Quantitative Prediction Model for Affinity of Drug-Target Interactions Based on Molecular Vibration and Overall System of Ligand-Receptor

xian rui wang

BUCM: Beijing University of Chinese Medicine

ting ting cao

BUCM: Beijing University of Chinese Medicine

cong min jia

BUCM: Beijing University of Chinese Medicine

xue mei tian

BUCM: Beijing University of Chinese Medicine

Yun Wang (✉ wangyun@bucm.edu.cn)

BUCM: Beijing University of Chinese Medicine

Research article

Keywords: molecular vibrations, random forest, drug-target affinity, chemical composition, drug-target interactions

Posted Date: June 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-641126/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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 were based either only on structure of drug molecules or on structure of targets with low accuracy and small scope of application. How to construct quantitative prediction models with high accuracy with 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 model with high accuracy-wide applicability based on K_d and EC_{50} , and to provide reference for quantifying affinity of DTIs.

Methods: Through parametric characterization based on molecular vibrations and protein sequences, taking molecule-target as whole system and feature selection of drug molecule-target, we constructed feature datasets of DTIs quantified by K_d and EC_{50} , respectively. Then, prediction models were constructed using above datasets and SVM, RF and ANN. In addition, optimal models were selected for application evaluation and comprehensive comparison.

Results: Under ten-fold cross-validation, evaluation parameters based on RF for EC_{50} dataset are as follows: R^2 (RF) of training and test sets are 0.9611, 0.9641; MSE (RF) of training and test sets are 0.0891, 0.0817. Evaluation parameters based on RF for K_d dataset are as follows: R^2 (RF) of training and test sets are 0.9425, 0.9485; MSE (RF) of training and test sets are 0.1208, 0.1191. After comprehensive comparison, the results showed that RF model in this paper is optimal model. In application evaluation of RF model, the errors of most prediction results were in range of 1.5-2.0.

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

1. 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 interaction between drug molecule and target-Drug-Target Interactions (DTIs). DTIs plays an important role in pharmacology, biology and mechanism [3–6]. For example, on the basis of DTIs research, the off-target toxicity of appetite suppressant Fen-Phen that can cause death is due to the activation of 5-HT_{2B} 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 result 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, research on DTIs focused on two directions, one is traditional experimental analysis and the other is DTIs prediction based on existing databases combined with learning algorithms [13]. Traditional experimental analysis of DTIs are expensive and inefficient and face 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 whether 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 includes 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-HT_{1A} Receptor, in which K_i 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 was often built for specific targets or several targets, making it difficult to achieve quantitative predictions 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 protein [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]. This limitation is severe because most useful drug targets are membrane proteins, such as ion channels and G protein-coupled receptors (GPCRs) [27, 28]. This led 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 ignored holistic nature of receptor-ligand interactions, which lead to low prediction accuracy and excessive bias. In addition, in constructing quantitative predictive 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 also led 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 was 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.

2. Results And Discussions

2.1 Data Collection

Based on multiple databases mentioned in 4.1, 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 drug molecule-target 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 11076 ligand-target-EC₅₀ pairs. At the same time, The Kd dataset contains 1870 ligands and 778 targets, and 10923 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 prediction models for DTIs affinity. It is worth noting that half-maximal effective concentration (EC50) refers to the concentration of a drug, antibody or toxicant that induces a response halfway between the baseline and maximum after a specified exposure time. It was commonly used as a measure of a drug's potency [31]. Dissociation constants (Kd) are often used to describe degree of binding of a ligand to a particular protein [32]. The smaller dissociation constant, the tighter ligand binding, otherwise the higher affinity between ligand and protein. Considering the practical significance of Kd and EC50, we finally chose both as quantitative indexes of DTIs affinity.

2.2 Results of descriptor calculation

2.2.1 Calculation of drug molecule descriptors

After calculation by PaDEL software, 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

Loading [MathJax]/jax/output/CommonHTML/jax.js 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.

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

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 method does not take into account properties of drug molecules and that may not be applicable in quantitative prediction of drug molecule-target interactions.

