Bioinformatics, 38(7), 2022, 1964–1971
https://doi.org/10.1093/bioinformatics/btac048
Advance Access Publication Date: 4 February 2022
Original Paper

Systems biology

NerLTR-DTA: drug–target binding affinity prediction based on neighbor relationship and learning to rank

Xiaoqing Ru [1,3], Xiucai Ye [1,*], Tetsuya Sakurai [1] and Quan Zou [2,3,*]

[1]Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan, [2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China and [3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang 324000, China

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on October 29, 2021; revised on December 20, 2021; editorial decision on January 14, 2022

Abstract

Motivation: Drug–target interaction prediction plays an important role in new drug discovery and drug repurposing. Binding affinity indicates the strength of drug–target interactions. Predicting drug–target binding affinity is expected to provide promising candidates for biologists, which can effectively reduce the workload of wet laboratory experiments and speed up the entire process of drug research. Given that, numerous new proteins are sequenced and compounds are synthesized, several improved computational methods have been proposed for such predictions, but there are still some challenges. (i) Many methods only discuss and implement one application scenario, they focus on drug repurposing and ignore the discovery of new drugs and targets. (ii) Many methods do not consider the priority order of proteins (or drugs) related to each target drug (or protein). Therefore, it is necessary to develop a comprehensive method that can be used in multiple scenarios and focuses on candidate order.

Results: In this study, we propose a method called NerLTR-DTA that uses the neighbor relationship of similarity and sharing to extract features, and applies a ranking framework with regression attributes to predict affinity values and priority order of query drug (or query target) and its related proteins (or compounds). It is worth noting that using the characteristics of learning to rank to set different queries can smartly realize the multi-scenario application of the method, including the discovery of new drugs and new targets. Experimental results on two commonly used datasets show that NerLTR-DTA outperforms some state-of-the-art competing methods. NerLTR-DTA achieves excellent performance in all application scenarios mentioned in this study, and the $r^2_{m(\text{test})}$ values guarantee such excellent performance is not obtained by chance. Moreover, it can be concluded that NerLTR-DTA can provide accurate ranking lists for the relevant results of most queries through the statistics of the association relationship of each query drug (or query protein). In general, NerLTR-DTA is a powerful tool for predicting drug–target associations and can contribute to new drug discovery and drug repurposing.

Availability and implementation: The proposed method is implemented in Python and Java. Source codes and datasets are available at https://github.com/RUXIAOQING964914140/NerLTR-DTA.

Contact: yexiucai@cs.tsukuba.ac.jp or zouquan@nclab.net

1 Introduction

Drugs are compounds that can bind to protein targets and alter disease states. Drug discovery has always been a topic of great concern because of the necessity of drugs to treat diseases (Bahuguna and Rawat, 2020; Theodoris *et al.*, 2021). Recently, with the large-scale outbreak of COVID-19, drug repurposing, which is characterized by shorter research cycles and lower costs (Liu *et al.*, 2021; Pushpakom *et al.*, 2019), has become an important means of discovering drugs to treat this disease. Thousands of FDA-approved drugs on the market and late-clinical drugs may have undiscovered targets (Xia *et al.*, 2010). Nearly, 70 FDA-approved drugs are being explored to determine whether they can treat COVID-19 (O'Meara *et al.*, 2020). Overall, drug repurposing and discovery can provide significant guarantees for saving lives and improving quality of life. Exploring which proteins can be targeted by which drugs is an important step in drug discovery and drug repurposing (Chen *et al.*, 2016; Ezzat *et al.*, 2019). Therefore, it is meaningful to develop drug–target association prediction studies.

With the development of sequencing technology, huge amounts of protein data have been generated. The functions of many sequences have not been determined by experimental methods, which means that such a large-scale protein sequence may contain significant proteins that can inhibit or promote the occurrence of diseases

(Chen *et al.*, 2016; Yamanishi *et al.*, 2008). Similarly, there are many unexplored compounds. For example, the PubChem database contains 111 million compounds, including many entries whose interactions with proteins are unknown, and among them there may be many drugs with good curative effects (Kim *et al.*, 2021). Therefore, obtaining drug–target association knowledge through computational methods has gained widespread interest. Machine learning can use limited data to complete large-scale predictions, and many machine learning models for predicting drug–target interactions have been proposed (Bleakley and Yamanishi, 2009; Gönen, 2012; Mousavian and Masoudi-Nejad, 2014; Yamanishi *et al.*, 2008).

