BMC Bioinformatics

RESEARCH ARTICLE Open Access

Quantitative prediction model for affinity of drug–target interactions based on molecular vibrations and overall system of ligand-receptor

Xian-rui Wang, Ting-ting Cao†, Cong Min Jia, Xue-mei Tian and Yun Wang*

*Correspondence: wangyun@bucm.edu.cn
†Ting-ting Cao contributed equally to this work and should be considered co-first author.
Key Laboratory of TCM-Information Engineer of State Administration of TCM, School of Chinese Pharmacy, Beijing University of Chinese Medicine, Beijing 100102, China

Abstract

Background: The study of drug–target interactions (DTIs) affinity plays an important role in safety assessment and pharmacology. Currently, quantitative structure–activity relationship (QSAR) and molecular docking (MD) are most common methods in research of DTIs affinity. However, they often built for a specific target or several targets, and most QSAR and MD methods were based either on structure of drug molecules or on structure of receptors with low accuracy and small scope of application. How to construct quantitative prediction models with high accuracy and wide applicability remains a challenge. To this end, this paper screened molecular descriptors based on molecular vibrations and took molecule-target as a whole system to construct prediction models with high accuracy-wide applicability based on dissociation constant (Kd) and concentration for 50% of maximal effect (EC50), and to provide reference for quantifying affinity of DTIs.

Results: After comprehensive comparison, the results showed that RF models are optimal models to analyze and predict DTIs affinity with coefficients of determination (R^2) are all greater than 0.94. Compared to the quantitative models reported in literatures, the RF models developed in this paper have higher accuracy and wide applicability. In addition, E-state molecular descriptors associated with molecular vibrations and normalized Moreau-Broto autocorrelation (G3), Moran autocorrelation (G4), transition-distribution (G7) protein descriptors are of higher importance in the quantification of DTIs.

Conclusion: Through screening molecular descriptors based on molecular vibrations and taking molecule-target as whole system, we obtained optimal models based on RF with more accurate-widely applicable, which indicated that selection of molecular descriptors associated with molecular vibrations and the use of molecular-target as whole system are reliable methods for improving performance of models. It can provide reference for quantifying affinity of DTIs.

Keywords: Molecular vibrations, Random forest, Drug–target affinity, Chemical composition, Drug–target interactions

Background

The rapid development of systems biology has proposed a new view that a single drug molecule acts on multiple targets or that multiple drug molecules act on a common target [1, 2]. That is to say, there are multiple interactions between targets and drug molecules-DTIs. DTIs plays an important role in pharmacology, biology and mechanism [3–6]. For example, on the basic of DTIs research, the off-target toxicity of appetite suppressant Fen-Phen that can cause death is due to the activation of 5-HT2B receptor by one of its metabolites-Norfenfluramine, leading to proliferative valvular heart disease [7]. In study of repositioning salicylanilide anthelmintic drugs to treat adenovirus infections, the results showed that Niclosanide and Rafoxanide target transport of HAdV particle from endosome to nuclear envelope, whilst oxyclozanide specifically targets adenovirus immediately early gene E1A transcription [8]. Therefore, the research of DTIs will help to understand mechanisms or toxic side effects of drugs and repositioning of drugs [9–12].

Currently, the research on DTIs focused on two directions, one is traditional experimental analysis and the other is DTIs predictive analysis based on existing databases [13]. Traditional experimental analysis of DTIs is expensive and inefficient, and faces many challenges such as financial, technical and time aspects. It is almost impossible for researchers to carry out experiments to identify mechanisms or toxic side effects for all drug compounds. In comparison, the prediction of DTIs that is efficient and low cost can make up for shortcomings of traditional trials [14]. In prediction of DTIs, prediction of drug–target affinity is becoming increasingly important. This is because prediction of affinity not only predicts weather there is an interaction between molecules and targets, but also obtains strength of interaction, which is useful for drug discovery, effect and toxic evaluation, etc. Computational approaches for DTIs affinity in most of current research mainly include two categories: ligand-based and receptor-based methods [15, 16]. In above methods, quantitative structure–activity relationship (QSAR) and molecular docking (MD) are most common methods. Such as Simeon S, et al., constructed QSAR models of Janus kinase 2 inhibitors based on machine learning algorithms to predict inhibitory potency [17]. Luo M, et al., used random forests (RF), support vector machine (SVM), and K Nearest Neighbors (KNN) to construct QSAR models of 5-HT1A Receptor, in which Ki value characterized affinity of receptor-ligand [18]. Van Den Driessche G and Fourches D used 3D molecular docking to reveal common HLA-B*57:01 variants that trigger adverse drug reactions [19]. In addition, there is also a similarity search-based approach, which utilizes chemical structure similarity to predict DTIs and DTIs affinity [20, 21].

However, quantitative structure–activity relationship (QSAR) and molecular docking (MD) have some limitations. QSAR or MD is often built for a specific target or several targets, making it difficult to achieve quantitative prediction for multiple targets at the same time, which leads to a small range of applications. Moreover, molecular docking and its evaluation methods are limited to 3D structure of target proteins [22–24]. Molecular docking is inaccurate when those proteins whose 3D structure is unknown, especially for membrane proteins whose 3D structure is difficult to crystallize [25, 26]. These limitations are severe because most useful drug targets are membrane proteins, such as ion channels and G protein-coupled receptors (GPCRs) [27, 28]. This leads to

low accuracy and low applicability of most DTIs prediction models, not to mention prediction of affinity for DTIs. The more serious fact is that most QSAR and MD were based either only on structure of ligands or on structure of receptors. By considering only structure of receptor or ligand, similarity-based analysis inevitably leads to inaccurate results that are inconsistent with experimental results. This fragmented approach ignores holistic nature of receptor-ligand interactions, which leads to low prediction accuracy and excessive bias. In addition, in constructing quantitative prediction models, researchers mostly used molecular descriptors to solve problem of quantifying abstract molecules, and solved mapping problem of best-described function by optimizing algorithm and parameters. However, researchers ignore problem of feature characterization. This can also lead to low accuracy and excessive bias for prediction of DTIs affinity [29, 30].

In this paper, with above limitations in mind, we took molecule-target as a whole system from systems biology perspective to construct prediction models for DTIs affinity with high accuracy and wide applicability, in which simultaneously considering both receptors and ligands. Molecular descriptors associated with molecular vibrations were combined with protein sequence descriptors to construct whole system of molecule-target, in which Kd and EC50 were used as quantitative indicators. On the premise of feature selection, combining machine-learning algorithms to predict DTIs affinity efficiently and accurately. These models consisted of internal cross-validation and external tests, which provided a predicted performance with high accuracy and wide applicability. In addition, optimal models were selected for application evaluation and comprehensive comparison. The new quantitative models will provide reference for prediction of DTIs affinity.