After comprehensive consideration, in this paper, based on properties of drug molecules, we screened characteristic descriptors of drug molecules from the perspective of molecular vibrations. This is because molecular vibrations was caused by vibrations of chemical bonds within molecule and molecular vibrations is a macroscopic representation of properties of drug molecules [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 remember that seven physicochemical properties are particularly relevant to the nature of chemical bonds, including electronegativity, π -atomic charge, total charge, and bond polarity [38]. Selecting all descriptors related with these seven properties based on meaning of each descriptor. For instance, Mpe-Constitution Descriptor-mean Atomic Pauling Electronegativity (scaled on carbon atom) was selected as feature descriptors to construct prediction models for DTIs affinity due to it was related 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 molecule.

2.2.2 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 its accuracy [39]. The target protein descriptors were shown in Table 2.

Table 2 Type and number of target protein descriptors		
Serial number	Descriptor type	Number of descriptors
1	Amino acid composition	20
2	Dipeptide composition	400
3	Normalized Moreau-Broto autocorrelation	240
4	Moran autocorrelation	240
5	Geary autocorrelation	240
6	Composition	21
7	Transition-Distribution	126
8	Sequence-order-coupling number	60
9	Quasi-sequence-order descriptors	100

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

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

2.3 Results of feature screening

The Boruta algorithm was used for feature filtering. If a feature attribute was marked as "Confirmed", it means that the attribute is "important". On the contrary, if feature attribute was marked as "Rejected", it

means that the attribute is "Not Important". 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". As shown in Fig. 1, 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 DTIs in the integrated EC50 dataset was characterized by 1259 feature attributes. Similarly, as shown in Fig. 2, 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 DTIs in integrated EC50 dataset was characterized by 827 feature attributes.

In the process of feature selection, we chose Boruta algorithm because this algorithm can filter the set of features that are associated with the response variable, rather than selecting the set of features that minimizes penalty factor only for a specific model.

The feature subsets of EC50 and Kd were obtained by feature screening for construction of quantitative prediction models for DTIs affinity.

2.4 Results of quantitative prediction model for DTIs affinity

2.4.1 Parameter optimization

The setting of algorithm parameters is crucial to construction of prediction models for DTIs affinity. In RF model, there are two important parameters need to be considered: Ntree and Mtry. After comparison and optimization of several parameters, we finalized RF algorithm parameters: Ntree = 500, Mtry = default value; As 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]. Kernel selected radial basis kernel function to produce minimum error rate. In same optimization way, ANN algorithm parameters were determined: size = 2, decay = 0.1, linout = T (non-linear function), maxit = 1000, the rest of parameters were default values.

3.4.2 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 10-fold cross validation for EC50 feature subset were shown in Table 3.

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

Model (EC ₅₀)	R ²		MSE		SSE	
	Training	Test	Training	Test	Training	Test
SVM	0.9317	0.5759	0.1270	0.8356	1249	8216
RF	0.9611	0.9641	0.0891	0.0817	876	803.3
ANN	0.7350	0.5211	0.4867	0.9590	4785	9429

As shown in Table 3 and Fig. 3, In RF model, R² 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. 3-a). As for SVM model, R² of training and test sets are 0.9317, 0.5759 respectively. SVM model exhibited some differ for training and test sets, but order of magnitude is the same and no greatly obvious overfitting can be observed from SVM model (Fig. 3-b). However, predictive performance of SVM model worse than that of RF model. For training and test sets in ANN model, no obvious overfitting can be observed (Fig. 3-c), 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 EC₅₀ data.

The same analysis was appropriate for Kd data, on the basic of data in Table 4 and scatter plot in Fig. 4, we completed selection of optimal model: RF model showed satisfactory predictive performance with R² of test set being 0.9485 (Fig. 4-a). The SVM model suffered from overfitting and its predictive performance was worse than that of RF model (Fig. 4-b). ANN models are the least effective model (Fig. 4-c). The results indicated that RF model was the optimal quantitative prediction model for KD data.

