


METHODOLOGY ARTICLE

Open Access



# MOST: most-similar ligand based approach to target prediction

Tao Huang<sup>1†</sup>, Hong Mi<sup>1,2†</sup>, Cheng-yuan Lin<sup>1,3</sup>, Ling Zhao<sup>1</sup>, Linda L. D. Zhong<sup>1,4</sup>, Feng-bin Liu<sup>2</sup>, Ge Zhang<sup>1</sup>, Ai-ping Lu<sup>1,4</sup>, Zhao-xiang Bian<sup>1,4\*</sup>  and for MZRW Group

## Abstract

**Background:** Many computational approaches have been used for target prediction, including machine learning, reverse docking, bioactivity spectra analysis, and chemical similarity searching. Recent studies have suggested that chemical similarity searching may be driven by the most-similar ligand. However, the extent of bioactivity of most-similar ligands has been oversimplified or even neglected in these studies, and this has impaired the prediction power.

**Results:** Here we propose the **MOst-Similar ligand-based Target inference** approach, namely **MOST**, which uses fingerprint similarity and explicit bioactivity of the most-similar ligands to predict targets of the query compound. Performance of MOST was evaluated by using combinations of different fingerprint schemes, machine learning methods, and bioactivity representations. In sevenfold cross-validation with a benchmark Ki dataset from ChEMBL release 19 containing 61,937 bioactivity data of 173 human targets, MOST achieved high average prediction accuracy (0.95 for  $pKi \geq 5$ , and 0.87 for  $pKi \geq 6$ ). Morgan fingerprint was shown to be slightly better than FP2. Logistic Regression and Random Forest methods performed better than Naïve Bayes. In a temporal validation, the Ki dataset from ChEMBL19 were used to train models and predict the bioactivity of newly deposited ligands in ChEMBL20. MOST also performed well with high accuracy (0.90 for  $pKi \geq 5$ , and 0.76 for  $pKi \geq 6$ ), when Logistic Regression and Morgan fingerprint were employed. Furthermore, the  $p$  values associated with explicit bioactivity were found to be a robust index for removing false positive predictions. Implicit bioactivity did not offer this capability. Finally,  $p$  values generated with Logistic Regression, Morgan fingerprint and explicit activity were integrated with a false discovery rate (FDR) control procedure to reduce false positives in multiple-target prediction scenario, and the success of this strategy was demonstrated with a case of fluanisone. In the case of aloemodin's laxative effect, MOST predicted that acetylcholinesterase was the mechanism-of-action target; in vivo studies validated this prediction.

**Conclusions:** Using the MOST approach can result in highly accurate and robust target prediction. Integrated with a FDR control procedure, MOST provides a reliable framework for multiple-target inference. It has prospective applications in drug repurposing and mechanism-of-action target prediction.

**Keywords:** Explicit bioactivity, False discovery rate, Logistic regression, Mechanism-of-action target, Most-similar ligand, Target prediction

\* Correspondence: bianzxiang@gmail.com

†Equal contributors

<sup>1</sup>Lab of Brain and Gut Research, School of Chinese Medicine, Hong Kong Baptist University, 7 Baptist University Road, Hong Kong, People's Republic of China

<sup>4</sup>Hong Kong Chinese Medicine Clinical Study Centre, Hong Kong Baptist University, 7 Baptist University Road, Hong Kong, People's Republic of China  
Full list of author information is available at the end of the article

## Background

Target identification is key to understanding the mechanism-of-action of active compounds discovered from phenotypic screening or found in traditional herbal medicines. Various experimental methods, including affinity chromatography, drug affinity responsive target stability, and proteomics have been used for target identification [1]. However, these experimental approaches are laborious, expensive, and often unsuccessful. In contrast, computational target identification (also called “target prediction” or “target inference”) approaches are inexpensive, and effective. It is readily integrated with experimental validation, and can quickly narrow down potential targets to a handful of most likely candidates. A number of computational tools are available for target prediction [2]; they can be classified by algorithms into four major classes, namely, machine learning, inverse docking, bioactivity spectra analysis, and chemical similarity searching; the merits and flaws of each approach can be found elsewhere [3]. In this study, we will focus on chemical similarity searching.

Chemical similarity searching is based on the observation by medicinal chemists that structurally similar compounds usually have similar biological activities [4]. In practice, compounds are represented by two-dimensional (2D) fingerprints, and the similarity can be measured by Tanimoto coefficient ( $T_c$ ) metrics [5]. Fingerprint-based similarity searching is widely used for target prediction. By fitting distribution of similarity between different ligand sets with extreme distribution, Keiser et al. developed the Similarity Ensemble Approach (SEA) to quantitatively calculate the correlation between different targets [6]. SEA has been successfully used in predicting new targets [7] and off-targets associated with side effects [8] of existing drugs. Similarity can also be measured by molecular shapes. For instance, Armstrong et al. proposed three-dimensional (3D) descriptors incorporating shape, chirality and charges to compare chemicals [9, 10]. One merit of molecular shape is that it can be used to detect similarity between structurally unrelated compounds, which is impossible for fingerprints. Work has been done to combine fingerprint similarity (2D) with shape similarity (3D) to improve target prediction [11, 12]. More recently, the chemical similarity network was used for target inference based on global comparison [13]. Discovering binders of a new structural class by chemical similarity searching is difficult because this approach requires high similarity with known ligands to make predictions.

Indeed, fingerprint-based similarity searching approaches have performed well in terms of accuracy and speed according to various benchmark tests of target prediction. Recently, the results of several studies imply that the most-similar counterpart of the target drives high predictive accuracy of fingerprint-based similarity searching

[12, 14]. Despite these advances, the explicit bioactivity data of the most-similar ligand were oversimplified as implicit values like “active” or “inactive”. Our insight is that, if the query compound has the same similarity as two most-similar ligands belonging to targets A and B, then the probabilities of query compound being active on target A or B should not be equal. Instead, the more potent known ligand should suggest better probability. To verify this insight, we will investigate, the effects on prediction performance by explicit or implicit bioactivity in current study.

Finally, we describe a method in which we use the fingerprint similarity and explicit bioactivity data of ligands most-similar to query compounds to make inferences about their targets. We name this method “MOST”, representing “**MO**st-Similar ligand-based **T**arget inference”. MOST showed high prediction accuracy with a reduced false positive rate.

