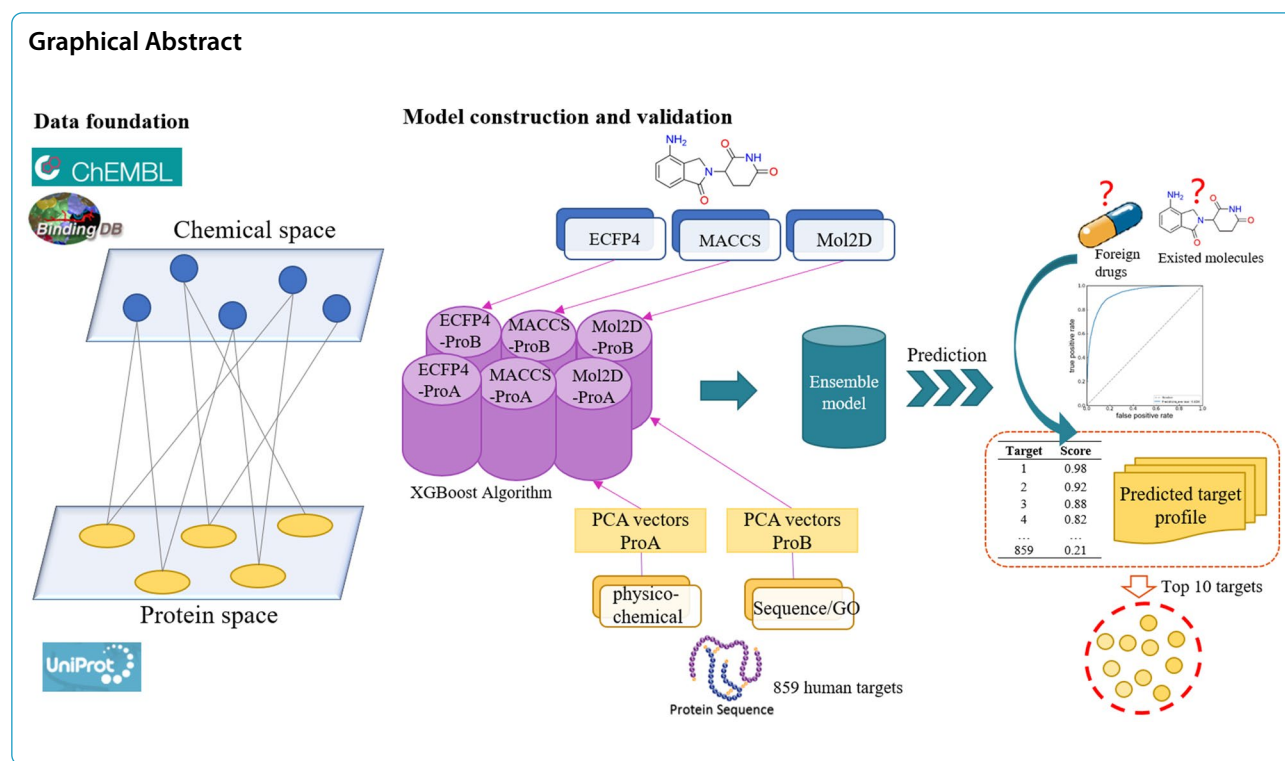
*Correspondence:

Ting-Jun Hou
tingjunhou@zju.edu.cn
Dong-Sheng Cao
oriental-cds@163.com

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



Introduction

It is estimated that 52% of clinical phase II failures are primarily due to insufficient efficacy, in which most are caused by poor targeting or unfavorable off-target effects [1, 2]. Apparently, identification of the potential targets for drug candidates in the early stages of drug discovery may reveal their adverse side effects, thereby reducing the attrition rate in clinical trials. Moreover, traditional drug discovery primarily followed the ‘one-compound, one-target, one-disease’ diagram, implying that a drug is designed to modulate a single target for a specific disease. But it is well known that most drugs bind to multiple targets, a general phenomenon known as polypharmacology [3]. The interactions between these secondary targets and drugs may be responsible for unexpected off-target effects, which usually induce unfavorable side effects but may also provide more opportunity for drug repositioning. For example, Sildenafil, the first selective type 5 phosphodiesterase (PDE₅) inhibitor for the treatment of angina pectoris, has been repurposed for the treatment of penile erectile dysfunction (PED) and pulmonary hypertension [4]. Other notable drug repositioning examples include Memantine [5], Buprenorphine [6], Requip [7, 8], Colesevelam [9] and so on. Nowadays, due to the great difficulty and financial strain of drug discovery, identifying new indications for old drugs is informed the best way to bring a drug to market [10].

With the development of experimental techniques, protein targets can be identified by chemical proteomics methods such as affinity chromatography and activity-based protein profiling. In such experiments, modified or labelled compounds would specifically bind to proteins, and then the related protein targets can be precipitated or traced [11–18]. However, the modification and labelling of the key functional groups of the query compound may hamper compound-protein interactions. Moreover, these experimental approaches are labor-intensive, time-consuming and costly, and may suffer from high false-positive rate. Alternatively, driven by massive bioactivity data deposited in public chemogenomic databases such as ChEMBL [19], DrugBank [20], and TTD [21], *in-silico* target prediction has shown promise in recent years. By screening a compound against a database, it is possible to identify potential target candidates for subsequent experimental validation [22–25].

Generally, computational target prediction methods are classified into two categories: structure-based and ligand-based. The former methods detect the possible targets based on the three-dimensional (3D) crystal structure information of proteins, focusing on docking a query compound either to a set of targets or mapping to the pharmacophores inferred from ligand-target complexes [26–29]. However, the necessity of the 3D structures of proteins makes these methods applicable to a small range. Moreover, the uncertain relation

between bioactivities and physicochemical properties served for scoring and the insufficient accuracy of scoring functions also show their weakness. Differently, the latter methods, mapping targets through the insight of the similarities between two compounds based on the hypothesis that similar compounds are likely to have similar target-binding profiles, are approved to achieve better predictive performance [30–34]. The most common implementation is machine learning (ML), which is accomplished by combining multiple independent binary classifiers. Each binary classifier trains on ligand information (i.e., descriptors) associated with a target and then learn knowledge that can correctly map descriptors to the target [35–38]. However, as ML methods do not take any protein information into account, the interactions between targets and compounds have not been fully explored. More importantly, if the number or structural diversity of the ligands for some targets is insufficient, the mapping functions cannot be guaranteed and well established.

Recently, the chemogenomic methods, by combining the protein sequence information with the compound-target interaction data to prediction models, making up for some key information of interactions and increasing the number of ligands for some targets by sharing ligands with targets having similar sequences, offset some weakness of ML methods discussed above [39–41]. These methods utilize both ligand and target spaces to extrapolate the bioactivities of compounds. Typically, a vector of descriptors representing each compound-target pair is taken as the input, and the output is whether there is an interaction between a compound and a target [42–44]. What inspires us is that, given the characteristic of this approach, a target prediction task can also be completed through combining a basket of protein targets for a compound and putting these compound-target pairs into a model to yield predictions. The score of each compound-target pair represents the probability of the association between the compound and protein, and finally the top-ranked targets are regarded as the potential targets. But as we already known, the prediction performance of the methods for target prediction has not been systematically evaluated.