Results and discussions

As shown in Fig. 1, this paper was completed under that research methodology. In the following section, we presented and discussed the data collection, results of descriptor calculation, optimal prediction models, the importance of descriptors and so on.

Data collection

Based on open source database: PubChem (https://pubchem.ncbi.nlm.nih.gov/), Drugbank (https://go.drugbank.com/), ChEMBL (https://www.ebi.ac.uk/chembl/) and Uniprot database (https://www.uniprot.org/), we performed data collection of drug molecules, target protein sequences, and Kd and EC50 values characterizing drug molecule-target affinity. Taking drug molecule and target as a whole system, we obtained the EC50 dataset-quantifying DTIs affinity by EC50 and the Kd dataset-quantifying drug molecule-target affinity by Kd, respectively. The EC50 dataset contains 8147 ligands and 544 targets, and 11,076 ligand-target-EC50 pairs. At the same time, The Kd dataset contains 1870 ligands and 778 targets, and 10,923 ligand-target-Kd pairs. The two datasets without redundancy were used as benchmark datasets.

In process of data collection, we kept to the following two criteria: (1) maintain entries as many as possible; (2) exclude redundant data as many as possible. Therefore, some drug molecules and targets were removed due to Kd, EC50 has no definite value, or their activity values are inconsistent. These redundant data may strongly affect the accuracy of

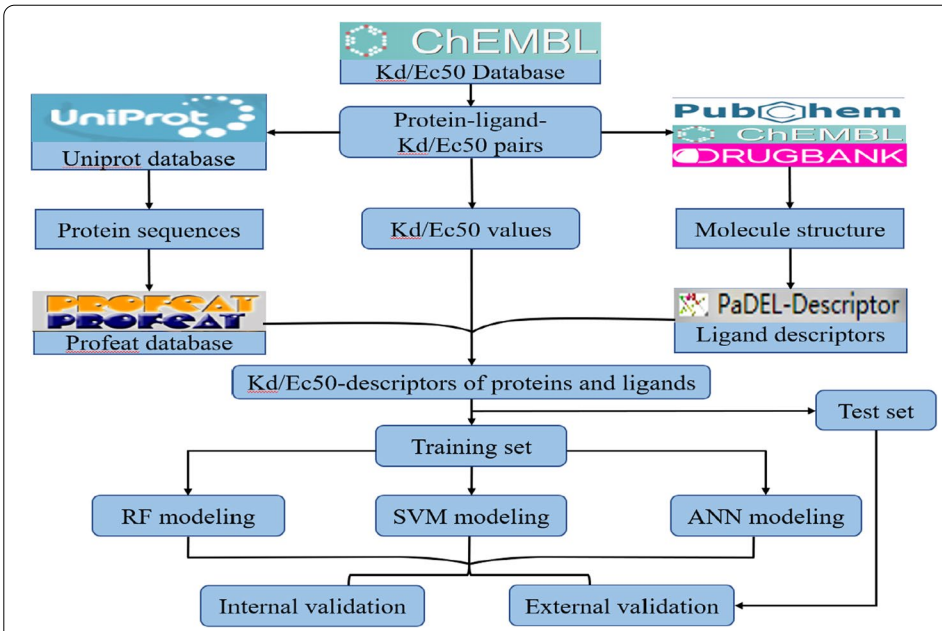

Fig. 1 The specific research methodology

prediction models for DTIs affinity. It is worth noting that EC50 refers to the concentration of a drug, antibody or toxicant that achieves 50% of the maximum biological effect after a specified exposure time. It was commonly used as a measure of a drug's potency [31]. Kd is often used to describe degree of binding of a compound to a particular target [32]. The smaller dissociation constant, the higher affinity between compounds and proteins. Considering the practical significances of Kd and EC50, we finally chose both as quantitative indexes of DTIs affinity.

Results of descriptor calculation

Calculation of drug molecule descriptors

After calculation by online platform-PaDEL, we obtained the molecular descriptors. The descriptors calculated in this article were shown in Table 1. There were 1874 descriptors for drug molecules and drug molecular descriptors can be divided into 16 categories, among which E-state descriptors, Autocorrelation descriptors and Topological type descriptors account for a relatively large number. Even though many descriptors in Table 1 are of the same type, each descriptor has its own specific meaning. However not all molecular descriptors are suitable for the construction of predictive models for DTIs affinity.

Therefore, how to measure importance of descriptors and filter out meaningful ones is the key to improve accuracy of prediction models. Some researchers used kernel functions, thresholds, and other methods to filter descriptors to improve accuracy of models [33, 34]. It is worth considering that these methods don't take into account properties of drug molecules and that may not be applicable in quantitative prediction of DTIs.

After comprehensive consideration, in this paper, based on properties of drug molecules, we screened characteristic descriptors of drug molecules from the perspective

Table 1 Type and number of drug molecule descriptors

Serial number	Descriptor type	Number of descriptors
1	Constitutional descriptors	120
2	Autocorrelation descriptors	346
3	Basak descriptors	42
4	BCUT descriptors	6
5	Burden descriptors	96
6	Connectivity descriptors	56
7	E-state descriptors	489
8	Kappa descriptors	3
9	Molecular property descriptors	15
10	Quantum chemical descriptors	5
11	Topological descriptors	265
12	CPSA descriptors	29
13	RDF descriptors	210
14	Geometrical descriptors	21
15	WHIM descriptors	91
16	3D Autocorrelation descriptors	80

of molecular vibrations. This is because molecular vibrations are caused by vibrations of chemical bonds within molecules and they are macroscopic representation of drug molecules properties [35, 36]. Moreover, molecular vibrations are affected by various factors such as conjugation effect, induction effect, spatial effect, hydrogen bonding, vibrational coupling effect, etc. Therefore, molecular vibrations can reflect drug molecular structure and physicochemical properties of drugs to a certain extent [37]. It should be remembered that seven physicochemical properties are particularly relevant to molecular vibrations, including electronegativity, π-atomic charge, total charge, and bond polarity [38]. Therefore, we choose molecular descriptors related to molecular vibrations based on above physicochemical properties. For instance, Mpe-Constitution Descriptor-mean Atomic Pauling Electronegativity (scaled on carbon atom) was selected as feature descriptor to construct prediction models for DTIs affinity due to its relation to atomic electronegativity. Finally, 813 descriptors associated with molecular vibrations were selected from 1874 descriptors in Table 1 to represent the feature characteristics of drug molecules. In addition, 813 molecular descriptors associated with molecular vibrations and their specific meanings were given in Additional file 3: Table S1.