## Methods

### Generation of $K_i$ dataset from ChEMBL database

The bioactivity data of all human targets in ChEMBL release 19 and 20 [15] were downloaded via an in-house script written in Python. The direct binding (confidence score=“9”) bioactivity data with type “ $K_i$ ” of each target were extracted and processed. Bioactivity data with unspecified concentration/activity values, unspecified concentration/activity units, unspecified references, and ambiguous operators were classified as “ineffective” and excluded. For multiple records for one target-ligand pair, if the  $K_i$  values were from the same publication, the smallest  $K_i$  (i.e. highest  $pK_i$ ) value was taken to reflect experimental optimization and/or remove unclear stereoisomer annotations [16]. After this step, if there were still multiple measurements for one target-ligand pair, which were from different publications (tested in the same or different labs), the mean  $K_i$  values were taken.

### Generation of sevenfold cross-validation dataset

Bioactivity data from preprocessed ChEMBL19 were used to generate cross-validation datasets. Targets which had less than 10 ligands and more than 10,000 ligands were filtered out. To make consistent comparison, only targets occurring in both ChEMBL19 and ChEMBL20 were kept. Finally, the benchmark  $K_i$  dataset was comprised of 173 targets annotated with 61,937 bioactivity data (Additional file 1: Table S1). These targets were covered by major drug target types, including 79 receptors, 60 enzymes, and 12 transporters (Additional file 1: Figure S1A). These targets were annotated with different number of ligands: 71 targets had 10–100 ligands; 82 targets had 100–1,000 ligands; and 20 targets had 1,000–10,000 ligands (Additional

file 1: Figure S1B). The annotated ligands were further categorized into three classes by their  $K_i$  values. Percentages of ligands of with  $K_i$  values less than 1  $\mu\text{M}$ , between 1- and 10  $\mu\text{M}$ , and greater than 10  $\mu\text{M}$  ligands were 76.7%, 17.2%, and 6.2%, respectively (Additional file 1: Figure S1C). However, such proportions were varied for specific targets (Additional file 1: Figure S1C).

15% of the ligands of each target were randomly selected to comprise the test set, and the rest were treated as the training set. This procedure was repeated seven times to make sure that the whole dataset was sampled [14]. In the training set, each ligand was selected and the most-similar ligand was generated by comparing the selected ligand with remaining ligands. In the test set, the most-similar ligand was acquired by comparing the query compound with ligand sets in the training set. To see how similar query compounds were with their most-similar counterparts, the distribution of  $Tc_{\text{most}}$  was determined as shown in Additional file 1: Figure S2A. A large fraction (85.4%) of query compounds had very similar ( $Tc_{\text{most}} \geq 0.8$ ) ligands, both in the training and test sets. Scatter plots of  $pKi_{\text{query}}$  vs  $pKi_{\text{most}}$  clearly showed that the principle, “structurally similar chemicals have similar bioactivities” [17], applies to the benchmark dataset; although there are also exceptions—some potent compounds had similar but weak binder counterparts, which is consistent with previous observations [18]. Moreover, the calculated Pearson correlation coefficient showed that structurally similar chemicals have more close bioactivities (Additional file 1: Figure S2B), suggesting that  $pKi_{\text{most}}$  may also be a strong predictor for the activity of a query compound.

Two threshold values of  $pKi$  were applied to label a ligand is “active” (represented by 1) or “inactive” (represented by 0) to a target. When  $pKi \geq 5$  was applied, 93.8% were categorized as “active”, while 6.2% were “inactive”. If  $pKi \geq 6$  was applied, 76.7% were labeled as “active”, while 23.3% were “inactive” (Additional file 1: Table S1). Inactive data were included in model training and testing since evidence has shown that negative information can improve the prediction performance [19].

#### Generation of temporal validation dataset

The whole  $Ki$  dataset from ChEMBL19 was used as training set for temporal validation. By comparing with ChEMBL19, which was released in 2014, the newly added bioactivity data in ChEMBL20 (released in 2015) was identified and used to generate the test set. In total, there were 173 targets annotated with 3,754  $Ki$  data. In this dataset, when  $pKi \geq 5$  was applied, 91.3% were categorized as “active”, while 8.7% were “inactive”. If  $pKi \geq 6$  was applied, 75.7% were labeled as “active”, while 24.3% were “inactive” (Additional file 1: Table S1).

#### Calculation of fingerprint and similarity

Two fingerprint schemes were used in this study—ECFP-4-like Morgan (radius = 2) [20] calculated by RDKit [21] and FP2 calculated by OpenBabel [22]. Once fingerprints were derived, the similarity between compound pairs was calculated by Tanimoto coefficient ( $Tc$ ) [5].

#### Machine learning methods

Machine learning models, including Naïve Bayes [23], Logistic Regression [24], and Random Forests [25], were used in this study for comparison. The probability to be active ( $p_a$ ) as calculated by Naïve Bayes (Eq. 1) or by Logistic Regression (Eq. 2) is expressed as follows:

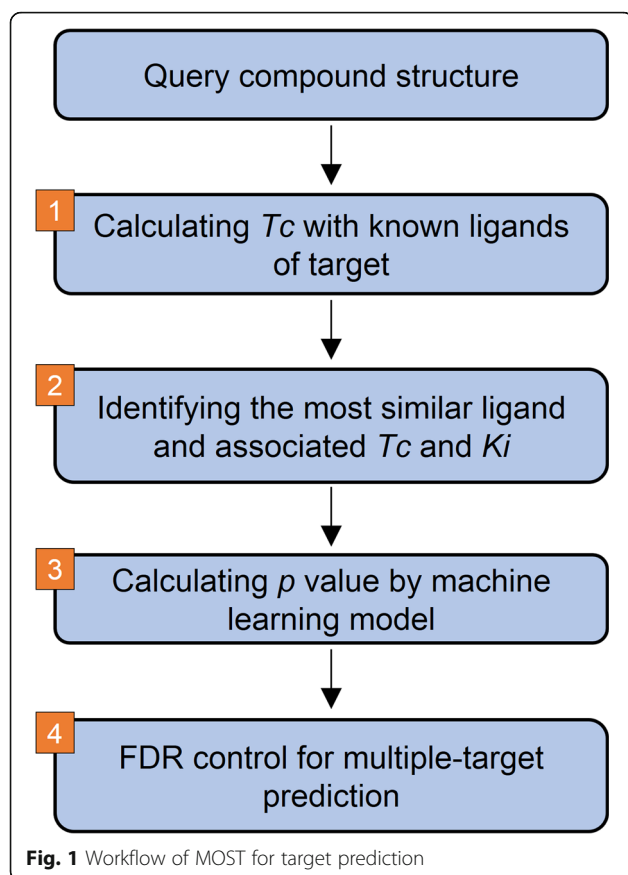
$$\begin{aligned} p(y = \text{active} | Tc_{\text{most}}, pKi_{\text{most}}) &= \frac{p(Tc_{\text{most}}, pKi_{\text{most}} | y = \text{active}) p(y = \text{active})}{p(Tc_{\text{most}}, pKi_{\text{most}})} \end{aligned} \quad (1)$$

$$\begin{aligned} p(y = \text{active} | Tc_{\text{most}}, pKi_{\text{most}}) &= \frac{1}{1 + e^{-(a_0 + a_1 Tc_{\text{most}} + a_2 pKi_{\text{most}})}} \end{aligned} \quad (2)$$

where  $Tc_{\text{most}}$  is the similarity between query compound and the most-similar ligand, while  $pKi_{\text{most}}$  is the activity of the most-similar ligand. The sum of probabilities to be active ( $p_a$ ) and inactive ( $p_i$ ) always equal to 1. Fitting Naïve Bayes, Logistic Regression, and Random Forests models were realized by a machine learning package in scikit-learn [26].