In this study we constructed several chemogenomic models by integrating two types of protein descriptors and three types of molecular descriptors. These models are equipped with good binary classification that can accurately differentiate the compound-target interactions with strong binding affinity from those with weak binding affinity. Driven by the fact that the more descriptors from different insights included in the models, the more excellent prediction will be derived, different ensemble ML models were established and fully assessed, and the

best one was selected as the final prediction model. The target prediction performance of the models (i.e., the fraction of known targets identified in the top-k (1 to 10) prediction list) was validated by various strategies and external datasets. Statistically, 26.78% and 57.96% of the known targets were enriched in top-1 and top-10 of the prediction list according to the stratified tenfold cross validation, respectively, suggesting approximately 230-fold and 50-fold enrichments. When validated by the external datasets including natural products, more than 45% targets were enriched in the top-10 of the prediction list. Compared with multiple state-of-art target prediction methods, our model yielded equivalent or better predictive ability on the top-k predictions.

Materials and method

Dataset collection

In this study, 859 human target proteins from the ChEMBL27 database [45] were collected for target prediction. Although some target proteins from other species also bind drug-like ligands, they were excluded because the prediction of targets against human species is our main focus. The collected targets mainly cover kinases, GPCRs, proteases, enzymes, and proteins from other detailed categories, among which 294 are FDA approved targets, 256 for clinical trial targets, 53 for patent targets, 236 for investigational targets and 20 targets documented in the literatures (Fig. 1). In addition, the target information including protein sequence and gene ontology (GO) terms with three subclass of biological process (BP), molecular function (MF) and cellular component (CC) was retrieved from the UniProt database [46] in order to facilitate the calculation of protein descriptors for constructing the prediction models. The detailed information of these 859 targets served for prediction can be seen in Additional file 1.

The entire dataset for modeling is composed of 153,281 compound-target interactions extracted from the Binding database v2020 [47] and the ChEMBL27 database [45], associated with 859 target proteins and 93,281 unique compounds. For each compound-target pair, its corresponding bioactivity data (K_i) were extracted from these two databases. It is possible that multiple bioactivity data may be found for one compound-target pair due to the integration of different sources or literatures. A median of these bioactivity data was used for such pairs whose difference is below one magnitude, and those pairs whose bioactivity difference exceeds one magnitude were excluded. The detailed information of modeling data can be seen in Additional file 2.

The K_i value of 100 nM was used as the threshold to tune the positive set (compound-target pairs with $K_i \leq 100$ nM) and the negative set (compound-target

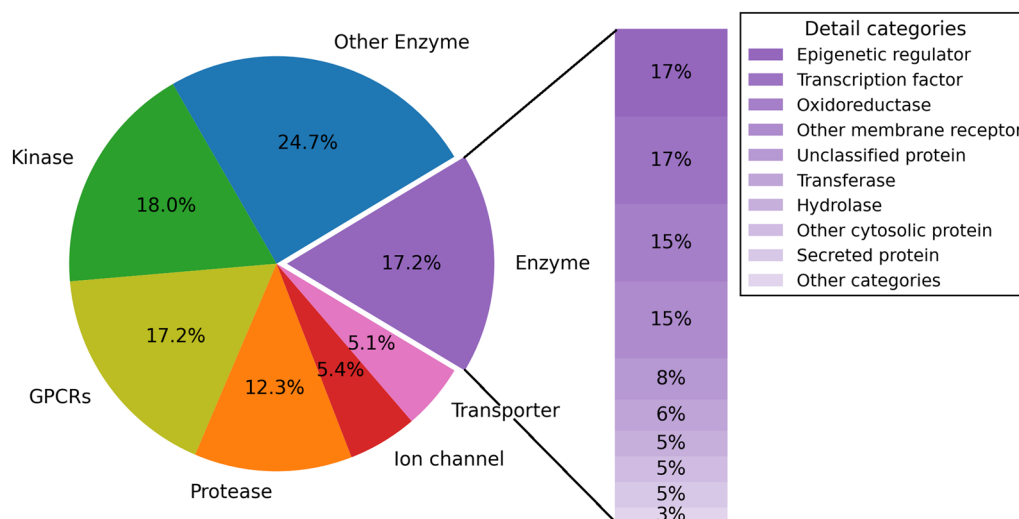


Fig. 1 Category distribution of 859 targets for prediction

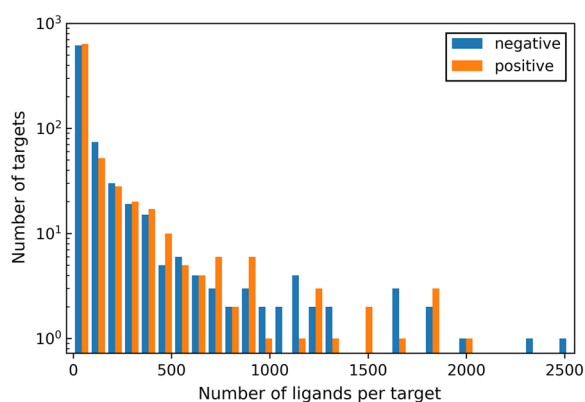


Fig. 2 The distribution of the numbers of the positive samples and negative samples associated with each target

pairs with $K_i > 100$ nM). Thus, the entire data set was firstly divided into 80,608 positive samples and 72,673 negative samples. Of these 859 targets, 549 had 10 or more known data points, 240 had more than 100 data points and 37 had more than 1000 data points. In more detail, 380 targets had 10 or more positive samples (namely active compound-target interactions), 145 targets had more than 100 positive samples, and 17 targets had more than 1000 positive samples (Fig. 2). On average, each target had 173.6 data points and 93.8 positive samples, the maximum numbers of the data points and positive samples of a target (carbonic anhydrase 2) were 3,351 and 2,013, respectively.

Chemical structure and protein sequence representation

The descriptors used to represent and describe data decide the application range and success of a ML model.

Structural descriptions under different levels sketch different compound/protein behaviors and provide diverse clues to inferring compound-target interactions. One-sided descriptors may not contain enough features to fully characterize the chemical and biological spaces of the data, provided the occurrence of “activity cliff” which presents pairs of compounds with high structural similarity but unexpectedly large activity (or property) difference. As a supplement, this gap may be captured by other types of descriptors. Therefore, to fully represent the comprehensive target-ligand interaction space, compounds were represented by three types of descriptors and proteins were described by three-level characterizations involving physicochemical properties, protein sequences and gene ontology (GO) terms.