Calculation of target protein descriptors

As was known to all, 3D structures of many proteins are unknown, especially for membranous proteins [27, 28]. Thus, the analysis based on protein sequences rather than 3D structures of proteins can ensure a wide range of applicability of models and accuracy [39]. The target protein descriptors were shown in Table 2.

As shown in Table 2, there are 1437 descriptors for each protein and the descriptors can be divided into 9 categories, among which Dipeptide composition, Moran autocorrelation as well as Normalized Moreau-Broto autocorrelation account for a relatively large number.

Table 2 Type and number of target protein descriptors

Serial number	Descriptor type	Number of descriptors
G1	Amino acid composition	20
G2	Dipeptide composition	400
G3	Normalized Moreau-Broto autocorrelation	240
G4	Moran autocorrelation	240
G5	Geary autocorrelation	240
G6	Composition	21
G7	Transition, distribution	126
G8	Sequence-order-coupling number	60
G9	Quasi-sequence-order descriptors	100

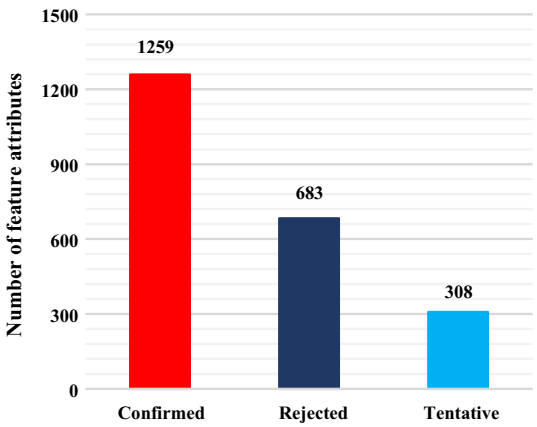

Fig. 2 Feature filtering results for EC50 dataset

813 drug molecule descriptors were integrated with 1437 protein descriptors and Kd, EC50 datasets to obtain the integrated Kd, EC50 datasets.

Results of feature screening

The Boruta algorithm was used for feature filtering [58]. As shown in Fig. 2, for integrated EC50 dataset, 1259 descriptors were marked as "Confirmed" and 683 descriptors were marked as "Rejected", with 308 descriptors being marked as "Tentative". That is, after feature selection, each DTI in the integrated EC50 dataset was characterized by 1259 feature attributes. Similarly, as shown in Fig. 3, for the integrated Kd dataset, 827 descriptors were marked as "Confirmed" and 1191 descriptors were marked as "Rejected" with 232 descriptors being marked as "Tentative". Each DTI in integrated Kd dataset was characterized by 827 feature attributes. The feature subsets of EC50 and Kd were obtained by feature screening for construction of quantitative prediction models for DTIs affinity.

The purpose of feature selection in machine learning is to filter out features set that minimize the cost function of currently selected model. However, Boruta algorithm can use a random forest approach to select the set of all features that are relevant to the dependent variable, rather than selecting the set of features that minimizes penalty

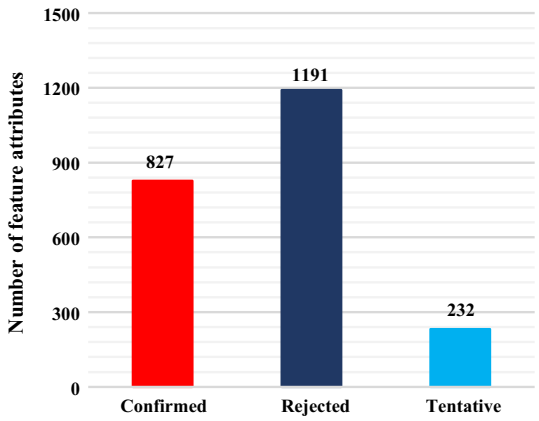

Fig. 3 Feature filtering results for Kd dataset

Table 3 Tenfold cross validation of three kinds of algorithms for EC50 feature subset

Model (EC50)	R^2		MSE		RMSE		MAE		SSE	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SVM	0.9317	0.5759	0.1270	0.8356	0.3564	0.9146	0.1960	0.5801	1249	8216
RF	0.9611	0.9641	0.0891	0.0817	0.2985	0.2858	0.1976	0.1989	876	803.3
ANN	0.7350	0.5211	0.4867	0.9590	0.6976	0.9793	0.5023	0.6792	4785	9429

factor only for a specific model such as SVM. It can also disrupt the order of features and calculate the importance of features. Boruta algorithm can help us to understand the factors influencing dependent variable more fully and make feature selection better and more efficient. Therefore, when it's not known upfront which algorithm is optimal, we chose Boruta algorithm for feature filtering.

Results of quantitative prediction model for DTIs affinity

Parameter optimization

In the RF model, there are important parameters need to be considered, such as ntree and max depth. After comparison and optimization of several parameters, we finalized RF algorithm parameters: ntree = 500, max depth = no limitation, min samples split = 2, min samples leaf = 1, max leaf nodes = none. For SVM model, we used "Tune" function to determine the optimal parameters of SVM algorithm, with the following algorithm parameters: cost = 1000, gamma = 0.0001 [40]. In same optimization way, ANN algorithm parameters were determined: size = 2, decay = 0.1, linout = T (non-linear function), maxit = 1000, max nwts = 10,000.

Optimal prediction model for DTIs affinity

Before attempting to construct prediction models for DTIs affinity, EC50 feature subsets were preprocessed to facilitate calculation. Then combined with SVM, RF and ANN to construct quantitative prediction models respectively. The results of tenfold cross validation for EC50 feature subset were shown in Table 3.

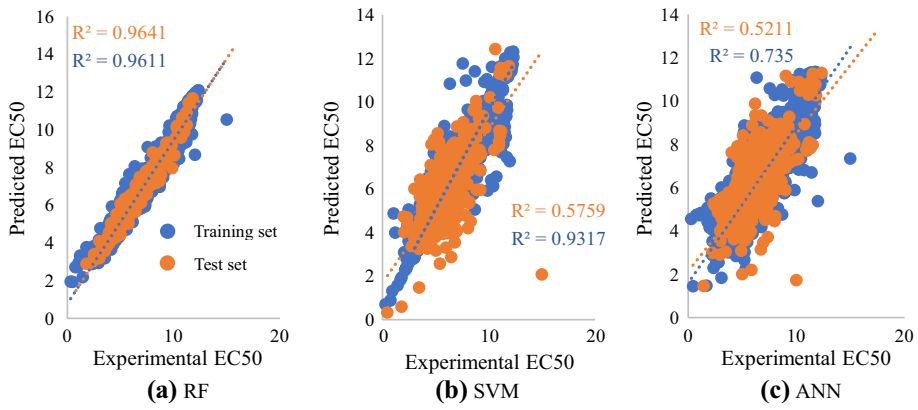

Fig. 4 Scatter plot of experimental and predicted EC50 values for three prediction models (**a**: EC50 scatterplot based on random forest model, **b**: EC50 scatterplot based on support vector machine model, **c**: EC50 scatterplot based on K Nearest Neighbors model.)