#### Workflow of MOST

The workflow adopted by MOST to make predictions for a query compound with reference to a series of targets is depicted in Fig. 1. Firstly, the  $Tc$  values between the query compound and annotated ligands of target are calculated. Secondly, the most-similar ligand is identified by ranking the  $Tc$  values. Thirdly, the  $Tc$  and  $pKi$  of the most-similar ligand ( $Tc_{\text{most}}$  and  $pKi_{\text{most}}$ ) are fed into a trained model to generate probabilities ( $p$  value) measuring how likely it is that the query compound is inactive. If explicit activity is used, the  $pKi_{\text{most}}$  is used “as-it-is” in model training and testing. If implicit activity is used,  $pKi_{\text{most}} \geq 5$  or 6 is then simplified as 1, and  $pKi_{\text{most}} < 5$  or 6 is simplified as 0. Once the probabilities have been generated by machine learning models, if  $p_a > p_i$ , the query compound is predicted to be active; otherwise, it is considered to be inactive. The probability to be inactive,  $p_i$ , is treated as  $p$  values in MOST for FDR control. If multiple-target predictions are made spontaneously, the FDR procedure is implemented to control the risk of false positives.



### Performance evaluation

The performance of MOST was evaluated by calculations of accuracy and Mathews Correlation Coefficient (MCC) [27, 28], according to the following equations:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where TP is true positives, TN is true negatives, FN is false negatives, and FP is false positives.

### FDR control procedure

FDR control was implemented by the ‘p\_adjust’ method in the ‘stats’ library of the R package (version 3.1.2) for the Benjamini-Hochberg [29] algorithm or by the ‘q value’ method in the ‘bioconductor’ library of the R package for the Storey-Tibshirani [30] algorithm.

### Animals and fecal pellet output

Male C57BL/6 J mice weighing around 22 g (6-week old) were purchased from the Laboratory Animal Services Center, The Chinese University of Hong Kong, Hong

Kong. The mice were fed with a standard rodent diet ad libitum with free access to water and were housed in rooms maintained at  $22 \pm 1$  °C with a 12 h light/dark cycle (lights on 6:00–18:00).

Mice were randomly divided into 6 groups with 10 mice per group. Saline and aloe-emodin (3.75 mg/kg, 7.5 mg/kg, 15 mg/kg, 30 mg/kg, and 60 mg/kg) were intragastrically administrated at 9:00 a.m. All the mice were placed in individual cages without water or food. The fecal pellets for each mouse were recorded continuously for 2 h. In a parallel study, mice were randomly divided into 4 groups with 10 mice for each group. Saline and atropine (2 mg/kg and 4 mg/kg) were intragastrically administrated 20 min before aloe-emodin (15 mg/kg) treatment. Then mice were placed in individual cages without water or food, and fecal pellets for each mouse were collected within 2 h.

## Results

### Performance of MOST in sevenfold cross-validation

To evaluate the performance of MOST, three factors were evaluated in a combinational way. These factors are (1) machine learning methods (Naïve Bayes, Logistic Regression or Random Forests), (2) fingerprint schemes (Morgan or FP2), and (3) representation of bioactivity of the most-similar ligand (explicit or implicit). Accuracy and MCC under different conditions are summarized (Table 1). Firstly, Logistic Regression or Random Forests performed better than Naïve Bayes in almost all cases in terms of average accuracy and MCC; there were only marginal differences between Logistic Regression and Random Forests. Secondly, Morgan fingerprint was slightly better than FP2 in most cases. Thirdly, explicit  $pKi$  were as good as implicit  $pKi$  in terms of average accuracy. The best performance of MOST were achieved when Logistic Regression/Random Forests and Morgan fingerprint were used. For active data defined by  $pKi \geq 5$ , the average accuracies were about 0.95, with MCC ranging from 0.50 to 0.59. While for active data defined by  $pKi \geq 6$ , the average accuracies were about 0.87, with MCC ranging from 0.61 to 0.63.

### Performance of MOST in temporal validation

To avoid overestimation of model quality with cross-validation, we used a temporal dataset to evaluate the performance of MOST. Newly added  $Ki$  data in ChEMBL20 were predicted by MOST trained with  $Ki$  data in ChEMBL19 and the results are summarized (Table 2). In general, the performance of MOST was slightly worse in temporal validation than in cross-validation. Logistic Regression outperformed Random Forests and Naïve Bayes in this temporal validation. Morgan fingerprints were better than FP fingerprints. Similar with the cross-validation results, models with explicit- and implicit  $pKi$

**Table 1** Overall performance of MOST in sevenfold cross-validation

Active data defined by	Performance						
$pKi \geq 5$	Accuracy						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit Ki	0.927 ± 0.003	0.929 ± 0.003	0.948 ± 0.003	0.949 ± 0.002	0.948 ± 0.001	0.946 ± 0.004
	Implicit Ki	0.939 ± 0.002	0.937 ± 0.003	0.948 ± 0.003	0.949 ± 0.003	0.950 ± 0.002	0.949 ± 0.003
	MCC						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit Ki	0.352 ± 0.013	0.345 ± 0.017	0.495 ± 0.010	0.484 ± 0.025	0.530 ± 0.017	0.504 ± 0.032
	Implicit Ki	0.554 ± 0.008	0.543 ± 0.022	0.558 ± 0.013	0.585 ± 0.023	0.585 ± 0.014	0.560 ± 0.023
$pKi \geq 6$	Accuracy						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit Ki	0.848 ± 0.004	0.842 ± 0.002	0.866 ± 0.002	0.862 ± 0.002	0.861 ± 0.004	0.853 ± 0.004
	Implicit Ki	0.860 ± 0.004	0.855 ± 0.004	0.867 ± 0.003	0.863 ± 0.003	0.867 ± 0.004	0.862 ± 0.004
	MCC						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit Ki	0.561 ± 0.014	0.540 ± 0.009	0.617 ± 0.009	0.602 ± 0.007	0.609 ± 0.013	0.581 ± 0.013
	Implicit Ki	0.624 ± 0.011	0.610 ± 0.612	0.632 ± 0.012	0.618 ± 0.011	0.632 ± 0.013	0.618 ± 0.012