Table 4
Tenfold cross validation of three kinds of algorithms

Model (KD)	R ²		MSE		SSE	
	Internal	External	Internal	External	Internal	External
SVM	0.9099	0.5083	0.1254	0.7290	1230	808.4
RF	0.9425	0.9485	0.1208	0.1191	1204	132.1
ANN	0.5857	0.2961	0.5612	1.0190	5593	1130

In summary, whether based on EC50 data or Kd data, the performance of RF model was the best. Therefore, in this paper, random forest (RF) model is more suitable for quantitative prediction of

2.5 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 model further, we used RF model 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- EC_{50} 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 optimal models based on Kd and EC50. Calculating absolute value of the difference between true value and predicted value- $|d|$ and dividing $|d|$ into 5 parts in which each part was divided on a scale of 0.5. Therefore, we obtained the results of distribution of $|d|$ (Fig. 5 and Fig. 6) in new EC50 and Kd dataset, reflecting prediction capability of RF model.

The predictive values of RF models were all greater than zero, suggesting that drug molecule-target interactions do exist, which is consistent with the data information gathered from datasets. This indicated that optimal model constructed in this paper could be accurately used for qualitative prediction of DTIs. However, as shown in Fig. 5 and Fig. 6, eighty percent of the $|d|$ 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. Further comparison of predicted and experimental data revealed that all predicted values are greater than experimental true values and within a certain margin of error. The reason for that maybe a systematic error due to different standards used to store data in different databases. The case can be set with a correction factor - average of all difference values. 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.

2.6 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 many reports for predicting DTIs, such as Xie L, et al., adopted transcriptome data and 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 methods, even current state of prediction analysis for DTIs, there are often only analysis based on structure of ligand or receptor rather than taking ligand-receptor as a whole system for DTIs analysis. This method of analysis, which separated ligands from receptors, can be limited by its own structure and produce non-reciprocal results, leading to poor accurate. Conversely, in this paper, the model was constructed to take full account of ligands and receptors. From perspective of taking molecule-target as a whole system, we integrated molecule-target for DTIs affinity, which is able to avoid unequal results based

on receptors or ligands only, thus increasing accuracy of prediction model. 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 specific targets 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 modes 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 model 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 model in this paper. The MSE of CNN model was high than 0.194, which was high than that of RF model 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 result showed that RF model developed based on K_d and EC_{50} 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 model for DTIs affinity.

3. Conclusion

In this paper, from perspective of overall systematic of ligand-receptor, after screening molecular descriptors associated with molecular vibrations, and feature selection, we constructed two feature subsets in which EC_{50} and K_d values were used to quantify affinity of the DTIs respectively. On the basis of that, we applied RF, SVM, and ANN for constructing quantitative prediction model for DTIs and the optimal models were selected for comprehensive application evaluation. The results showed that comparing to SVM, ANN models in this paper and previously reported models, RF models had more accurate and more widely applicable. Moreover, quantitative prediction of affinity for DTIs based on K_d or

Loading [MathJax]/jax/output/CommonHTML/jax.js

 which also showed a certain reliability. It can provide a

reference for DTI's affinity prediction. In addition, 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.

4. Methods And Materials

In this paper, we constructed prediction model for DTIs affinity from the perspective of taking molecule-target as a whole system. Firstly, drug molecules and protein sequences of targets as well as ligand-receptor-Kd/EC50 were screened on the basis 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 descriptor obtained in step 2, we constructed Kd and EC50 quantified drug molecule-target feature datasets by taking drug molecules and targets as a whole system, respectively. Finally, combining above datasets with machine learning algorithms SVM, RF, ANN for construction of prediction models of DTIs affinity. The research methodology was shown in Fig. 1.

4.1 Datasets

This paper carried out construction of prediction model of DTIs affinity, which requires a large amount of data support. The drug molecules (ligand) 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/>) [47–49]. The target protein sequences (receptor) were collected from open source Uniprot database (<https://www.uniprot.org/>) [50]. 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.

4.2 Drug molecules and target sequence descriptors

Descriptors can effectively solve problem of parametric characterization of drug molecules and protein sequences of targets, which facilitate the construction of predictive models for DTIs affinity. In this paper, using PaDEL to calculate the descriptors of drug molecules [51]. Each descriptor calculated by PaDEL 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>) [52, 53].

4.3 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.

4.4 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.