Table 4 Tenfold cross validation of three kinds of algorithms for Kd feature subset

Model (Kd)	R^2		MSE		RMSE		MAE		SSE	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SVM	0.9099	0.5083	0.1254	0.7290	0.3541	0.8538	0.2116	0.6406	1230	808.4
RF	0.9425	0.9485	0.1208	0.1191	0.3476	0.3451	0.2640	0.2594	1204	132.1
ANN	0.5857	0.2961	0.5612	1.0190	0.7491	1.0095	0.5792	0.7390	5593	1130

As shown in Table 3 and Fig. 4, In RF model, R^2 of training and test sets are 0.9611, 0.9641 respectively indicated a good fit of RF model to data. MSE of training and test sets were both less than 0.09 and were in same order of magnitude, which indicated that there is no overfitting problem existing, and demonstrated that RF model showed satisfactory predictive performance (Fig. 4a). As for SVM model, R^2 of training and test sets are 0.9317, 0.5759 respectively. SVM model exhibited some differences between training and test sets, but order of magnitude is the same and no greatly obvious overfitting can be observed from SVM model (Fig. 4b). However, predictive performance of the SVM model was worse than that of the RF model. For training and test sets in ANN model, no obvious overfitting can be observed (Fig. 4c), but the performance of ANN model in training and test set were lower than both RF model and SVM model. By comparing predictive performance of three models based on evaluation indicators, it can be observed that the performance of RF model is best selection for EC_{50} data.

The same analysis was appropriate for Kd dataset, on the basic of data in Table 4 and scatter plot in Fig. 5, we completed selection of optimal model: RF model showed satisfactory predictive performance with R^2 of test set being 0.9485 (Fig. 5a). The SVM model suffered from overfitting and its predictive performance was worse than that of RF model (Fig. 5b). ANN model was the least effective model (Fig. 5c). The results indicated that RF model is the optimal quantitative prediction model for KD dataset.

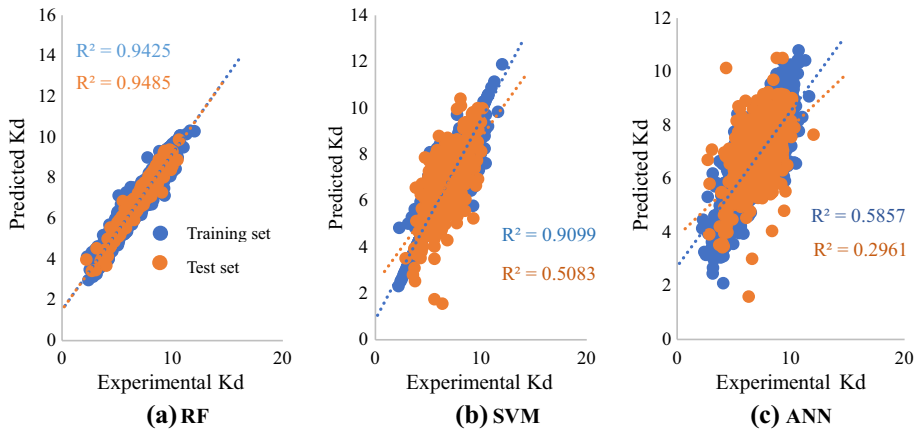

Fig. 5 Scatter plot of experimental and predicted Kd values for three prediction models (**a**: Kd scatterplot based on random forest model, **b**: Kd scatterplot based on support vector machine model, **c**: Kd scatterplot based on K Nearest Neighbors model.)

In summary, whether based on EC50 dataset or Kd dataset, the performance of RF models are the best. Therefore, in this paper, random forest (RF) models are more suitable for quantitative prediction of biological activities for DTIs affinity.

Evaluation of application for optimal models

By comparing analysis in 2.4, we obtained RF optimal models. To demonstrate the reliability and applicability of RF models further, we used RF models for analysis of DTIs in Binding DB database, in which Kd and EC50 quantified affinity of DTIs.

Using same data collection methods and eliminating duplicate data, we collected 1045 ligand-receptor-EC50 pairs and 89 ligand-receptor-Kd pairs from Binding DB database for quantitative analysis of DTIs affinity. Quantitative analysis of new dataset was carried out using RF models based on Kd and EC50. Calculating absolute value of the difference between true value and predicted value (referred as '|d|' from now) and dividing |d| into 5 parts in which each part was divided on a scale of 0.5.

Therefore, we obtained the distribution histogram of |d| (Fig. 6) in new EC50 and Kd datasets, reflecting prediction capability of RF models.

The predictive values of RF models were all greater than zero, suggesting that drug molecule-target interactions do exist, which is consistent with data information gathered from datasets. This indicated that optimal models constructed in this paper could be accurately used for qualitative prediction of DTIs. However, as shown in Fig. 6, eighty percent of |d| distribution was 1.5–2.0. The range of differences was within 2.0 for 98.95% (EC50) 96.63% (Kd) of |d| respectively. This indicated that there is error between predicted value and experimental value. The above demonstrated that quantitative RF prediction model developed in this paper can predict affinity of DTIs to a certain extent based on Kd and EC50.

Comprehensive comparisons of models

Besides evaluation of application of RF models, comprehensive comparisons were made with predictive models for DTIs previously reported. In recent years, there have been