**Table 2** Overall performance of MOST in temporal validation

Active data defined by	Performance						
$pKi \geq 5$	Accuracy						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit $pKi$	0.750	0.670	0.905	0.901	0.893	0.871
	Implicit $pKi$	0.741	0.696	0.896	0.894	0.896	0.897
	MCC						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit $pKi$	0.275	0.110	0.272	0.184	0.283	0.136
	Implicit $pKi$	0.267	0.138	0.292	0.213	0.256	0.192
$pKi \geq 6$	Accuracy						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit $pKi$	0.633	0.554	0.755	0.736	0.724	0.709
	Implicit $pKi$	0.632	0.556	0.761	0.737	0.759	0.726
	MCC						
		Naïve Bayes		Logistic Regression		Random Forests	
		Morgan	FP2	Morgan	FP2	Morgan	FP2
	Explicit $pKi$	0.357	0.225	0.382	0.321	0.319	0.273
	Implicit $pKi$	0.334	0.212	0.370	0.307	0.381	0.300

had almost the same accuracy. The best performance was achieved when Logistic Regression and Morgan fingerprint were employed. The average accuracy was about 0.90 with MCC ranging from 0.27 to 0.29, when active data was defined by  $pK_i \geq 5$ , and it was about 0.76 with MCC ranging from 0.37 to 0.38, when active data was defined by  $pK_i \geq 6$ .

#### Benefits of using explicit $pK_i$

To investigate the differences resulting from explicit- and implicit activity modes, one of the sevenfold cross-validation results was analyzed (Additional file 1: Table S2). Logistic Regression and Morgan fingerprint were chosen because they achieved the best performance. It was clear that more positive predictions (for both TPs and FPs) were made by explicit activity mode, compared with implicit activity mode. When all predictions were displayed in a scatter plot of  $pK_{i_{\text{most}}}$  vs  $Tc_{\text{most}}$ , many FPs were found to be predicted by most-similar ligand with weak affinity, and using explicit activity enhanced this tendency (Fig. 2a and b, left panels).

Former study suggested that setting the lower threshold ( $k$ ) of  $Tc$  could reduce FPs [14]. Thus we calculated the fraction of data ( $f$ ) when  $Tc \geq k$ , while the difference between  $f_{\text{TP}}$  and  $f_{\text{FP}}$  was used as a trade-off index. Ideally, a best  $k$  means keeping as many TPs as possible and as few FPs as possible at the same time, which is, the maximum difference between  $f_{\text{TP}}$  and  $f_{\text{FP}}$ . The  $f_{\text{TP}}$  started to fall when  $Tc \geq 0.4$ , indicating that  $Tc \geq 0.4$  was a minimum requirement for removing substantially unrelated compound pairs (Fig. 2a and b, middle panels). The difference between  $f_{\text{TP}}$  and  $f_{\text{FP}}$  reached a maximum when  $Tc \geq 0.85$  in both explicit- and implicit bioactivity modes. However, the extent of difference ( $f_{\text{TP}} - f_{\text{FP}}$ ) at this point was only about 0.2, suggesting that increasing the  $Tc$  threshold may not be a robust way to reduce FPs predicted by MOST.

We then wondered if setting the  $p$  value threshold could be a better way to reduce FPs without losing too much TPs. A decreased  $p$  value threshold led to rapidly decreased  $f_{\text{FP}}$  and slowly decreased  $f_{\text{TP}}$ , which was only observed with explicit-, but not implicit bioactivity mode (Fig. 2a and b, right panels). Moreover, the maximum difference between  $f_{\text{TP}}$  and  $f_{\text{FP}}$  occurred when  $p \leq 0.1$ : 0.60 for active data defined by  $pK_i \geq 5$ , or 0.40 for active data defined by  $pK_i \geq 6$ . These results suggested that setting the upper threshold of  $p$  value in explicit bioactivity mode was a better way to reduce FPs than  $Tc$ .

#### Multiple-target prediction by MOST integrated with FDR control

One important application of MOST is to predict novel targets of known drugs, which is key to repurposing drugs and inferring side effects. In such cases, the drug

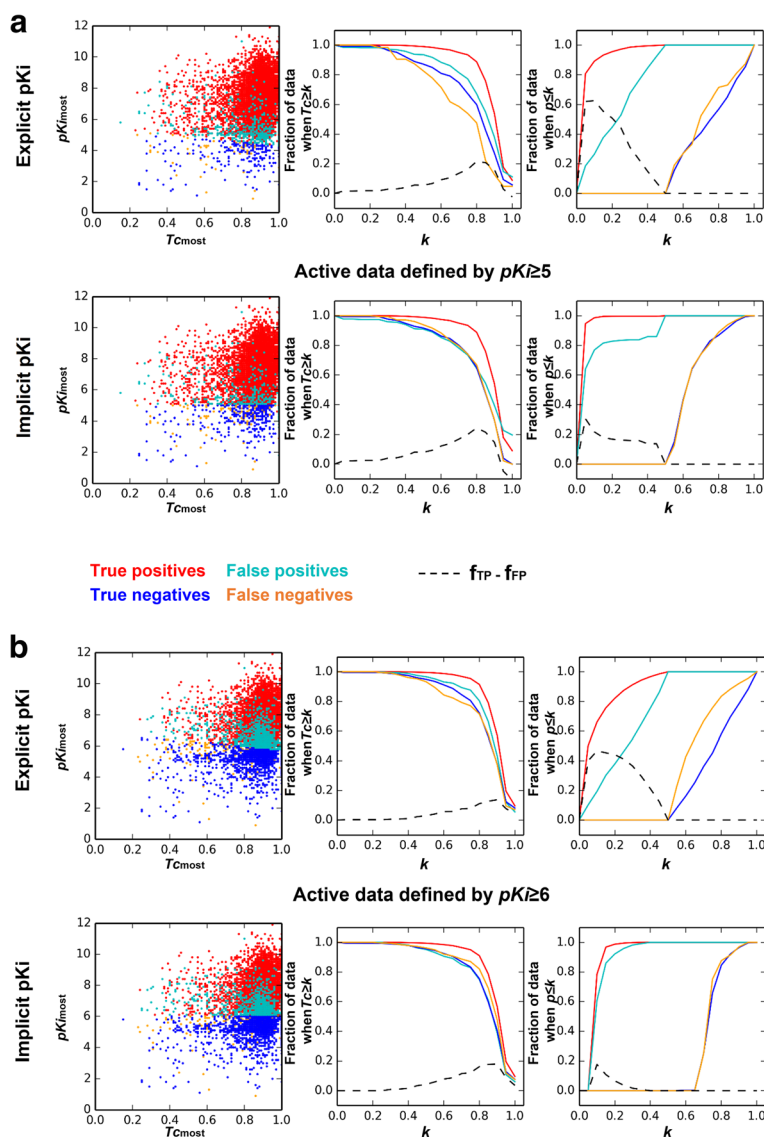
will be compared with known ligands of thousands of human targets. Encouraged by the benchmark results, we evaluated MOST in multiple-target prediction, where the query compound was searched against 1,439 human targets. To avoid too many false positive predictions, FDR control procedures were introduced to correct the  $p$  values generated by the Logistic Regression model (Fig. 3a).