For compound representation, three different molecular descriptors without the requirements of the three-dimensional conformations were used: (1) 188 Mol2D descriptors derived from the article proposed by Dong et al., including 30 molecular constitutional descriptors, 25 topological descriptors, 44 molecular connectivity indices, 7 kappa shape descriptors, 21 Basak descriptors, 25 charge descriptors, and 60 MOE-type descriptors [48, 49]; (2) Extended Connectivity Fingerprint with a bond diameter of 4 (ECFP4), a class of 1024 bit circular fingerprints developed specifically for SAR modeling [50]; (3) MACCS fingerprints recording the occurrence of 166 predefined substructures reported to effectively encode molecular structure [51]. The above descriptors were calculated using MOE [52], ChemDes [53], ChemoPy [54], PyBioMed [48], and PyDPI [55].

The used protein descriptors were classified into two parts: protein type A (ProA) and protein type B (ProB).

Protein sequences from the Uniprot database were used as the source for calculation. ProA was designed to execute the computation of seven physicochemical descriptor groups including amino acid composition descriptors, autocorrelation coefficient descriptors, CTD (composition, transition, and distribution) descriptors, conjoint triad descriptors, quasi-sequence-order descriptors, pseudo-amino acid composition descriptors, and proteochemometrics descriptors [48]. In order to reduce the model load, the multivariate descriptors (more than 50 dimensionality) from different sub-groups are projected to a lower-data space (50 dimensionality) from its most informative viewpoint by principal component analysis (PCA) [56]. Thus, each protein is described by 762 ProA descriptors. ProB is the descriptors derived from similarity matrix, which records the similarity between each pair of 859 protein pairs, including protein sequence similarity and the GO term similarity. Technically, the sequence similarities between each pair of proteins were calculated using the Resnik algorithm and GO semantic similarity measures including BP, MF and CC were obtained using the BLOSUM62 algorithm. In this manner, four similarity matrices of 859×859 were obtained, and each row of the matrix is the descriptors for each protein [57, 58]. Through PCA as before, ProB descriptors of each sample were reduced to 200 (4×50) dimensionality [56]. The detailed information of the descriptors and the percentage of explained variance (%VAR) of PCA were shown in Additional file 3. Here, the introduction of the matrix-derived descriptors limits the target prediction application of the models to these 859 targets [43, 59–61]

In this work, we employed the chemogenomic approach to encode the compound-target interactions by using both compounds and target proteins. An interaction can be efficiently represented by simultaneously considering the structural information from this compound and this protein. Through the combination of the structures from related compound and related protein (i.e., six combined descriptors of ECFP4-ProA, ECFP4-ProB, Mol2D-ProA, Mol2D-ProB, MACCS-ProA, and MACCS-ProB), each interaction sample (positive or negative) is finally characterized as a 1786, 1044, 950, 388, 928, 366 dimensional vectors, respectively.

Machine learning methods

The extreme gradient boosting (XGBoost) algorithm was employed to construct the classification models [62]. XGBoost is an efficient and scalable implementation of the gradient boosting framework, and it provides insights on cache access patterns, data compression, and fragmentation. XGBoost develops the model in a sequential stage-wise fashion like other boosting methods do, and generalizes them by allowing optimization of an arbitrary

differentiable loss function [63–65]. It has been regarded as a new generation of ensemble learning algorithms, which has become the winners for several ML competitions in recent years [66].

In the implementation, Konstanz Information Miner (KNIME) [67], a platform integrating data processing, data analysis, data exploration and Python package [68], was applied to construct models. The main hyperparameters, including learning rate (eta), the maximum depth of a tree (maximum depth), the minimum loss reduction required to make a further partition on a leaf node of the tree (gamma) and the number of models to train in the boosting ensemble (boosting rounds), were optimized by using the grid search method and the stratified tenfold cross-validation.

Performance evaluation

The primary task of the model is to distinguish compound-target interactions with strong binding affinity from those with weak binding affinity, namely binary classification. Only when the model is equipped with satisfied binary classification performance, the target prediction performance can be guaranteed because it requires not only the binary classification capability of the classifier but also the ability to enrich the potential active targets of the compound at the top of the prediction list, namely early retrieval, typically the top 10 targets (top 0.1–1%) so that users are able to obtain a reasonable number of targets to be experimentally tested. Therefore, the binary classification performance of the model is firstly evaluated, and subsequently the target prediction ability is measured.

To ensure that the derived model has good generalization ability, the stratified tenfold cross-validation (CV) was used where the stratification process guaranteed that samples from each target were present in both the training and test dataset and samples from some targets which have a small quantity of ligands (≤ 5) were present only in the training datasets. By definition, the compounds or targets in the training set are called 'known', whereas those not existing in the training set are called 'new'. Compared with the training set, there are two types of test set: (1) known compounds and known targets (intend to identify more possible targets for known active compounds); (2) new compounds and known targets (intend to identify targets for new compounds). Therefore, we conducted two levels of validation: pair-split validation and compound-split validation. As for the pair-split validation, the training and test sets were generated by randomly splitting the dataset according to the stratification. It measures the average performances of our models as the test datasets include both two types of pairs. As for the compound-split validation, it splits the

compounds into 10 parts, and therefore the compound-target interactions associated with 1 out of these 10 parts were used as the test set and the interactions associated with the remaining 9 parts were kept in the training set. It assumes the situation where we want to detect the targets for external compounds. In order to evaluate the robustness of the model, the stratified tenfold CV was repeated 50 times and the obtained mean values and variance values were used to quantify the performance.

In the assessment, the binary classification performance of compound-target interactions was evaluated by several commonly used statistical parameters: true positives (TP), false negatives (FN), true negatives (TN), false positives (FP), the overall prediction accuracy ($ACC = (TP + TN) / (TP + TN + FP + FN)$), the prediction accuracy of the positive set (Sensitivity, $SE = TP / (TP + FN)$), and the prediction accuracy of the negative set (Specificity, $SP = TN / (TN + FP)$). Besides, the receiver operating characteristic (ROC) curve was plotted, and the area under the receiver operating characteristic curve (AUC) was used to assess each of the models.

The target prediction performance was verified by the recall rate, namely the fraction of known targets identified in the top k of the prediction list. For each compound to be predicted in the test set, the features from its compound descriptors combined with 859 protein targets, namely 859 compound-target integrated descriptors (i.e., six combined descriptors of ECFP4-ProA, ECFP4-ProB, Mol2D-ProA, Mol2D-ProB, MACCS-ProA, and MACCS-ProB), were inputted into the corresponding prediction model and then 859 compound-target interaction scores could be outputted. The targets ranked top-k of the prediction list are recognized as the potential targets, whereas the other targets are assumed inactive. An arbitrary cutoff of k (1–10) predictions was feasible

number of protein targets that could be screened and differences in classifier performance after this cutoff will be missed. The recall rate is relatively harsh as it requires a classifier to have placed a correct target for a compound in the top 0.1%–1% of the lists but it gives an indication for the practicability of a model for target prediction.

Results and discussion

Compound-target interactions can be accurately predicted from integrated features

Our first concern in this study is to construct a predictive model that can accurately differentiate compound-target interactions with strong binding affinity from those with weak binding affinity. To represent compound-target interactions, we used a chemogenomic framework. In brief, an interaction is represented by simultaneously considering the structure content from this compound and this protein. Thus, each interaction sample (positive or negative) is finally characterized by a fixed dimensional vector by combining the structural content from the related compound and protein. Each of these factors can be considered as a separate coordinate spanning a multidimensional space, and in this sense a compound-target interaction is an event in this type of multidimensional space.