At present, many studies regard drug–target association prediction as a binary classification task (Bleakley and Yamanishi, 2009; Cao *et al.*, 2014; Öztürk *et al.*, 2016), that is, to explore whether compounds and proteins can interact or not. Various principles, characteristics and algorithms have been used in such tasks to build models. For example, MultiDTI (Zhou *et al.*, 2021) proposes a joint learning framework based on heterogeneous networks to make full use of the sequence information of drugs/targets and the interaction or association information in heterogeneous networks containing drugs, targets, side effects and diseases. Ding *et al.* (2017) use drug molecular substructure fingerprints, protein multivariate mutual information and network topology to characterize the properties of drugs, targets and drug–target pairs. iDTI-ESBoost (Rayhan *et al.*, 2017) uses structure, evolutionary characteristics and an AdaBoost classifier to predict drug–protein interactions. RFDT (Wang *et al.*, 2018) is a predictor based on the rotating forest algorithm. It encodes protein sequences as position-specific scoring matrix (PSSM) descriptors and encodes drug molecules as fingerprint feature vectors. Liu *et al.* (2015) improve the prediction of compound-protein interactions by establishing highly reliable negative samples. NRLMF (Liu *et al.*, 2016) proposes a neighborhood regularized logistic matrix factorization method to predict drug–protein interactions. Tsubaki *et al.* (2019) predict compound–protein interaction by combining a graph neural network for compounds and a convolutional neural network for proteins.

Binding affinity indicates the strength of drug–target interactions. Predicting drug–target binding affinity (DTA) helps to select a small number of drug–target interaction pairs that are conducive to later experimental verification and provide promising candidates for biologists (Corsello *et al.*, 2017; Ragoza *et al.*, 2017). KronRLS (Pahikkala *et al.*, 2015) builds an affinity prediction model based on the kronecker regularized least squares algorithm and evaluates the impact of different chemical structures and sequence similarities on the prediction accuracy. SimBoost (He *et al.*, 2017) is a novel nonlinear method that inputs three types of features based on drug/protein similarity and drug–target affinity information into the gradient boosting regression tree to predict DTA. DeepDTA (Öztürk *et al.*, 2018) only uses the sequence information of the drug/target and convolutional neural network to construct a model with good performance. WideDTA (Öztürk *et al.*, 2019) uses protein sequence, ligand SMILES, protein domains and motifs and maximum common substructure words to predict DTA. ML-DTI (Yang *et al.*, 2021) introduces a mutual information mechanism to improve the performance and comprehensibility of DTA prediction. OnionNet (Zheng *et al.*, 2019) extracts features based on rotation-free element-pair-specific contacts between ligands and protein atoms and uses deep convolutional neural network to predict DTA. Although previous research has achieved improved results, there are still some challenges. Many existing methods only predict the missing relationship between drugs and proteins for which part of the drug–target pair information is known, without paying attention to new drugs/targets that do not have any interaction information. Moreover, ranking the candidates according to the interaction strength helps to select a small number of promising drug–target pairs, but this point is not considered in these methods.

In this study, we treat DTA prediction as a search ranking task, aiming to solve these problems through a ranking framework. We propose a new method called NerLTR-DTA, which effectively completes the related investigation of the DTA relationship. First, NerLTR-DTA extracts features with the help of neighboring drugs (or proteins) information, these neighbors meet certain similarity and sharing with the target drug (or protein), and then treats these features as the input to the learning to rank (LTR) algorithm. Taking advantage of the one-to-many relationship between the query and the document in LTR, NerLTR-DTA realizes multi-scenario applications by setting different query types and dividing different types of training and test sets. The multiple scenarios implemented include drug repurposing, protein new function prediction, new drug discovery and new target discovery.

Davis and KIBA datasets are used to evaluate NerLTR-DTA. The experimental results show that in the common scenario of drug repurposing, NerLTR- DTA is superior to some state-of-the-art DTA prediction methods. Furthermore, NerLTR-DTA can also achieve good performance in other three scenarios, and the $r^2_{m(\text{test})}$ values indicate that models in all scenarios are robust and acceptable. Finally, thanks to the characteristics of the search ranking task, this model can provide a ranking list of related proteins (or drugs) for each query drug (or query protein). Through the statistics of the association relationship of each query drug (or query protein), it can be observed that NerLTR-DTA can accurately rank the relevant results of most queries.