Bioactivity data of some drugs can be found in references but not the ChEMBL database, which gave us the opportunity to test if MOST can predict novel targets of such drugs in a multi-target prediction scenario. We used fluanisone, an antipsychotic and sedative drug approved for schizophrenia [31], as an example to illustrate how MOST can be used to predict novel targets for approved drugs (Fig. 3b). The  $p$  values of fluanisone against 1,439 human targets were calculated by Logistic Regression model with Morgan fingerprint trained by the ChEMBL 20 benchmark dataset in explicit bioactivity mode. The distribution of  $p$  values suggested that either the Benjamini-Hochberg or Storey-Tibshirani methods are suitable for correction (Fig. 3c). The predicted targets were ranked by adjusted  $p$  values and  $Tc_{\text{most}}$ . Among the top 5 predicted targets, adrenoceptor alpha 1B (ADRA1B) and adrenoceptor alpha 1D (ADRA1D) are known human targets of fluanisone ( $pK_i$  equals to 7.85 and 8.15, respectively), documented in literature [7] but not in ChEMBL database. ADRA1B and ADRA1D were ranked as the 2nd and 3rd targets according to adjusted  $p$  values and  $Tc_{\text{most}}$  (Fig. 3d). Fluanisone was related to the two targets because it was quite similar ( $Tc = 0.70$ ) to a common ligand (Fig. 3e), ChEMBL8618, which potently acts on ADRA1B ( $pK_i = 9.02$ ) and ADRA1D ( $pK_i = 9.52$ ). MOST assigned low  $p$  values to both targets ( $3.4E-04$  and  $8.2E-04$ ) and made them top hits.

#### Investigating mechanism-of-action target of aloemodin for laxative effect with MOST

Another important application of MOST is to predict the mechanism-of-action targets of active compounds discovered from phenotypic screening and traditional medicine. We used the laxative aloemodin to illustrate how MOST can be used to predict mechanism-of-action targets.

Aloemodin belongs to anthraquinone, a class of chemicals commonly found in the Traditional Chinese Medicine (TCM) herbs Aloe vera and Rhubarb. Aloemodin is found to have antibacterial, antiviral, hepatoprotective, anticancer, and anti-inflammation effects [32]. More interestingly, aloemodin has a laxative effect, which is in line with the traditional TCM use of Rhubarb as a laxative; however, the mechanism-of-action target of neither the herb nor aloemodin is not fully understood.