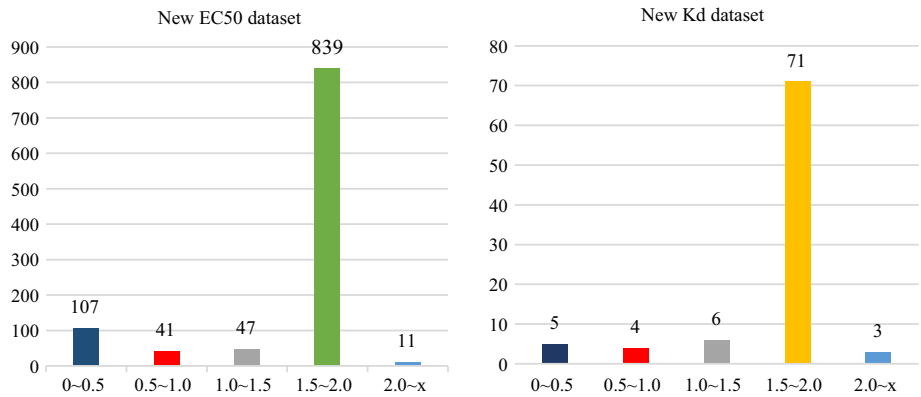

Fig. 6 Distribution histogram of |d| for EC50 dataset and Kd dataset

many reports for predicting DTIs, such as Xie L, et al., that integrated transcriptomic data by a deep-learning algorithm to predict the potential DTIs [41]. Olyan R S, et al., developed a novel method based on RF model to improve DTIs prediction accuracy [42]. Chen N, et al., carried out a quantitative analysis of antioxidant activity of antioxidant tripeptides in free radical systems based on QSAR [43]. In above analysis, there are only analysis based on structure of ligand or receptor, rather than taking ligand-receptor as a whole system for DTIs analysis. These methods of analysis, with separated ligands from receptors, can be limited by their own structure and produce non-reciprocal results, leading to poor accuracy. Conversely, in this paper, the models were constructed to take full account of ligands and receptors. From perspective of taking molecule-target as a whole system, we integrated molecule-target descriptors to construct predictive models for DTIs affinity, which is able to avoid unequal results based on receptors or ligands only, thus increasing accuracy of prediction models. At the same time, based on whole system of ligand-receptor, we can collect a large amount of molecule-target data rather than building for a specific target or several targets, expanding scope of application.

There were related reports on quantitative prediction of DTIs affinity. Based on 9948 DTIs quantified by Ki, 1589 molecular descriptors and 1080 protein descriptors, Shar P A, et al., constructed quantitative prediction models for DTIs using RF and SVM model, respectively [44]. However, the Coefficient of Determination-R^2 of RF and SVM models in training set are 0.88 and 0.86, at the same time that of RF and SVM models in test set are 0.63 and 0.61, which showed that there exists over-fitting. That is to say, predictive models have low accuracy. The main reasons for that would be improper characterization of drug molecules-targets and lack of feature screening. Considering this situation, in this paper, we screened characteristic descriptors of drug molecules from the perspective of molecular vibrations [35, 38]. Moreover, the analysis based on protein sequences rather than 3D structure of protein can ensure a wide range of applicability of models and its accuracy [39]. Therefore, the SVM and RF models in this paper had good results better than above research. In addition, the two datasets in this paper involved 544 and 778 targets respectively, which guaranteed that the models had some broad applicability. Likewise, Hakime Öztürk, et al., constructed DeepDTA to quantify the affinity of ligands-receptors, in which the results were not ideal. In process of building

model, more attention was paid to amount of data and neglecting molecular feature representation. The R^2 of Convolutional Neural Network (CNN) model was less than 0.70, which was lower than optimal RF models in this paper. The MSE of CNN model was high than 0.194, which was high than that of RF models in this paper (0.119) [45]. Abbasi W A, et al., proposed a sequence-based novel protein binding affinity predictor called ISLAND, in which the SVR model for LA kernel was the best model with R = 0.44, MSE = 6.55 [46]. Above comparative results showed that RF models developed based on Kd and EC50 in this paper can perform quantitative prediction of DTIs affinity more accurately with certain applicability and reliability. Moreover, literature already reported has not characterized drug molecules from the perspective of molecular vibrations. Based on the methods and good results of this paper, it was also shown that parametric characterization based on molecular vibrations is crucial for construction of prediction models for DTIs affinity with more accurately.

Analysis of molecular descriptors and protein descriptors

Molecular descriptors and protein descriptors are essential for construction of quantitative models for DTIs. Judged on the importance of descriptors, we can obtain feature descriptors that have higher importance in the quantification of DTIs affinity based on EC50 and Kd values, which can help us to analyze the importance of different molecular descriptors for quantification of DTIs and provide us with biological insights. Therefore, in the process of feature screening, we filtered the descriptors according to their importance scores to obtain the important descriptors. Importance score of single feature is equal to (oob_accuracy - oob_accuracy_after_perputation), in which the oob_acc_after_perputation is the accuracy of samples on the singletree count after shuffling the dimensional feature with out_of_bag.

For the EC50 datasets, we obtained the top-ranking molecular descriptors and protein descriptors according to importance scores. The top-ranking protein descriptors and molecular descriptors were shown in Tables 5 and 6. In addition, it can be seen in Additional files 1 and 2 for more information on ranking the importance of molecular descriptions and protein descriptors.

As shown in Tables 5 and 6, we retained descriptors with importance scores greater than 0.85 in the feature screening process with maximum value of 1. A higher importance score means that the corresponding descriptor is more important for quantification of DTIs.

Table 5 The top-ranking protein descriptors in EC50 datasets

Protein descriptors	Important scores	Protein descriptors	Important score
[G7.1.1.66]	1.00	[G3.3.4.1.8]	0.99
[G4.1.15.1]	1.00	[G3.3.2.1.19]	0.97
[G4.1.23.3]	1.00	[G5.2.2.13]	0.94
[G4.2.8.1]	1.00	[G3.3.2.1.22]	0.90
[G4.2.11.1]	1.00	[G3.3.4.1.27]	0.93
[G7.1.1.47]	0.98	[G7.1.1.43]	0.92

Table 6 The top-ranking molecular descriptors in EC50 datasets

Molecular descriptors	Important scores	Concrete meaning
JGI5	0.96	Mean topological charge index of order 5
minaaSe	0.94	Minimum atom-type E-State: aSea
maxaaS	0.91	Maximum atom-type E-State: aSa
minHsSH	0.91	Minimum atom-type H E-State: -SH
maxssssSn	0.88	Maximum atom-type E-State: > Sn <
nHdsCH	0.87	Count of atom-type H E-State: = CH-
maxsNH2	0.86	Maximum atom-type E-State: -NH2
maxssPH	0.86	Maximum atom-type E-State: -PH-
ETA_Beta_s	0.86	A measure of electronegative atom count of molecule
maxddssSe	0.86	Maximum atom-type E-State: = Se =

Based on Tables 2 and 5, it can be found that protein descriptors with high importance are concentrated in Normalized Moreau-Broto autocorrelation (G3), Moran autocorrelation (G4), Transition-Distribution (G7). It is well known that DTIs include a variety of interaction modes, such as electrostatic interaction, hydrophobic interaction, spatial interaction and hydrogen bond. G3 and G4 are the autocorrelation functions combining above physicochemical properties and they can reflect the action strength of DTIs to some extent [47, 48]. G7 represents the amino acid distribution pattern of a specific structural or physicochemical property along a protein or peptide sequence, which directly influence ligand-receptor interactions and it has been used for recognition of protein folds and prediction of ligand-receptor interactions [48]. To sum up, G3, G4 and G7 descriptors are closely related to DTIs, therefore, they have higher importance scores in feature screening.