Firstly, the classification performance of compound-target interactions was evaluated. The statistics of the predictions on the stratified tenfold CV are summarized in the “Integrated” rows of Table 1. The ROC curves are shown in Fig. 3.

From Table 1, both the pair-split and compound-split models performed well with an average ACC up to 0.81 and an average AUC up to 0.90, and the low standard deviations obtained from the 50 repetitions of the model shows the robust predictive performance of the models.

Table 1 Statistical results of the models derived from different descriptors (integrated or separated groups) on the stratified tenfold CV

Descriptors		Pair-split				Compound-split			
		ACC	SE	SP	AUC	ACC	SE	SP	AUC
Integrated	ECFP4-ProA	83.96 ± 0.12	85.74 ± 0.11	82.00 ± 0.17	92.67	83.58 ± 0.06	85.35 ± 0.09	81.65 ± 0.09	92.33
	ECFP4-ProB	83.99 ± 0.16	85.87 ± 0.12	81.91 ± 0.22	92.68	83.55 ± 0.05	85.41 ± 0.08	81.51 ± 0.06	92.33
	Mol2D-ProA	82.11 ± 0.09	84.68 ± 0.10	79.29 ± 0.11	92.04	81.47 ± 0.05	83.95 ± 0.08	78.69 ± 0.03	91.46
	Mol2D-ProB	82.17 ± 0.13	84.85 ± 0.10	79.23 ± 0.18	92.02	81.59 ± 0.07	84.17 ± 0.07	78.75 ± 0.11	91.35
	MACCS-ProA	82.89 ± 0.25	85.00 ± 0.21	80.53 ± 0.34	90.86	82.24 ± 0.05	84.23 ± 0.08	80.04 ± 0.08	90.19
	MACCS-ProB	82.83 ± 0.27	85.02 ± 0.20	80.40 ± 0.33	90.93	82.09 ± 0.07	84.16 ± 0.11	79.81 ± 0.09	90.34
Separated	ECFP4	74.99 ± 0.03	76.21 ± 0.04	73.66 ± 0.09	85.08	75.28 ± 0.09	77.53 ± 0.12	72.76 ± 0.18	84.65
	Mol2D	74.09 ± 0.06	76.73 ± 0.07	71.17 ± 0.07	83.18	73.61 ± 0.07	76.99 ± 0.13	69.85 ± 0.16	82.33
	MACCS	72.83 ± 0.07	75.30 ± 0.09	70.12 ± 0.08	82.88	72.94 ± 0.08	76.45 ± 0.08	69.03 ± 0.18	83.30
	ProA	66.20 ± 0.01	72.51 ± 0.09	59.22 ± 0.13	72.57	66.21 ± 0.03	72.41 ± 0.20	59.34 ± 0.22	72.53
	ProB	66.21 ± 0.03	72.53 ± 0.08	59.20 ± 0.13	72.57	66.21 ± 0.03	72.44 ± 0.12	59.34 ± 0.15	72.53

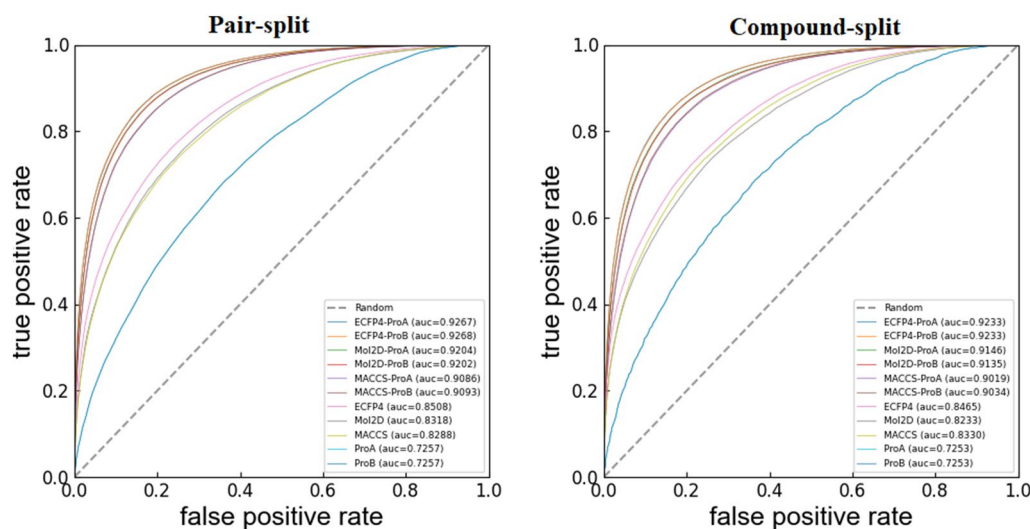


Fig. 3 ROC curves of models derived from different descriptors (integrated or separated groups) on the stratified tenfold CV

These results above indicated that our models built with the six integrated descriptor groups and XGBoost algorithm could effectively distinguish the compound-target interactions with strong binding affinity from those with weak binding affinity. Unsurprisingly and reasonably, the performance of the compound-split validation is slightly worse than that of the pair-split validation (e.g., Mol2d-prob model ACC: 82.11 vs. 81.47) since these two strategies simulate different situations that actual predictions may encounter where the former means the prediction of brand-new “new” compounds while the latter additionally includes the prediction of “known” compounds whose associated compound-target interaction(s) in the training set may provide prediction clues. The statistical values of the models built on the individual descriptor group were as follows in a decreasing order: ECFP4-ProA > ECFP4-ProB > Mol2D-ProA > Mol2D-ProB > MACCS-ProB > MACCS-ProA. The model utilizing the ECFP4-ProA descriptors yielded the best performance, with ACC = 0.832 and AUC = 0.913.

The chemogenomic approach, aiming at integrating the chemical space with the genomics space, is demonstrated to be strikingly helpful for representing compound-target associations. A demonstrable feature of our approach is that the information from compounds and targets were integrated to represent compound-target associations. We assume that compound-target interactions can be determined by the structural features from compounds and targets, which comprise of a pharmacological space. To demonstrate the reliability of our assumption, we re-established our model using only the structural information from a single space (i.e., chemical space or genomics space), that is, models are constructed only using the

compound features or protein features, respectively. The statistics of these models on the stratified tenfold CV were summarized in the “Separated” rows of Table 1. The ROC curves of the re-established models were shown in Fig. 3.