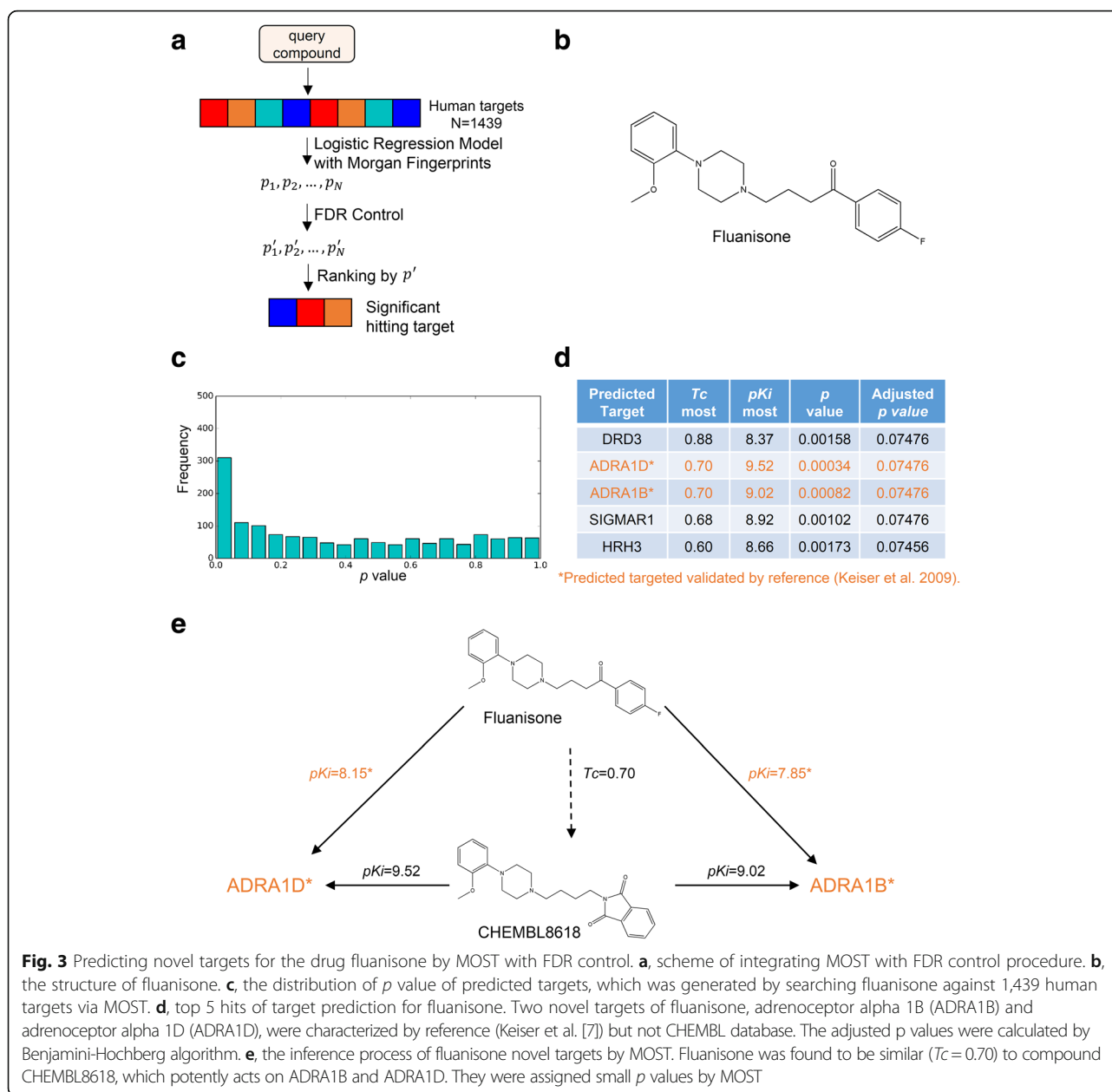


**Fig. 2** Prediction results of MOST in one dataset of sevenfold cross-validation with Logistic Regression method and Morgan fingerprint. **a** and **b**, the predicted results derived from different “active” data definition:  $pK_i \geq 5$  and  $pK_i \geq 6$ . Results generated by using explicit and implicit  $K_i$  of most-similar ligand in model training are compared. *Left panels*, the predicted results in  $T_{c\_most}$  vs  $pK_{i\_most}$  scatter plot. *Middle panels*, the fraction of data regarding to the increasing threshold of  $T_c$ . *Right panels*, the data fraction regarding to the decreasing threshold of  $p$  values. The difference between  $f_{TP}$  and  $f_{FP}$  was plotted in black, dash line. In all panels, true positives are colored red, while true negatives are blue; false positives are cyan, while false negatives are orange

By using MOST, we searched aloes-emodin against 1,439 human targets, and found that acetylcholinesterase (ACHE) was the top target (Fig. 4a). Aloe-emodin was similar ( $T_c = 0.50$ ) to CHEMBL3233826, the rhein-derived compound which potently ( $pK_i = 8.97$ ) inhibits ACHE [33]. Actually, aloe-emodin was shown to inhibit ACHE with  $pK_i = 4.57$  in an early study [34]. ACHE is the enzyme involved in the rapid hydrolysis of acetylcholine in numerous cholinergic pathways. Inhibition of ACHE results in accumulation of acetylcholine and hyperstimulation of the gastrointestinal smooth muscles

via muscarinic M2 and M3 receptors [35, 36]. Indeed, one of the ACHE inhibitors, acotiamide is approved as a prokinetic agent for treating functional dyspepsia [37].

Given these facts, we tested whether the laxative effect of aloe-emodin is mediated by the acetylcholine signaling pathway. In C57 mice, aloe-emodin significantly increased the production of fecal pellets within 2 h after treatment with doses of 15-, 30-, and 60 mg/kg (Fig. 4b). The intragastric pre-treatment of atropine (2- and 4 mg/kg), given 20 min beforehand, totally abolished the stimulatory effect of aloe-emodin (15 mg/kg) on



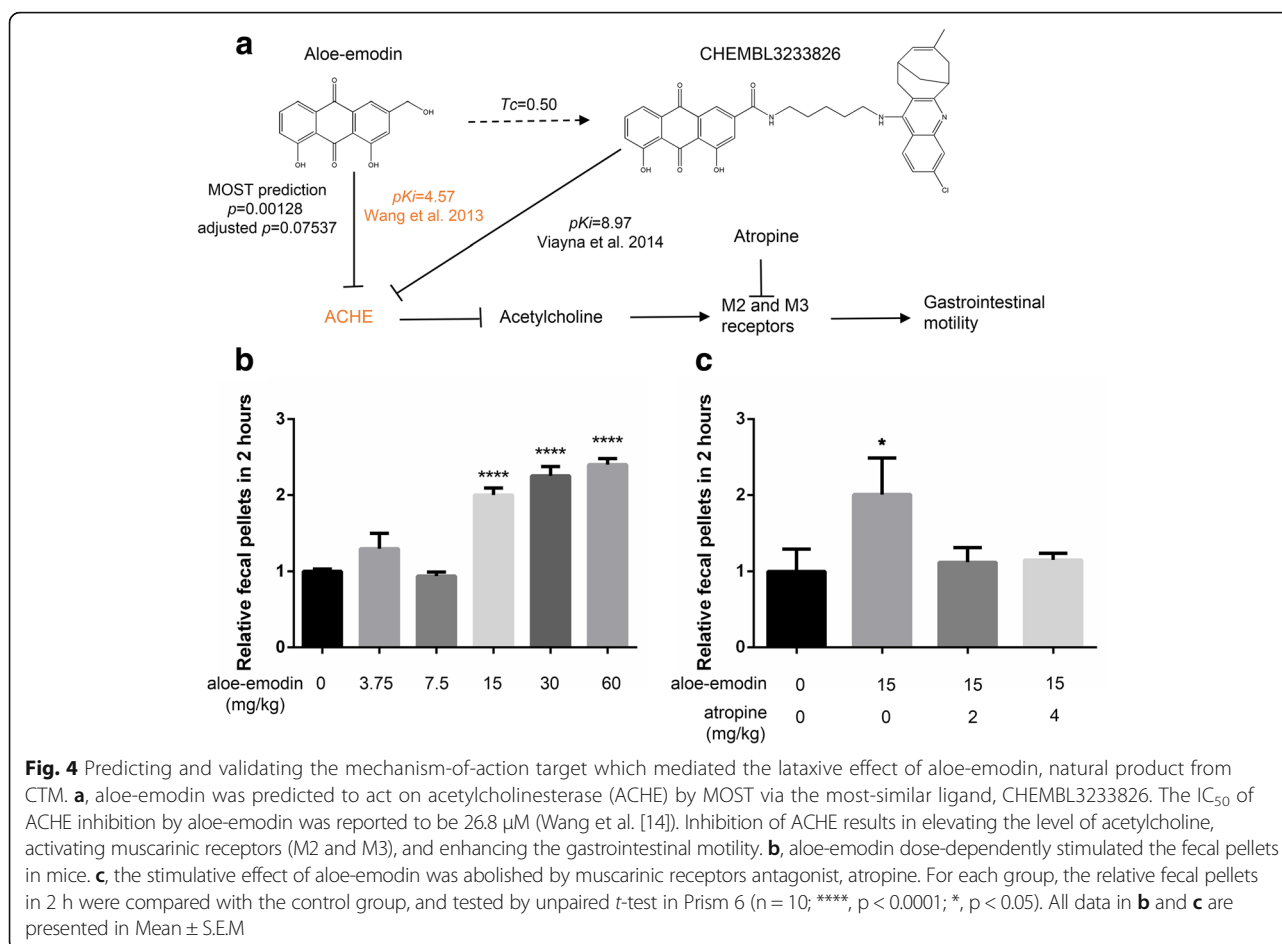
mice fecal pellet output (Fig. 4c). These results suggest that the cholinergic pathway is involved in the laxative effect of aloë-emodin on mice colonic motility.