2 Materials and methods

2.1 Datasets

We perform our study on Davis *et al.* (2011) and KIBA (Tang *et al.*, 2014) datasets, which the samples included are kinase-inhibitor pairs and have been used in many DTA prediction studies. Table 1 lists basic information of these two datasets.

In this study, we process the affinity data as in SimBoost (He *et al.*, 2017), DeepDTA (Öztürk *et al.*, 2018) and GraphDTA (Nguyen *et al.*, 2021). For KIBA, the affinity values are processed by taking the negative of each value, and adding the absolute value of the minimum to all negative values. For Davis, the affinity values are transformed into log space, as shown in formula (1). The major advantage of the processed data is that they make the experimental results more intuitive and easier to compare.

$$pK_d = -\log_{10}^{\frac{K_d}{1e9}} \tag{1}$$

2.2 Method overview

The framework of our study with five main steps (S1. sample collection and processing, S2. dataset division, S3. feature extraction, S4. model training, S5. model test) is illustrated in Figure 1a. Formulating the DTA prediction as a search ranking task, we explain our method by using the process of searching and ranking the targeted proteins for a new drug, as shown in Figure 1b.

2.3 Feature engineering

Five types of associations are used to extract drug and protein features, including drug–drug similarity, drug–drug sharing, protein–protein similarity, protein–protein sharing, drug–protein binding affinity. The notations of these associations are shown in Figure 2a.

In this study, $Sim^{D-D}_{i,j}$ is the 2D chemical structure similarity between drug i and drug j, calculated by using the structure clustering server at PubChem. $Sim^{P-P}_{i,j}$ is the sequence similarity between protein i and protein j, represented by normalized Smith-Waterman

Table 1. Basic information of the datasets

Dataset	No. of Proteins	No. of Compounds	Affinity measure	No. of Pairs
Davis	442	68	K_d	30056
KIBA	229	2111	K_d, K_i, IC_{50}	118254

Note: Low K_d, K_i, IC_{50} values indicate high binding affinity (Benson *et al.*, 2007). K_d, the dissociation constant; K_i, the inhibition constant; IC_{50}, the half-maximal inhibitory concentration.

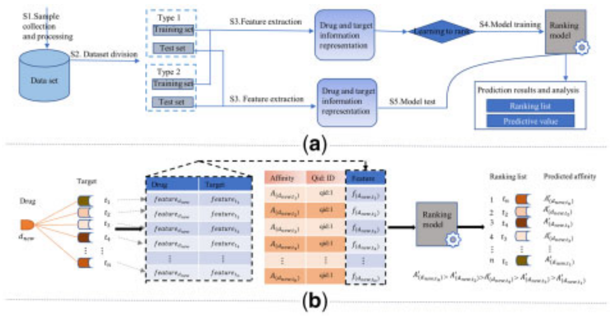

Fig. 1. The framework of this study. (**a**) The five main steps. S1. Two datasets from previous studies are used in this study. S2. The samples are divided into different types of test and training sets to explore the performance of NerLTR-DTA in multiple scenarios. S3. The properties of drugs and proteins are represented in numerical form. S4. The features are fed into LTR algorithm to train model. S5. The performance of the ranking model is tested, a ranking list and predicted drug–target affinity can be provided by ranking model. (**b**) Procedures of our method for predicting DTA with a given new drug. To obtain the list of proteins that a new drug d_{new} can target and the specific affinity values between them. We calculate the feature vector of drug d_{new} and proteins $t_{1\sim n}$ and organize the information into the form of *[affinity, qid: id, feature]*, and then input these information into the ranking model. Finally, the affinity values of $d_{\text{new}} - t_{1\sim n}$ and the protein list ranked according to the predicted affinity value can be returned

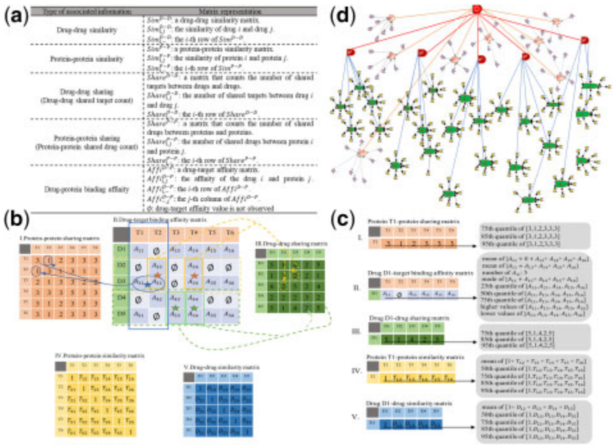
(a)
Type of associated information
Matrix representation
Drug-drug similarity
Protein-protein similarity
Drug-drug sharing (Drug-drug shared target count)
Protein-protein sharing (Protein-protein shared drug count)
Drug-protein binding affinity
(b)
(c)
(d)