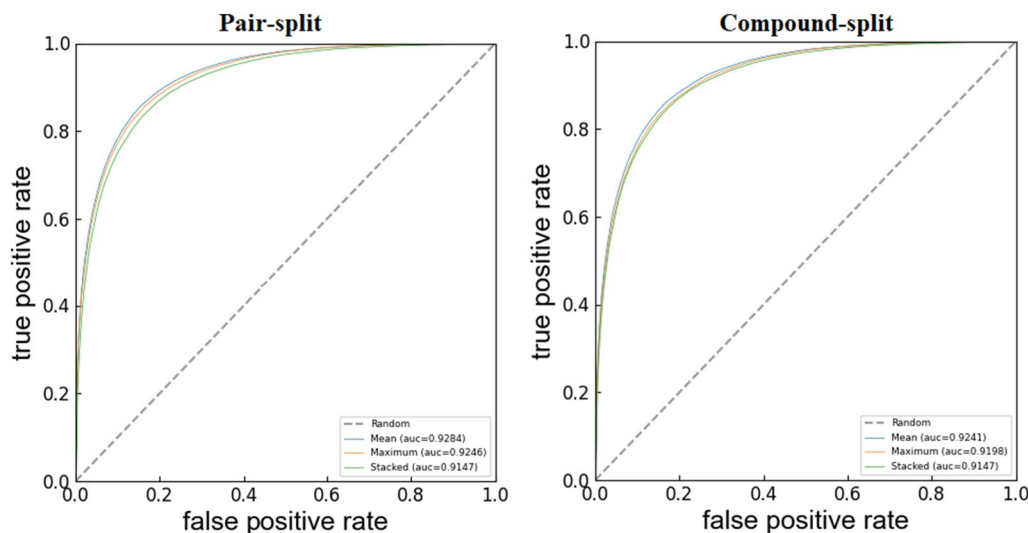
As can be seen from Table 1, the models with the compound features or protein features provided noticeably inferior predictions. The comparison between the models with the separated features and those with the integrated features sufficiently indicates that the structural information from compounds and targets contributes to the discrimination of compound-target associations cooperatively. Somewhat surprising, our comparison also illustrated that the features from compounds seem to be more predictive than those from target proteins.

The ensemble model performs well than individual models

Due to the different strengths in compound-target interaction prediction caused by different descriptor groups, we attempt to improve the prediction performance through their combination. We built three types of ensemble models by averaging the predictions given by the six individual models (Mean) [69], taking the maximum value given by the six individual models (Maximum) and obtaining new scores using the stacked models reported by Nicholas (Stacked) [70]. The performance statistics are summarized in Table 2 and the ROC curves are shown in Fig. 4. The result shows that the ensemble model (Mean) yielded better predictive ability than any individual model, with the improved ACC of 0.01–0.1 and AUC of 0.01–0.1. It appeared that it could capture the relationship between compound-target interaction patterns and the interaction endpoint more efficiently

Table 2 Prediction results of different ensemble models on the stratified tenfold CV

Methods	Pair-split				Compound-split			
	ACC	SE	SP	AUC	ACC	SE	SP	AUC
Mean	84.83 ± 0.16	86.96 ± 0.10	82.44 ± 0.24	92.84	84.41 ± 0.03	86.5 ± 0.04	82.13 ± 0.04	92.41
Maximum	80.73 ± 0.19	94.53 ± 0.08	65.39 ± 0.32	92.46	79.97 ± 0.05	94.63 ± 0.05	63.71 ± 0.08	91.98
Stacked	83.80 ± 0.23	85.07 ± 0.20	82.37 ± 0.30	91.47	82.93 ± 0.08	84.33 ± 0.08	81.40 ± 0.11	91.47

**Fig. 4** ROC curves of three ensemble models on the stratified tenfold CV

than any individual model. Therefore, the ensemble model (Mean) was used as our final model and applied for the subsequent analysis.

Evaluation of the target prediction performance of our model

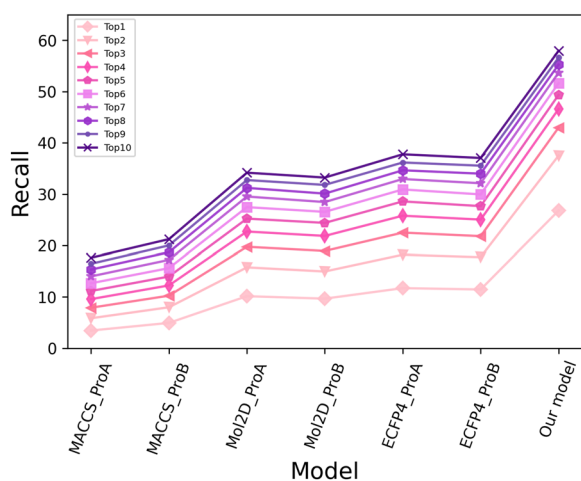
Under the premise of ensuring good classification performance of compound-target interactions, the target prediction performance of our model was then evaluated, which was the focus of our study that we attempted to verify whether our method could be expanded to the application of target prediction. For each compound to be predicted, a vector of 859 compound-target interaction scores could be outputted by our models and the targets with higher scores are considered as the target prediction result. Therefore, the target prediction performance was verified here using the recall rate, namely the fraction of the known targets identified in the top k of the prediction list. Undoubtedly, the performance improved with the increasing number of the picked targets. However, if the threshold of the number of selected targets is high, the number of the targets to be experimentally tested increases and thus the efficiency of the model application decreases. Inversely, if the threshold of the

number of the selected targets is low, many targets recognized as inactive might be actually active. For practicality, approximately the top 1–10 targets out of the total 859 targets are proposed as the candidate targets.

The target prediction result on the stratified tenfold CV was showed in Table 3. The average recall rates of the top-1 and top-10 metrics for pair-split validation datasets were 28.54% and 59.50%, respectively, implying that there are 28.54% and 59.50% of known targets were enriched to the top-1 and top-10 of the ranked list by our model. Given that predictions were made among the 859 possible human targets, these recall rates of the top-1 and top-10 metrics correspond to approximately the 245-fold [28.54%/(1/859)] and 51-fold [59.50%/(10/859)] enrichment compared to random picking, respectively [30]. As for the compound-split validation datasets, the average recall rates of the top-1 and top-10 metric were 26.78% and 57.96%, respectively, which refer to approximately the 230-fold [26.78%/(1/859)] and 50-fold [57.96%/(10/859)] enrichments [30]. By the way, the targets to be correctly predicted evenly distributed across different target classes, which recognized the unbiased prediction performance for different target classes. This indicated that our model based on the ensemble chemogenomic

Table 3 Recall rates of our model measured on the stratified tenfold CV datasets

	Top1	Top3	Top5	Top7	Top9	Top10
Pair-split	28.54 ± 1.22	43.92 ± 0.56	50.63 ± 0.46	55.00 ± 0.35	58.18 ± 0.28	59.50 ± 0.26
Compound-split	26.78 ± 0.12	42.80 ± 0.23	49.42 ± 0.22	53.59 ± 0.17	56.69 ± 0.17	57.96 ± 0.14

**Fig. 5** The recall rates of six individual models and our model within various top k values ($k = 1-10$) measured on the stratified 10-CV

approach could push true targets at the top of the ranking list and make some efforts to narrow down the potential targets to be tested.