## Discussion

Utilizing the fact that similar compounds have similar bioactivity profiles [17, 38], similarity searching is one of the most simple, but robust, approaches to ligand-based target prediction. The earliest example was PASS, in which chemicals were represented by MNAs (multilevel neighborhoods of atoms) descriptors [39]. A Bayesian model was employed to train 31,000 bioactive substances, and the biological activity spectra, including 319

types of pharmacological effects, action mechanisms and toxicities, were predicted in the form of probabilities. By analyzing the similarity between ligand sets of various targets, Keiser et al. proposed SEA, which uses ensemble similarity to make target prediction [6]. In SEA, the relationship of a compound with a biological target is determined by calculating the sum of fingerprint similarities of known ligands annotated with that target, and  $T_c \geq 0.57$  was used to remove substantially unrelated ligands. The prediction significance was accessed by "BLAST-like" Z-score and  $p$  values according to a pre-fitted probability distribution in SEA. Unlike SEA, the mean of similarity to multiple ligands of target is utilized





with multi-category Bayes classifiers to improve the performance of ligand-based target inference [40, 41]. However, it seemed not to be necessary to use all ligand information, whether in sum or mean, when relating compound with target annotated with multiple ligands. It was firstly demonstrated with MCM algorithm [42], and later proved in TargetHunter, that performance of similarity searching can be dominated by top similar ligands [14]. Except Bayes classifiers, other machine learning methods such as Support Vector Machines (SVMs) [43], Logistic Regression [12, 44], and Random Forests [28], were also employed for task of target prediction.

In the current study, we demonstrated that, solely using the information of most-similar ligands, MOST achieves high prediction accuracy. We also investigated the effects of using explicit bioactivity of most-similar ligand, which has usually been oversimplified as category values (implicit bioactivity) in previous similarity searching approaches. There was only little difference in prediction accuracy between MOST using explicit- and implicit bioactivities. In both cases, a large fraction of FPs were found to result from most-similar ligand with high  $T_c$  ( $>0.8$ ) but low  $pK_i$  values. This is an important

finding; it suggests that simply using a  $T_c$  threshold cannot reduce the major part of FPs. In such case, using explicit bioactivity of most-similar ligand provides a significant advantage over implicit bioactivity because, in explicit bioactivity mode, more potent ligands will generate better probability, while less potent ligands will give worse. That's why when  $p$  value threshold was applied, a large fraction of FPs were removed, while most TPs remained.

One limitation of the current study was the unbalanced training dataset—that is, the dataset included more “active” data and less “inactive data”. Since we extracted all  $K_i$  data with well-annotated references from the ChEMBL database, it seemed that researchers may be more likely to report positive, rather than negative results in their publications. The effects of skewed dataset were evaluated by MCC, which is more suitable for unbalanced datasets. If more negative data from other sources is included, the prediction performance can be further improved, as demonstrated in the work by Mervin *et al* [19].

We also demonstrated the application of MOST in the ‘real-world’ case of aloë-emodin. Considering there is

large unmet need to elucidate the mechanism-of-action targets of traditional medicine, MOST can be optimized for specific application domains, like biological function networks or disease pathways, which are influenced by traditional medicine therapies [45].

## Conclusions

Taken together, the results reported here show that MOST is a highly accurate approach to predicting targets. Logistic Regression and Random Forests learning methods performed better than Naïve Bayes in cross-validation, while Logistic Regression outperformed the other two in temporal validation. MOST has more power to detect more positive results with explicit activity. The  $p$  value, rather than  $Tc$ , is a robust way to filter out false positives. Integrated with the FDR control procedure, MOST provides a reliable framework to predict novel targets for known drugs and to predict the mechanism-of-action targets for active compounds from traditional medicines. These capabilities have been demonstrated via the examples of fluanisone and aloe-emodin. Success of MOST as reported here may have been partly because many query compounds had highly similar counterparts in datasets used in this study. If the query compounds are from a very different structural class than the ones in training dataset, MOST may be less accurate. Despite this potential limitation, MOST is a powerful approach to relating pharmaceuticals with their potential targets.

## Additional file

**Additional file 1: Figure S1.** Benchmark Ki dataset for evaluating the performance of MOST generated from ChEMBL19. **Figure S2.** Distribution of  $Tc_{most}$  and correlation between  $pKi_{most}$  and  $pKi_{query}$  in the training and test sets. **Table S1.** Statistics of Ki datasets generated from ChEMBL19 and ChEMBL20. **Table S2.** Prediction results of MOST with Logistic Regression method and Morgan fingerprint in sevenfold cross-validation. (DOCX 583 kb)

## Abbreviations

ACHE: acetylcholinesterase; ADRA1B: adrenoceptor alpha 1B; ADRA1D: adrenoceptor alpha 1D; FDR: false discovery rate; FN: false negatives; FP: false positives; MCC: Matthews Correlation Coefficient;  $Tc$ : Tanimoto coefficient; TCM: Traditional Chinese Medicine; TN: true negatives; TP: true positives

## Acknowledgements

The authors thank other members of MZRW Group, including Shu-hai Lin, Man Zhang, Yan-hong Li, Dong-dong Hu, and Chung-Wah Cheng, for their great help and support.

## Funding

This study was supported by Hong Kong Baptist University [grant number RC-IRMC/1213/01A]. The funding body did not play any role in the design or conclusion of this study.

## Availability of data and materials

Supporting information is available online. The datasets used and/or analyzed during the current study are available in the GitHub repository, <https://github.com/thuangsh/most>.

## Authors' contributions

ZXB designed the whole study. TH performed data collection, analysis, and building the model. HM performed animal testing with help from CYL and LZ. TH, CYL and ZXB wrote the manuscript. HM, LDZ, LZ, FBL, GZ and APL made substantial contributions to discussion and content. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Consent to publish

Not applicable.

## Ethics approval and consent to participate

The experimental protocols were approved by the Animal Ethics Committee of Hong Kong Baptist University, in accordance with "Institutional Guidelines and Animal Ordinance" from the Department of Health, Hong Kong Special Administrative Region.

## Publisher's Note

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

## Author details

<sup>1</sup>Lab of Brain and Gut Research, School of Chinese Medicine, Hong Kong Baptist University, 7 Baptist University Road, Hong Kong, People's Republic of China. <sup>2</sup>Department of Gastroenterology, the First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510405, People's Republic of China. <sup>3</sup>YMU-HKBU Joint Laboratory of Traditional Natural Medicine, Yunnan Minzu University, Kunming 650500, People's Republic of China. <sup>4</sup>Hong Kong Chinese Medicine Clinical Study Centre, Hong Kong Baptist University, 7 Baptist University Road, Hong Kong, People's Republic of China.