Fig. 2. Related knowledge involved in feature engineering. (**a**) Notations of five associations used in feature extraction. (**b**) Matrix representation of associated information and examples of calculation method for sharing information. (**c**) Self-associated feature representation of drug D1 and the setting threshold used in this study. (**d**) The determination process of 25 additional neighbors to the target object. Note: To improve comprehensibility, this part does not show similar drug (protein) relationships that may overlap between drugs (proteins). Straight lines connect the same substance, double arrow lines connect different substances

value. $Share_{i,j}^{D-D}$ is the number of shared targets between drug i and drug j. $Share_{i,j}^{P-P}$ is the number of shared drugs between protein i and protein j. The calculation methods of protein–protein sharing and drug–drug sharing are explained in Figure 2b, specifically, each element in the matrix is the number of the same protein (drug) that two drugs (two proteins) can target.

Based on these five associations, we define self-associated features and adjacent associated features to characterize the properties of drugs and targets.

2.3.1 Self-associated features (SAF)

SAF is derived from the similarity and affinity associated with the object (drug or target) itself. Table 2 summarizes the specific representation of SAF.

2.3.2 Adjacent associated features (AAF)

AAF is motivated and inspired by two common assumptions. (i) Similar drugs tend to target similar proteins, and vice versa. (ii) Drug d_s can bind to a new target, it is likely that the drug d_a that shares numerous targets with d_s can also bind to that new target. Inspired by assumption (i), SimBoost(He *et al.*, 2017), which is the first non-linear read-across method for DTA prediction, uses weight scores from drug(protein) to the neighbors of target (drug) as features. Following the two common assumptions and SimBoost (He *et al.*, 2017), we extract AAF using ASAF of the k neighbors that meet high similarity and high sharing with the object (drug or target). Furthermore, AAF are divided into general AAF (GAAF) and additional AAF (AAAF).

GAAF: For drug i (or target j), we first select a set of drugs (or targets) that similarity/sharing with drug i (or target j) are greater than the prescribed threshold, then concatenate ASAF of this set of drugs (or targets) as initial vector, finally take the minimum, maximum, upper quartile, median, lower quartile, mean and mode of the elements in initial vector as GAAF of drug i (or target j).

In this study, all the thresholds used are shown in Figure 2c. The value of each dimension feature in SSAF, that is, each special value of all similarity scores for drug i/target j in $Sim_{i,-}^{D-D}/Sim_{j,-}^{P-P}$, is used as threshold to select the general similarity neighbors of drug i (or target j). The 75th , 85th , 95th quantile of the $Share_{i,-}^{D-D}/Share_{j,-}^{P-P}$ are used to obtain general sharing neighbors.

AAAF: we concatenate ASAF of 25 additional neighbors as AAAF. The selection process of 25 additional neighbors is depicted in Figure 2d. For drug i (or target j), we first determine the top 5 most similar drugs (or targets), these 5 drugs (or targets) are represented by O^1–O^5, then obtain the top 5 most similar drugs (or targets) to O^1–O^5, respectively. Therefore, we get 25 additional neighbors of drug i (or target j). It is worth noting that drug i (or target j) is not included in each case.

2.4 Learning to rank (LTR)

LTR is originally applied in information retrieval, with the purpose of receiving the required information quickly and accurately from a huge amount of information data with subtle relationships (Trotman, 2005). At present, it has been applied to many fields of bioinformatics, such as protein-phenotype association prediction (Liu *et al.*, 2020), circRNA-disease association prediction (Wei *et al.*, 2021), protein remote homology detection (Chen *et al.*, 2017; Jin *et al.*, 2021) and drug–target interaction detection (Yuan *et al.*, 2016). The ultimate intent of DTA prediction is to find a small number of proteins (or compounds) that strongly bind to the target drug (or protein). Therefore, LTR is suitable for DTA prediction research. LTR is mainly divided into three types: pointwise, pairwise and listwise (Burges *et al.*, 2005; Cao *et al.*, 2007; Xia *et al.*, 2008). In search task, pointwise focuses on a point-by-point score regression, which outputs the URL ranking list by using the relevance of each query-URL pair. Pairwise applies the relevance of each query-URL pair and the priority between query-URL pairs to output URL ranking list, that is, pairwise shows solicitude for pairing preference satisfaction. The URL ranking list output by listwise mainly relies on the overall ranking of all query-URL relevance corresponding to each query.