As for molecular descriptors, according to Tables 1 and 6, it can be found that molecular descriptors with high importance were concentrated in E-state descriptors. E-state descriptors characterize both topological information of each atom and electronic relationships between atoms in the molecule [49]. The three molecular forces, dispersion, dipole moment and hydrogen bonding, which influence the strength of DTIs affinity, are closely related to the electronic relationships characterized by E-state descriptors [49, 50]. Due to their above natures, E-state descriptors have been widely used in the analysis of DTIs [51]. This suggests that E-state descriptors are a good choice for analyzing and predicting DTIs affinity.

For the Kd datasets, the filtering results of descriptors were generally consistent with the EC50 datasets. The top-ranking molecular descriptors were concentrated in E-state descriptors and the top-ranking protein descriptors were concentrated in Normalized Moreau-Broto autocorrelation (G3), Moran autocorrelation (G4), Transition-Distribution (G7). More information can be seen in supplementary data.

In summary, E-state molecular descriptors associated with molecular vibrations and G3, G4 and G7 protein descriptors are of higher importance in the quantification of DTIs. They are important for the analysis and prediction of DTIs affinity.

All in all, in this paper, the ligand and receptor were used as a whole system to analyze and predict DTIs affinity. But the method will not be limited to the known receptor-ligand interaction space and it enables the identification of new action targets of

chemical components and the prediction of active affinities. Although there is a margin of error, it can provide clues and guidance to elucidate the action mechanism of drugs. In addition, it is possible to identify unknown potential compounds for the treatment of diseases based on their relevant targets or to reposition existing drugs.

Conclusion

In this paper, from perspective of overall systematic of ligand-receptor, through screening descriptors based on molecular vibrations and protein sequences, we obtained optimal models based on RF with more accuracy-widely applicability. This method can provide a reference for DTI's affinity prediction. It also indicated that describing molecular features based on molecular vibrations, taking drug molecule-target as whole system were reliable approaches for construction of prediction model for DTIs affinity and improving its accuracy. In addition, E-state molecular descriptors associated with molecular vibrations and G3, G4 and G7 protein descriptors are important for the analysis and prediction of DTIs affinity.

Methods and materials

In this paper, we constructed prediction models for DTIs affinity from the perspective of taking molecule-target as a whole system. Firstly, drug molecules and protein sequences as well as ligand-receptor-Kd/EC50 were screened on the basic of existing databases. Secondly, descriptors of drug molecules and protein sequences were calculated separately, and descriptors associated with molecular vibrations were selected from drug molecule descriptors. Thirdly, based on descriptors obtained in step 2, we constructed Kd and EC50 quantified drug molecule-target feature datasets by taking drug molecule and target as a whole system, respectively. Finally, combining above datasets with machine learning algorithms SVM, RF, ANN for construction of prediction models of DTIs affinity.

Datasets

This paper carried out construction of prediction model of DTIs affinity, which requires a large amount of data support. The drug molecules (ligands) were collected from open source database: PubChem (https://pubchem.ncbi.nlm.nih.gov/), Drugbank (https://go.drugbank.com/) and ChEMBL (https://www.ebi.ac.uk/chembl/) [52–54]. The target protein sequences (receptor) were collected from open source Uniprot database (https://www.uniprot.org/) [55]. In addition, the Kd and EC50 values used to quantify protein–ligand affinity were also obtained from ChEMBL database. All the data as of 10 June 2020.

Drug molecules and protein sequence descriptors

Descriptors can effectively solve problem of parametric characterization of drug molecules and protein sequences, which facilitate the construction of predictive models for DTIs affinity. In this paper, using PaDEL to calculate the descriptors of drug molecules [56]. Each descriptor has a specific explanation and we screening descriptors of drug molecules from the perspective of molecular vibrations. In addition, protein sequence

descriptors such as peptide composition and dipeptide composition were calculated by using PROFEAT web server (https://bio.tools/profeat) [48, 57].

Feature selection

The Kd, EC50 datasets and molecular descriptors and protein sequence descriptors were integrated separately to obtain the integrated Kd, EC50 datasets. The feature subsets of integrated Kd, EC50 datasets were obtained by using Boruta algorithm (R 3.5.2 version) in feature selection.

Boruta algorithm flow is as follows [58]: first, the features of feature matrix [$\mathbf{X}$] are shuffled to obtain shadow feature [$\mathbf{X}^0$], and the shadow features are stitched together with true features to form a new feature matrix [$\mathbf{Y}$]. Then, using the new feature matrix as input to output feature importance and calculate the "Z-score" of true and shadow features. Further, taking the largest "Z score" among shadow features as "Z-max", and marking the real features with "Z-score" greater than "Z-max" as "Important". At the same time, marking the real features with "Z-score" significantly smaller than "Z-max" as "Rejected". Finally, repeating five times until we can obtain all features that are marked as "Important":

$$[\mathbf{X}] \supseteq [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots\ldots, \mathbf{X}_{n-2}, \mathbf{X}_{n-1}, \mathbf{X}_n]$$

$$\left[\mathbf{X}^0\right] \supseteq \left[\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \ldots\ldots, \mathbf{X}^{n-2}, \mathbf{X}^{n-1}, \mathbf{X}^n\right]$$

$$[\mathbf{X}] \rightarrow \left[\mathbf{X}^{\mathbf{0}}\right]$$

$$[\mathbf{Y}] = [\mathbf{X}] \cup \left[\mathbf{X}^{\mathbf{0}}\right]$$

$$\text{Feature importance}[\mathbf{Y}] = \text{Z - score}$$

$$\text{If } \mathbf{X}^i \in \left[\mathbf{X}^{\mathbf{0}}\right], \text{Max feature importance}\left[\mathbf{X}^i\right] = \text{Z - max}$$

$\mathbf{X}_i \in [\mathbf{X}]$, so $\mathbf{X}_i \in [\mathbf{Y}]$, feature importance $[\mathbf{X}_i] =$ Z-score; when Z-score > Z-max, $\mathbf{X}_i =$ "Important"; If Z-score < Z-max, $\mathbf{X}_i =$ "Rejected".

$\mathbf{X}_1 \sim \mathbf{X}_n$ and $\mathbf{X}^1 \sim \mathbf{X}^n$ are the attribute indicators in the feature matrix.

In this paper, Z-score (important score) was defined as 0.6. In addition, some data was marked as "Tentative", which means importance of the data is not clear. To ensure reliability of feature filtering, we excluded the data marked "Tentative" and "Rejected".

The quantitative prediction model for DTIs affinity

The feature subsets was first pre-processed, and then combined with machine learning algorithms for construction of quantitative prediction models for DTIs affinity.