4.4.1 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 affinity of drug molecules-targets were processed in logarithmic form-Log₂ (Kd), Log₂ (EC50). In other words, we obtained feature subsets in which took Log₂ (Kd) and Log₂ (EC50) values characterize drug molecule-target affinity, respectively.

4.4.2 Construction of quantitative prediction model for DTIs affinity

The subsets obtained by feature selection were combined with random forest (RF) [54], support vector machine (SVM) [55] and artificial neural network (ANN) [56] 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.

4.5 Evaluation and application of quantitative prediction model for DTIs affinity

The internal and external 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 model for DTIs affinity is regression model. Therefore, we used the Error Sum of Squares (SSE), Mean Square Error (MSE) and Coefficient of Determination (R²) to evaluate the performance of models. The R², MSE and SSE can be expressed in the form as follows:

$$SSE = \sum (Y_{actual} - Y_{predict})^2$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_{actual} - Y_{predict})^2$$

$$R^2 = 1 - \frac{\sum (Y_{actual} - Y_{predict})^2}{\sum (Y_{actual} - Y_{mean})^2}$$

Y_{actual} and $Y_{predict}$ denoted experimental value and predicted value, respectively. n is number of samples in the training sets or test sets. A higher R² value means model is more reliable. A lower MSE or

SSE value means model has higher accuracy.

Loading [MathJax]/jax/output/CommonHTML/jax.js

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

Abbreviations	Full names
ANN	Artificial Neural Network
DTIs	Drug-target interactions
MD	Molecular Docking
MSE	Mean Square Error
QSAR	Quantitative Structure-Activity Relationship
R ²	Coefficient of Determination
RF	Random Forest
SSE	Error Sum of Squares
SVM	Support Vector Machine

Declarations

Conflict of Interest

We declare that there are no financial and personal interest conflict with other people or organizations in the manuscript and the exposition of above article.

Ethics Approval and Consent to Participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The algorithm processing and applications involved in this paper were all done in R (version 3.5.2). The data used for quantitative prediction model construction in this paper are available at <https://zenodo.org/record/4699610>.

Funding

Loading [MathJax]/jax/output/CommonHTML/jax.js

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.

Authors' contributions

XW and TC conducted data analysis and drafted manuscript. TC and CJ were responsible for collecting data. XT performed data pre-processing. XW and YW designed the study. HL were responsible for software technology. All authors read and approved the manuscript.

Acknowledgement

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.

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(Suppl-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 RA, 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-HT_{1A} receptor binding to virtual screening identifies novel and potent 5-HT_{1A} 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 Bioinf*. 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 on Comput Bio Bioinform*. 2011;8(5):1373–84.
35. Muller EA, Pollard B, Bechtel HA, et al. Nanoimaging and Control of Molecular Vibrations through Electromagnetically Induced Scattering Reaching the Strong Coupling Regime. *ACS Photonics*. 2018;5(9):3594–600.
36. Wang S. Intrinsic molecular vibrations and rigorous vibrational assignment of benzene by first-principles molecular dynamics. *Sci Rep*. 2020;10(1):17875.
37. Okabayashi N, Peronio A, Paulsson M, et al. Vibrations 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 Infor 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 Bioinformatics*. 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. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*. 2019;47(D1):D1102–9.
48. 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.
49. Bühlmann S, Reymond JL. ChEMBL-Likeness Score and Database GDBChEMBL. *Front Chem*. 2020;8:46.
50. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480–9.
51. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32(7):1466–74.
52. 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.
53. 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(Web Server issue): W385-W390.
54. 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.
55. 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.
56. Świetlik D, Białowas 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.

Figures

Figure 1

Feature filtering results for EC50 dataset

Figure 2

Feature filtering results for Kd dataset

Loading [MathJax]/jax/output/CommonHTML/jax.js

Figure 3

Scatter plot of experimental and predicted EC50 values for three prediction models

Figure 4

Scatter plot of experimental and predicted Kd values for three prediction models

Figure 5

Distribution results of $|d|$ for EC50 dataset

Figure 6

Distribution results of $|d|$ for Kd dataset

Figure 7

The specific research methodology

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [highlight.pdf](#)
- [tableS1.docx](#)
- [tableS2.docx](#)