To further verify whether our model had better target prediction performance than individual models based on various integrated descriptor groups, the prediction abilities of the individual models were prerecorded and compared with that of our model. As for the compound-split validation, the average recall rates of the top-k targets for different models on the stratified tenfold CV datasets, were plotted in Fig. 5. As shown in Fig. 5, the performance of the individual models was greatly inferior to that of our model. The recall value of each individual model for top 1 was lower than 20% even lower than 10%, while that of our model was 26.78%. The recall rate of top 10 for each individual model was lower than 40%, while that for our model was 57.96%. The recall values of the models in decreasing order were as follows: our model >> ECFP4_ProA > ECFP4_ProB > Mol2d_ProA > Mol2d_ProB > MACCS_ProB > MACCS_ProA, which further illustrated the robustness and predictivity of our model based on the ensemble chemogenomic approach for target prediction.

Target prediction performance for external test sets

To validate the generalization ability of our model on the external test dataset, we collected nonduplicated

Table 4 The target prediction results of the external test sets

Top k threshold	PDSP Ki (778)		NPASS (181)	
	Count	Recall (%)	Count	Recall (%)
Top1	147	18.89	16	8.84
Top3	257	33.03	49	27.07
Top5	322	41.39	68	37.57
Top7	363	46.66	72	39.78
Top10	415	53.34	82	45.30

compound-target interactions with K_i less than 100 nM from the PDSP Ki (Psychoactive Drugs Screening Program Ki Database, <https://pdsp.unc.edu/databases/kidb.php>) and NPASS databases (Natural Product Activity & Species Source Database) [71] to evaluate the ability of the model. After compound filtering and preprocessing, we finally obtained 442 compounds with 778 compound-target interactions from the PDSP Ki database and 122 compounds with 181 compound-target interactions from the NPASS database. The two test datasets include 94 and 113 proteins, respectively. The detailed information of validation data can be seen in Additional file 4.

The target prediction results were shown in Table 4. For the compounds from the PDSP Ki database, 147 targets (out of 778) were ranked at the top-1 of the predicted target list, with a recall rate of 18.89%. The NPASS dataset obtained a recall rate of 8.84%, indicating that 16 targets (out of 181) were successfully predicted in the top-1 list. The performance gap between these two datasets might be explained by the fact that the enough knowledge about natural products didn't be well learned by the model constructed by datasets mostly composed of synthetic compounds. However, whether for the PDSP Ki dataset or NPASS dataset, more than 45% targets were enriched in the top-10 of the predicted ranking list (a recall rates of 53.34% and 45.30% for PDSP Ki and NPASS for the top-10 prediction, respectively). Although the performance of these external datasets was fractionally inferior to that of the stratified tenfold CV, it highlighted the capability of our

Table 5 Comparison results with alternative state-of-the-art prediction methods

TopK	^a HitPickV2	^a PPB2	^a PPB				^a TargetNet	^a Our model	^b Our model	^b Swisstarget prediction			
			NB(ECfp4)	NN(ECfp4) + NB(ECfp4)	NN(MQN) + NB(ECfp4)	NN(Xfp) + NB(ECfp4)					NN(Xfp)		
1	24.69	16.49	14.89	16.59	21.87	10.65	21.49	16.49	5.18	23.20	26.96	57.00	28.00
3	56.74	35.06	52.31	52.88	52.40	22.43	52.40	30.91	18.85	41.85	56.36	-	-
5	58.43	47.03	60.92	57.96	57.21	26.96	60.30	35.34	25.82	46.37	59.33	-	-
7	60.82	53.35	62.76	61.29	60.04	30.16	61.30	39.21	29.78	48.91	60.89	-	-
10	62.20	60.98	64.75	63.62	63.05	34.68	62.58	45.62	34.40	50.99	63.99	-	-
15	-	-	-	-	-	-	-	-	-	-	-	76.00	72.00

^a Recall rate defined in our article (%); ^b the fraction of compounds for which at least one known target was identified in the top-1 or top-15 of the prediction lists (%)

model to enrich active targets for different sets of compounds, even for natural products.

Comparison with alternative approaches

Our model was compared with some state-of-the-art target prediction tools including SwissTargetPrediction (the updated 2019 version) [30], HitPickV2 [72], PPB2 [73], PPB [74] and TargetNet [36]. The comparison dataset was from the validation data from SwissTargetPrediction (from ChEMBL24) annotated as direct binders with the high activity (K_i , K_D , IC_{50} or EC_{50}) < 1 nM, associated with 1,061 ligand-target interactions. The ligands present in our model were firstly removed from the model to rebuild a new one in order to avoid potential bias. The recall rate defined in this study was used in the comparison between our ensemble model and four web tools. As the performance of SwissTargetPrediction on this dataset is public and its statistics metric is different from that of our research, our model was re-measured by the same statistics metric to ensure fair comparison, that is, the fraction of compounds for which at least one known target was identified in the top-1 or top-15 of the prediction lists. The comparison results are listed in Table 5, and related detailed performance of individual methods can be seen in Additional file 5.

The comparison results with HitPickV2, PPB2, PPB and TargetNet showed that our model performed better than any other method for the recall rate on top-1 predictions, including the popular HitPickV2 (Recall: 26.96% vs. 24.69%) and PPB2 method NN(MQN)+NB(ECfp4) (Recall: 26.96% vs. 14.89%). For the top-10 predictions, the performance of our model was better than those of all other models except PPB2 method NN(ECfp4)+NB(ECfp4) (Recall: 63.99% vs. 64.75%). The above results are very encouraging, especially since it is not clear whether the tested interaction pairs have been used in the construction of other models.

Comparison results with SwissTargetPrediction showed that for 360 molecules (72%), at least one of the experimentally known targets can be found among the predicted top-15 of SwissTargetPrediction, while for 379 molecules (76%), at least one of the experimentally known targets can be found among the predicted top-15 of our method. More importantly, our model detected at least one known target at top-1 prediction for 57.0% of ligands, with 28% for SwissTargetPrediction. These excellent results supported that our model is a strongly powerful target prediction engine to enrich active targets which may strongly bind/associate to compounds. It is expected to make some efforts for narrowing down the set of potential targets to be

experimentally tested and to be of interest to the audiences for wider scientific community.

Conclusion

Predicting targets of active compounds can augment modern drug discovery efforts in a range of applications, from the elaboration of molecular mechanisms and side-effect to the repurposing of existing drugs, and to designing novel drugs with lower toxicity and higher efficacy. However, identifying a direct target for active compounds remains a challenging task as a significant investment of time and resources is required for the experiments. Here, the chemogenomic modeling using the integrated features of compounds and proteins can be considered as a promising method for target identification.

We developed a model with the multi-scale information of chemical structures and protein sequences through the chemogenomic framework and the ensemble method to predict targets. It shows excellent target prediction statistics, which means to approximately 230-fold and 50-fold enrichment. The performance of our model was greatly superior to the individual models. When the model was validated by external datasets including natural products, more than 45% targets were enriched in the top-10 of the prediction list. Moreover, compared with multiple state-of-art target prediction methods, our model yielded equivalent or better predictive ability on the top-k predictions. In summary, the ensemble chemogenomic model constructed by us is expected to make some efforts for narrowing down the set of potential targets to be tested and speed up the process of the target identification.