Pre-processing of feature subsets of descriptors

We normalized descriptors of the feature subsets in the range from -1 to 1. Meanwhile, the EC50 and Kd values that quantify DTIs affinity were processed in logarithmic form-Log_2 (Kd), Log_2 (EC50). In other words, we obtained feature subsets in which took Log_2 (Kd) and Log_2 (EC50) values characterize drug molecule-target affinity, respectively.

Construction of quantitative prediction model for DTIs affinity

The subsets obtained by feature selection were combined with random forest (RF) [59], support vector machine (SVM) [60] and artificial neural network (ANN) [61] to construct quantitative prediction model of DTIs affinity respectively. On the basic of ten-fold cross-validation, the feature subsets were randomly and equally divided into 10 data sets, where 9 groups of data were rotated as training sets for model construction, and the remaining 1 group of data will be used as a test set for model validation.

Evaluation and application of quantitative prediction model for DTIs affinity

The train and test validation were made use to wholly assess these models. Briefly, (1) the feature subsets were divided into 10 subsets randomly and equally as mentioned previously and 9 subsets were selected as training sets for modeling while the remaining subset served as test set for validating models. This process was repeated ten times until every subset served as test set. (2) Using different test sets to exert ten external independent validation. The nature of quantitative prediction models for DTIs affinity is regression models. Therefore, we used the Error Sum of Squares (SSE), Mean Square Error (MSE), Mean Absolute Error (MAE), Relative Mean Square Error (RMSE) and Coefficient of Determination (R^2) to evaluate the performance of models. The evaluation parameters can be expressed in the form as follows:

$$SSE = \sum (Y\_actual - Y\_predict)^2$$

$$MSE = \frac{1}{n}\sum_{i-1}^{n} (Y\_actual - Y\_predict)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n} (Y\_actual - Y\_predict)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Y\_actual - Y\_predict\right|$$

$$R^2 = 1 - \frac{\sum (Y\_actual - Y\_predict)^2}{\sum (Y\_actual - Y\_mean)^2}$$

Y_actutal and *Y_predict* denoted experimental value and predicted value, respectively. *n* is number of samples in the training sets or test sets. A higher R^2 value means model is more reliable. A lower MSE or SSE value means that model has higher accuracy.

Through above parametric evaluation to select the optimal models to predict quantitative affinity between drug molecules and targets collected in Binding DB database and comprehensive comparison were made with predictive models for DTIs affinity previously reported.

Abbreviations
ANN: Artificial neural network; DTIs: Drug–target interactions; MAE: Mean absolute error; MD: Molecular docking; MSE: Mean square error; QSAR: Quantitative structure–activity relationship; R^2: Coefficient of determination; RF: Random forest; RMSE: Relative mean square error; SSE: Error sum of squares; SVM: Support vector machine.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04389-w.

Additional file 1. Molecular descriptor information.

Additional file 2. Protein descriptor information.

Additional file 3. Implications of molecular descriptions related to molecular vibrations.

Acknowledgements
First, we are grateful to the editor and reviewers for their comments and suggestions. Then we thank National Natural Science Foundation of China for their support. Finally, we are grateful to open source databases such as PubChem, Drugbank and Uniprot.

Authors' contributions
XW and TC conducted data analysis and drafted manuscript. TC and CJ were responsible for collecting data and feature screen. XT performed data pre-processing. XW and YW designed the research. XW, TC and CJ were responsible for article correction and revision. HL and CJ were responsible for software technology. All authors read and approved the manuscript.

Funding
Publication costs are funded by National Natural Science Foundation of China under Grant Nos.81973495. The funder played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The algorithm processing and applications involved in this paper were all done in R (version 3.5.2). All data and materials such as raw data, EC50 dataset, Kd dataset, feature datasets, software, code, etc. used for quantitative prediction model construction and research results in this paper are available on open source data repository-Zenodo.org (https://zenodo.org/) with Zenodo_ numbers: 4699610 and 5510335.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 15 May 2021 Accepted: 20 September 2021
Published online: 14 October 2021

References

1. Suhail Y, Cain MP, Vanaja K, et al. Systems biology of cancer metastasis. Cell Syst. 2019;9(2):109–27.
2. Yeh SJ, Lin CY, Li CW, et al. Systems biology approaches to investigate genetic and epigenetic molecular progression mechanisms for identifying gene expression signatures in papillary thyroid cancer. Int J Mol Sci. 2019;20(10):2536.
3. Zhou M, Zheng C, Xu R. Combining phenome-driven drug–target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery. Bioinformatics. 2020;36(1):i436–44.
4. Fang J, Wu Z, Cai C, et al. Quantitative and systems pharmacology. 1. In silico prediction of drug–target interactions of natural products enables new targeted cancer therapy. J Chem Inf Model. 2017;57(11):2657–71.
5. Burstein B, Wieruszewski PM, Zhao YJ, et al. Anticoagulation with direct thrombin inhibitors during extracorporeal membrane oxygenation. World J Crit Care Med. 2019;8(6):87–98.
6. Zhou M, Chen Y, Xu R. A drug-side effect context-sensitive network approach for drug target prediction. Bioinformatics. 2019;35(12):2100–7.
7. Rothman RB, Baumann MH, Savage JE, et al. Evidence for possible involvement of 5-HT (2B) receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. Circulation. 2000;102(23):2836–41.
8. Marrugal-Lorenzo JA, Serna-Gallego A, Berastegui-Cabrera J, et al. Repositioning salicylanilide anthelmintic drugs to treat adenovirus infections. Sci Rep. 2019;9(1):17.
9. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8(1):573.