Received: 8 November 2016 Accepted: 4 March 2017

Published online: 11 March 2017

## References

- Lomenick B, Olsen RW, Huang J. Identification of direct protein targets of small molecules. *ACS Chem Biol*. 2011;6(1):34–46.
- Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Pujadas G, Garcia-Vallve S. Tools for in silico target fishing. *Methods*. 2015;71:98–103.
- Mohd Fauzi F, Koutsoukas A, Lowe R, Joshi K, Fan TP, Glen RC, Bender A. Chemogenomics approaches to rationalizing the mode-of-action of traditional Chinese and Ayurvedic medicines. *J Chem Inf Model*. 2013;53(3):661–73.
- Matter H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem*. 1997;40(8):1219–29.
- Tanimoto TT. IBM Internal Report 17th. 1957.
- Keiser MJ, Roth BL, Armbuster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007; 25(2):197–206.
- Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kujjer MB, Matos RC, Tran TB, et al. Predicting new molecular targets for known drugs. *Nature*. 2009;462(7270):175–81.
- Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. 2012;486(7403):361–7.
- Armstrong MS, Finn PW, Morris GM, Richards WG. Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J Comput Aided Mol Des*. 2011;25(8):785–90.
- Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI, Richards WG. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aided Mol Des*. 2010;24(9):789–801.
- Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res*. 2014;42(Web Server issue):W32–38.
- Gfeller D, Michielin O, Zoete V. Shaping the interaction landscape of bioactive molecules. *Bioinformatics*. 2013;29(23):3073–9.

13. Lo YC, Senese S, Li CM, Hu Q, Huang Y, Damoiseaux R, Torres JZ. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comput Biol*. 2015;11(3), e1004153.
14. Wang L, Ma C, Wipf P, Liu H, Su W, Xie XQ. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J*. 2013;15(2):395–406.
15. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012;40(Database issue):D1100–1107.
16. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The Experimental Uncertainty of Heterogeneous Public K-i Data. *J Med Chem*. 2012;55(11):5165–73.
17. Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem*. 2004;2(22):3204–18.
18. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? *J Med Chem*. 2002;45(19):4350–8.
19. Mervin LH, Afzal AM, Drakakis G, Lewis R, Engkvist O, Bender A. Target prediction utilising negative bioactivity data covering large chemical space. *J Cheminform*. 2015;7:51.
20. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742–54.
21. RDKit: Open-source cheminformatics. In: <http://www.rdkit.org>. Accessed 15 Mar 2016.
22. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform*. 2011;3:33.
23. Chan TF, Golub GH, LeVeque RJ. Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. In: Caussinus H, Ettinger P, Tomassone R. (eds) COMPSTAT 1982 5th Symposium held at Toulouse 1982. Heidelberg: Physica-Verlag HD; 1982.
24. Yu HF, Huang FL, Lin CJ. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach Learn*. 2011;85(1-2):41–75.
25. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
27. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405(2):442–51.
28. Fu G, Nan X, Liu H, Patel RY, Daga PR, Chen Y, Wilkins DE, Doerksen RJ. Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinforma*. 2012;13 Suppl 15:S3.
29. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met*. 1995;57(1):289–300.
30. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003;100(16):9440–5.
31. Inoue M, Ates N, Vossen JMH, Coenen AML. Effects of the Neuroleptanalgesic Fentanyl-Fluanisone (Hypnorm) on Spike-Wave Discharges in Epileptic Rats. *Pharmacol Biochem Be*. 1994;48(2):547–51.
32. Park MY, Kwon HJ, Sung MK. Evaluation of Aloin and Aloe-Emodin as Anti-Inflammatory Agents in Aloe by Using Murine Macrophages. *Biosci Biotech Bioch*. 2009;73(4):828–32.
33. Viayna E, Sola I, Bartolini M, De Simone A, Tapia-Rojas C, Serrano FG, Sabate R, Juarez-Jimenez J, Perez B, Luque FJ, et al. Synthesis and Multitarget Biological Profiling of a Novel Family of Rhein Derivatives As Disease-Modifying Anti-Alzheimer Agents. *J Med Chem*. 2014;57(6):2549–67.
34. Wang Y, Pan WL, Liang WC, Law WK, Ip DTM, Ng TB, Waye MMY, Wan DCC. Acetylshikonin, a Novel AChE Inhibitor, Inhibits Apoptosis via Upregulation of Heme Oxygenase-1 Expression in SH-SY5Y Cells. *Evid-Based Compl Alt*. 2013.
35. Colovic MB, Krstic DZ, Lazarevic-Pasti TD, Bondzic AM, Vasic VM. Acetylcholinesterase Inhibitors: Pharmacology and Toxicology. *Curr Neuropharmacol*. 2013;11(3):315–35.
36. Eglen RM. Muscarinic receptors and gastrointestinal tract smooth muscle function. *Life Sci*. 2001;68(22-23):2573–8.
37. Nolan ML, Scott LJ. Acotiamide: First Global Approval. *Drugs*. 2013;73(12):1377–83.
38. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular Similarity in Medicinal Chemistry. *J Med Chem*. 2014;57(8):3186–204.
39. Lagunin A, Stepanchikova A, Filimonov D, Porokov V. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics*. 2000;16(8):747–8.
40. Cleves AE, Jain AN. Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem*. 2006;49(10):2921–38.
41. Bender A, Mikhailov D, Glick M, Scheiber J, Davies JW, Cleaver S, Marshall S, Tallarico JA, Harrington E, Cornella-Taracido I, et al. Use of Ligand Based Models for Protein Domains To Predict Novel Molecular Targets and Applications To Triage Affinity Chromatography Data. *J Proteome Res*. 2009;8(5):2575–85.
42. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model*. 2006;46(3):1124–1133.
43. Plewczynski D, von Grothuss M, Spieser SAH, Rychlewski L, Wyrwicz LS, Ginalski K, Koch U. Target specific compound identification using a support vector machine. *Comb Chem High T Scr*. 2007;10(3):189–96.
44. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Research*. 2014;42(W1):W32–8.
45. Zobir SZM, Fauzi FM, Liggi S, Drakakis G, Fu XJ, Fan TP, Bender A. Global Mapping of Traditional Chinese Medicine into Bioactivity Space and Pathways Annotation Improves Mechanistic Understanding and Discovers Relationships between Therapeutic Action (Sub)classes. *Evid-Based Compl Alt*. 2016;2016:2106465.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