For practicability, our model was public to facilitate predicting targets of interested compounds: <https://github.com/SorchaYang00/Chemogenomic-Model>. By inputting a compound, 859 targets can be hit and assigned by corresponding probabilities. Given the complexity of the interaction pattern between compounds and targets and derived numerous features/variables, neural network may provide progressive progress of the chemogenomic method for target prediction.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-023-00720-0>.

Additional file 1. Targets that can be predicted by our model.

Additional file 2. The datasets supplied for model construction.

Additional file 3. The detailed information of the descriptors and the percentage of explained variance of PCA.

Additional file 4. External test sets and corresponding validation results.

Additional file 5. Results and performance of our model and alternative state-of-the-art target prediction tools on the validation data from SwissTargetPrediction.

Acknowledgements

We acknowledge Haikun Xu and the High-Performance Computing Center of Central South University for support. The study was approved by the university's review board.

Author contributions

SQY and DSC conceived of the idea. SQY designed and implemented the methodology, analyzed the results, and drafted the manuscript. DSC and TJH supervised the current study. LXZ, YJG, JWZ, JXH, CYS and APL provide valuable suggestions and technical supports. All authors read and approved the final manuscript.

Funding

This work was supported by National Natural Science Foundation of China (22173118), National Key Research and Development Program of China (2021YFF1201400), Hunan Provincial Science Fund for Distinguished Young Scholars (2021JJ10068), the science and technology innovation Program of Hunan Province (2021RC4011), the Natural Science Foundation of Hunan Province (2022JJ80104), Changsha Science and Technology Bureau project (kq2001034), and The 2020 Guangdong Provincial Science and Technology Innovation Strategy Special Fund (2020B1212030006, Guangdong-Hong Kong-Macau Joint Lab). We acknowledge Haikun Xu, and the High-Performance Computing Center of Central South University for support. The study was approved by the university's review board.

Availability of data and materials

All the datasets supporting the conclusion of this article are available in Additional files.

Declarations

Competing interests

The authors declare neither competing financial interests nor non-financial competing interests.

Author details

¹Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, Hunan, People's Republic of China. ²Department of Pharmacy, Jiangxi Provincial People's Hospital, The First Affiliated Hospital of Nanchang Medical College, Nanchang 330006, Jiangxi, People's Republic of China. ³The First Hospital of Hunan University of Chinese Medicine, Changsha 410007, Hunan, People's Republic of China. ⁴Departments of Biomedical Engineering and Pathology, School of Basic Medical Science, Central South University, Changsha 410013, Hunan, People's Republic of China. ⁵Institute for Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, People's Republic of China. ⁶Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, People's Republic of China.

Received: 14 May 2022 Accepted: 8 April 2023

Published online: 23 April 2023

References

- Rautio J, Meanwell NA, Di L, Hageman MJ (2018) The expanding role of prodrugs in contemporary drug design and development. *Nat Rev Drug Discovery* 17(8):559–587
- Harrison RK (2016) Phase II and phase III failures: 2013–2015. *Nat Rev Drug Discovery* 15(12):817
- Peón A, Naulaerts S, Ballester PJ (2017) Predicting the reliability of drug-target interaction predictions with maximum coverage of target space. *Sci Rep* 7(1):1–11
- Houslay MD (2016) Melanoma, Viagra, and PDE5 inhibitors: proliferation and metastasis. *Trends in cancer* 2(4):163–165
- Reisberg B, Doody R, Stöffler A, Schmitt F, Ferris S, Möbius HJ (2003) Memantine in moderate-to-severe Alzheimer's disease. *N Engl J Med* 348(14):1333–1341
- Bodkin JA, Zornberg GL, Lukas SE, Cole JO (1995) Buprenorphine treatment of refractory depression. *J Clin Psychopharmacol* 15(1):49–57
- Tompson DJ, Vearer D (2007) Steady-state pharmacokinetic properties of a 24-hour prolonged-release formulation of ropinirole: results of two randomized studies in patients with Parkinson's disease. *Clin Ther* 29(12):2654–2666
- Eden R, Costall B, Domeneq A, Gerrard P, Harvey C, Kelly M, Naylor R, Owen D, Wright A (1991) Preclinical pharmacology of ropinirole (SK&F 101468-A) a novel dopamine D2 agonist. *Pharmacol Biochem Behav* 38(1):147–154
- Davidson MH, Dillon MA, Gordon B, Jones P, Samuels J, Weiss S, Isaacsohn J, Toth P, Burke SK (1999) Colesevelam hydrochloride (cholestagel): a new, potent bile acid sequestrant associated with a low incidence of gastrointestinal side effects. *Arch Intern Med* 159(16):1893–1900
- Gfeller D, Michielin O, Zoete V (2013) Shaping the interaction landscape of bioactive molecules. *Bioinformatics* 29(23):3073–3079
- Szardenings K, Li B, Ma L, Wu M (2004) Fishing for targets: novel approaches using small molecule baits. *Drug Discov Today Technol* 1(1):9–15
- Bantscheff M, Drewes G (2012) Chemoproteomic approaches to drug target identification and drug profiling. *Bioorg Med Chem* 20(6):1973–1978
- Lee J, Bogoy M (2013) Target deconvolution techniques in modern phenotypic profiling. *Curr Opin Chem Biol* 17(1):118–126
- Terstappen GC, Schlüpen C, Raggiaschi R, Gaviraghi G (2007) Target deconvolution strategies in drug discovery. *Nat Rev Drug Discovery* 6(11):891–903
- Rix U, Superti-Furga G (2009) Target profiling of small molecules by chemical proteomics. *Nat Chem Biol* 5(9):616–624
- Chen Z, Jiang Z, Chen N, Shi Q, Tong L, Kong F, Cheng X, Chen H, Wang C, Tang B (2018) Target discovery of ebselen with a biotinylated probe. *Chem Commun* 54(68):9506–9509
- Chen X, Wong YK, Wang J, Zhang J, Lee YM, Shen HM, Lin Q, Hua ZC (2017) Target identification with quantitative activity based protein profiling (ABPP). *Proteomics* 17(3–4):1600212
- Martell J, Weerapana E (2014) Applications of copper-catalyzed click chemistry in activity-based protein profiling. *Molecules* 19(2):1378–1393
- Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46:D1074–D1082
- Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 46(D1):D1121–D1127
- Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Pujadas G, Garcia-Vallve S (2015) Tools for in silico target fishing. *Methods* 71:98–103
- Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, IJzerman AP, van Westen GJP, Volkamer A (2019) Advances and challenges in computational target prediction. *J Chem Inf Model* 59(5):1728–1742
- Liu X, Xu Y, Li S, Wang Y, Peng J, Luo C, Luo X, Zheng M, Chen K, Jiang H (2014) In Silico target fishing: addressing a "Big Data" problem by ligand-based similarity rankings with data fusion. *J Cheminform* 6(1):1–14
- Wei H, Guan Y-D, Zhang L-X, Liu S, Lu A-P, Cheng Y, Cao D-S (2020) A combinatorial target screening strategy for deorphaning macromolecular targets of natural product. *Eur J Med Chem* 204:112644
- Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J et al (2006) TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* 34:W219–224
- Lee A, Lee K, Kim D (2016) Using reverse docking for target identification and its applications for drug discovery. *Expert Opin Drug Discov* 11(7):707–715
- Wang J-C, Chu P-Y, Chen C-M, Lin J-H (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust

- scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res* 40(W1):W393–W399
29. Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, Zheng S, Li Z, Li H, Jiang H (2010) PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res* 38:W609–614
 30. Daina A, Michielin O, Zoete V (2019) SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res* 47(W1):W357–W364
 31. Wang L, Ma C, Wipf P, Liu H, Su W, Xie X-Q (2013) TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J* 15(2):395–406
 32. Peon A, Li H, Ghislat G, Leung KS, Wong MH, Lu G, Ballester PJ (2019) MolTarPred: A web tool for comprehensive target prediction with reliability estimation. *Chem Biol Drug Des* 94(1):1390–1401
 33. Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, Xing J, Luo X, Jiang H, Zheng M (2015) TarPred: a web application for predicting therapeutic and side effect targets of chemical compounds. *Bioinformatics* 31(12):2049–2051
 34. Kinnings SL, Jackson RM (2011) ReverseScreen3D: a structure-based ligand matching method to identify protein targets. *J Chem Inf Model* 51(3):624–634
 35. Nidhi Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46(3):1124–1133
 36. Yao ZJ, Dong J, Che YJ, Zhu MF, Wen M, Wang NN, Wang S, Lu AP, Cao DS (2016) TargetNet: a web service for predicting potential drug-target interaction profiling via multi-target SAR models. *J Comput Aided Mol Des* 30(5):413–424
 37. Dahl GE, Jaitly N, Salakhutdinov R (2014) Multi-task neural networks for QSAR predictions. *Arxiv Prepr*. <https://doi.org/10.48550/arXiv.1406.1231>
 38. Lee K, Lee M, Kim D (2017) Utilizing random forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinformatics* 18(Suppl 16):567
 39. Klabunde T (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol* 152(1):5–7
 40. Ezzat A, Wu M, Li X-L, Kwok C-K (2019) Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 20(4):1337–1357
 41. Mousavian Z, Masoudi-Nejad A (2014) Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin Drug Metab Toxicol* 10(9):1273–1287
 42. Cao D-S, Liang Y-Z, Deng Z, Hu Q-N, He M, Xu Q-S, Zhou G-H, Zhang L-X, Deng Z-x, Liu S (2013) Genome-scale screening of drug-target associations relevant to Ki using a chemogenomics approach. *PLoS ONE* 8(4):e57680
 43. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, Li X, Zhou W, Wang W, Wang Y (2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE* 7(5):e37608
 44. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H (2017) Deep-learning-based drug–target interaction prediction. *J Proteome Res* 16(4):1401–1409
 45. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42(D1):D1083–D1090
 46. Consortium U (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(D1):D204–D212
 47. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 35:D198–D201
 48. Dong J, Yao Z-J, Zhang L, Luo F, Lin Q, Lu A-P, Chen AF, Cao D-S (2018) PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *J Cheminform* 10(1):1–11
 49. Dong J, Zhu M-F, Yun Y-H, Lu A-P, Hou T-J, Cao D-S (2021) BioMedR: an R/CRAN package for integrated data analysis pipeline in biomedical study. *Brief Bioinform* 22(1):474–484
 50. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754
 51. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 42(6):1273–1280
 52. Vilar S, Cozza G, Moro S (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* 8(18):1555–1572
 53. Dong J, Cao D-S, Miao H-Y, Liu S, Deng B-C, Yun Y-H, Wang N-N, Lu A-P, Zeng W-B, Chen AF (2015) ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J Cheminform* 7(1):1–10
 54. Cao D-S, Xu Q-S, Hu Q-N, Liang Y-Z (2013) ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29(8):1092–1094
 55. Cao DS, Liang YZ, Yan J, Tan GS, Xu QS, Liu S (2013) PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J Chem Inf Model* 53(11):3086–3096
 56. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2(1–3):37–52
 57. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S (2010) GOsemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26(7):976–978
 58. Pages H, Aboyoum P, Gentleman R, DebRoy S (2016) Biostrings: string objects representing biological sequences, and matching algorithms. R Package Version 2:1018129
 59. Cao D-S, Liu S, Xu Q-S, Lu H-M, Huang J-H, Hu Q-N, Liang Y-Z (2012) Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Anal Chim Acta* 752:1–10
 60. Cao D-S, Zhou G-H, Liu S, Zhang L-X, Xu Q-S, He M, Liang Y-Z (2013) Large-scale prediction of human kinase–inhibitor interactions using protein sequences and molecular topological structures. *Anal Chim Acta* 792:10–18
 61. Cao DS, Zhang LX, Tan GS, Xiang Z, Zeng WB, Xu QS, Chen AF (2014) Computational prediction of drug target interactions using chemical, biological, and network features. *Mol Inf* 33(10):669–681
 62. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM (2016) Extreme gradient boosting as a method for quantitative structure–activity relationships. *J Chem Inf Model* 56(12):2353–2360
 63. Babajide Mustapha I, Saeed F (2016) Bioactive molecule prediction using extreme gradient boosting. *Molecules* 21(8):983
 64. Lei T, Sun H, Kang Y, Zhu F, Liu H, Zhou W, Wang Z, Li D, Li Y, Hou T (2017) ADMET evaluation in drug discovery. 18. Reliable prediction of chemical-induced urinary tract toxicity by boosting machine learning approaches. *Mol Pharm* 14(11):3935–3953
 65. Lei T, Chen F, Liu H, Sun H, Kang Y, Li D, Li Y, Hou T (2017) ADMET evaluation in drug discovery. Part 17: development of quantitative and qualitative prediction models for chemical-induced respiratory toxicity. *Mol Pharm* 14(7):2407–2421
 66. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Statist* 29:1189–1232
 67. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME—the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explor Newsl* 11(1):26–31
 68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
 69. Lei B, Li J, Yao X (2013) A novel strategy of structural similarity based consensus modeling. *Mol Inform* 32(7):599–608
 70. Cockroft NT, Cheng X, Fuchs JR (2019) STarFish: a stacked ensemble target fishing approach and its application to natural products. *J Chem Inf Model* 59(11):4906–4920
 71. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, Wang Y, Tan Y, Gao D, Wang B (2018) NPASS: natural product activity and species source

database for natural product research, discovery and tool development. *Nucleic Acids Res* 46(D1):D1217–D1222

72. Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, Campillos M (2019) HitPickV2: a web server to predict targets of chemical compounds. *Bioinformatics* 35(7):1239–1240
73. Awale M, Reymond J-L (2018) Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. *J Chem Inf Model* 59(1):10–17
74. Awale M, Reymond JL (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J Cheminform* 9:11

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