10. Chen H, Cheng F, Li J. iDrug: integration of drug repositioning and drug–target prediction via cross-network embedding. PLoS Comput Biol. 2020;16(7):e1008040.
11. Li J, Wu Z, Cheng F, et al. Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology. Sci Rep. 2014;4:5576.
12. Ivanov S, Lagunin A, Filimonov D, et al. Assessment of the cardiovascular adverse effects of drug-drug interactions through a combined analysis of spontaneous reports and predicted drug–target interactions. PLoS Comput Biol. 2019;15(7):e1006851.
13. Bagherian M, Sabeti E, Wang K, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform. 2021;22(1):247–69.
14. Wang H, Wang J, Dong C, et al. A novel approach for drug–target interactions prediction based on multimodal deep autoencoder. Front Pharmacol. 2020;10:1592.
15. Moumbock AFA, Li J, Mishra P, et al. Current computational methods for predicting protein interactions of natural products. Comput Struct Biotechnol J. 2019;17:1367–76.
16. Alaimo S, Pulvirenti A, Giugno R, et al. Drug–target interaction prediction through domain-tuned network-based inference. Bioinformatics. 2013;29(16):2004–8.
17. Simeon S, Jongkon N. Construction of quantitative structure activity relationship (QSAR) Models to predict potency of structurally diversed janus kinase 2 inhibitors. Molecules. 2019;24(23):4393.
18. Luo M, Wang XS, Roth BL, et al. Application of quantitative structure–activity relationship models of 5-HT1A receptor binding to virtual screening identifies novel and potent 5-HT1A ligands. J Chem Inf Model. 2014;54(2):634–47.
19. Van Den Driessche G, Fourches D. Adverse drug reactions triggered by the common HLA-B*57:01 variant: virtual screening of Drugbank using 3D molecular docking. J Cheminform. 2018;10(1):3.
20. Li Z, Han P, You ZH, et al. In silico prediction of drug–target interaction networks based on drug chemical structure and protein sequences. Sci Rep. 2017;7(1):11174.
21. Thafar MA, Olayan RS, Ashoor H, et al. DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. J Chem inform. 2020;12(1):44.
22. Guedes IA, Pereira FSS, Dardenne LE. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. Front Pharmacol. 2018;9:1089.
23. Li H, Leung KS, Wong MH, et al. Improving autodock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. Mol Inf. 2015;34(2–3):115–26.
24. Xu X, Huang M, Zou X. Docking-based inverse virtual screening: methods, applications, and challenges. Biophys Rep. 2018;4(1):1–16.
25. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–40.
26. Koehler LJ, Ulmschneider MB, Gray JJ. Computational modeling of membrane proteins. Proteins Struct Funct Bioinform. 2015;83(1):1–24.
27. Jones AJY, Gabriel F, Tandale A, et al. Structure and dynamics of GPCRs in lipid membranes: physical principles and experimental approaches. Molecules. 2020;25(20):4729.
28. Hutchings CJ, Colussi P, Clark TG. Ion channels as therapeutic antibody targets. MAbs. 2019;11(2):265–96.
29. Garcia-Chimeno Y, Garcia-Zapirain B, Gomez-Beldarrain M, et al. Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data. BMC Med Inform Decis Mak. 2017;17(1):38.
30. Jiang J, Wang N, Chen P, et al. DrugECs: an ensemble system with feature subspaces for accurate drug–target interaction prediction. Biomed Res Int. 2017;2017:6340316.
31. Krieger KL, Hu WF, Ripperger T, et al. Functional impacts of the BRCA1-mTORC2 Interaction in breast cancer. Int J Mol Sci. 2019;20(23):5876.
32. Hytönen VP, Määttä JA, Kidron H, et al. Avidin related protein 2 shows unique structural and functional features among the avidin protein family. BMC Biotechnol. 2005;5:28.
33. Cano G, Garcia-Rodriguez J, Garcia-Garcia A, et al. Automatic selection of molecular descriptors using random forest: application to drug discovery. Expert Syst Appl. 2017;72:151–9.
34. Wong WWL, Burkowski FJ. Using kernel alignment to select features of molecular descriptors in a QSAR study. IEEE/ACM Trans Comput Bioinform. 2011;8(5):1373–84.
35. Muller EA, Pollard B, Bechtel HA, et al. Nanoimaging and control of molecular vibrationss through electromagnetically induced scattering reaching the strong coupling regime. ACS Photon. 2018;5(9):3594–600.
36. Wang S. Intrinsic molecular vibrations and rigorous vibrationsal assignment of benzene by first-principles molecular dynamics. Sci Rep. 2020;10(1):17875.
37. Okabayashi N, Peronio A, Paulsson M, et al. Vibrationss of a molecule in an external force field. Proc Natl Acad Sci USA. 2018;115(18):4571–6.
38. Zhang QY, João AS. Structure-based classification of chemical reactions without assignment of reaction centers. J Chem Inform Model. 2005;45(6):1775–83.
39. Liu L, Zhu X, Ma Y, et al. Combining sequence and network information to enhance protein–protein interaction prediction. BMC Bioinform. 2020;21(Suppl 16):537.
40. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing. 2003;55:169–86.
41. Xie L, He S, Song X, et al. Deep learning-based transcriptome data classification for drug–target interaction prediction. BMC Genom. 2018;19(S7):667.
42. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics. 2018;34(7):1164–73.
43. Chen N, Chen J, Yao B, et al. QSAR study on antioxidant tripeptides and the antioxidant activity of the designed tripeptides in free radical systems. Molecules. 2018;23(6):1407.
44. Shar PA, Tao W, Gao S, et al. Pred-binding: large-scale protein–ligand binding affinity prediction. J Enzyme Inhib Med Chem. 2016;31(6):1443–50.

45. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821–9.
46. Abbasi WA, Yaseen A, Hassan FU, et al. ISLAND: in-silico proteins binding affinity prediction using sequence information. BioData Min. 2020;13(1):20.
47. Ding Y, Tang J, Guo F. Predicting protein–protein interactions via multivariate mutual information of protein sequences. BMC Bioinform. 2016;17(1):398.
48. Li ZR, Lin HH, Han LY, et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2006;34:W32–7.
49. Souza ES, Zaramello L, Kuhnen CA, et al. Estimating the octanol/water partition coefficient for aliphatic organic compounds using semi-empirical electrotopological index. Int J Mol Sci. 2011;12(10):7250–64.
50. Jiao L, Liu H, Qu L, et al. QSPR studies on the octane number of toluene primary reference fuel based on the electrotopological state index. ACS Omega. 2020;5(8):3878–88.
51. Wang C, Wang W, Lu K, et al. Predicting drug–target interactions with electrotopological state fingerprints and amphiphilic pseudo amino acid composition. Int J Mol Sci. 2020;21(16):5694.
52. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–9.
53. Wishart DS, Feunang YD, Guo AC, et al. Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
54. Bühlmann S, Reymond JL. ChEMBL-likeness score and database GDBChEMBL. Front Chem. 2020;8:46.
55. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.
56. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
57. Rao HB, Zhu F, Yang GB, et al. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2011;39:W385–90.
58. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36(11):1–13.
59. Dai JY, LeBlanc M. Case-only trees and random forests for exploring genotype-specific treatment effects in randomized clinical trials with dichotomous endpoints. J R Stat Soc Ser C Appl Stat. 2019;68(5):1371–91.
60. Xu L, Liang G, Shi S, et al. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. Int J Mol Sci. 2018;19(6):1773.
61. Świetlik D, Białowąs J. Application of artificial neural networks to identify Alzheimer's disease using cerebral perfusion SPECT data. Int J Environ Res Public Health. 2019;16(7):1303.

Publisher's Note