In this study, we use multiple additive regression trees (MART), which have been implemented in Ranklib (https://sourceforge.net/p/lemur/wiki/RankLib/), for training model. As a pointwise algorithm with regression properties, MART can not only rank the relevant samples under each query, but also fit the predicted value and the real value as much as possible.

The required input data format of LTR is *[affinity, qid: id, feature]*, where *affinity* represents the affinity value of the drug–target pair. In this study, the processed K_d, K_i, IC_{50} are used as *affinity*. *qid: id* represents the query id corresponding to a given drug (or protein). We assign a unique id to each drug (or protein) for query work. For example, all drug–target pairs $(D_n - T_1, D_n - T_2, D_n - T_3, \ldots, D_n - T_x)$ corresponding to drug D_n are assigned the same *qid: n*. *feature* is the features of drugs and targets, we concatenate SAF and AAF of drugs and targets as *feature*.

Table 2. Specific representation of SAF

SAF type	Specific representation (For drug *i*/target *j*)
SSAF	Special values of all similarity scores for drug *i*/target *j* in Sim^{D-D}/Sim^{P-P} $\left.\begin{array}{c}\textit{the mean}\\ \textit{the 50th quartile}\\ \textit{the 75th quartile}\\ \textit{the 85th quartile}\\ \textit{the 95th quartile}\end{array}\right\}$ of $Sim^{D-D}_{i,-}/Sim^{P-P}_{j,-}$
ASAF	The mean of affinity values between drug *i* -all proteins/target *j*-all drugs Special values of the observed values for drug *i*/target *j* in $Affi^{D-P}$ $\left.\begin{array}{c}\textit{the mean}\\ \textit{the number}\\ \textit{the mode}\\ \textit{the 25th quartile}\\ \textit{the 50th quartile}\\ \textit{the 75th quartile}\\ \textit{5 higher values}\\ \textit{5 lower values}\end{array}\right\}$ of $Affi^{D-P}_{i,-}/Affi^{D-P}_{-,j}$

SSAF, SAF extracted using drug–drug (or protein–protein) similarity; ASAF, SAF extracted using DTA.

2.5 Evaluation criteria

The performance of NerLTR-DTA is evaluated by three metrics: Concordance Index (CI), Mean Square Error (MSE) and $r^2_{m(\text{test})}$.

CI measures whether the predicted order of two random drug–target pairs is same as their real order, where the predicted order and real order are determined by comparing the predicted and real affinity values of the two drug–target pairs, respectively (Steck *et al.*, 2008). CI $\in [0.50, 0.70]$ means low accuracy, CI $\in [0.71, 0.90]$ means medium accuracy, and CI $\in [0.91, 1.0]$ means high accuracy. The formula is as follows (Gönen and Heller, 2005):

$$\mathrm{CI} = \frac{1}{M}\sum_{y_i > y_j} \rho(p_i - p_j) \tag{2}$$

and,

$$\rho(x) = \begin{cases} 1, \ x > 0 \\ 0.5, \ x = 0 \\ 0, \ x < 0 \end{cases} \tag{3}$$

where M is a normalization constant. y_i and y_j represent the larger and smaller affinity values, respectively. p_i and p_j represent predicted values of y_i and y_j, respectively.

MSE reflects the degree of difference between the real and predicted values (Köksoy, 2006). Low MSE value indicates high accuracy. The formula is as follows (Marmolin, 1986):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2 \tag{4}$$

where P and Y contain the predicted and real values of n data points, respectively. P_i and Y_i represent the predicted and real values of the ith sample, respectively.

$r^2_{m(\text{test})}$ is an external validation parameter to evaluate the external prediction performance and decide acceptability of QSAR model. It determines the extent of deviation of the real values from the predicted values of test samples. A model can be determined acceptable, robust and not obtained by chance when the value of $r^2_{m(\text{test})}$ for the test set is greater than 0.5. The formula is as follows (Pratim Roy *et al.*, 2009):

$$r^2_{m(\text{test})} = r^2(1 - \sqrt{r^2 - r_0^2}) \tag{5}$$

where r^2 denotes squared correlation coefficient, and r_0^2 denotes squared correlation coefficient with intercept to zero.

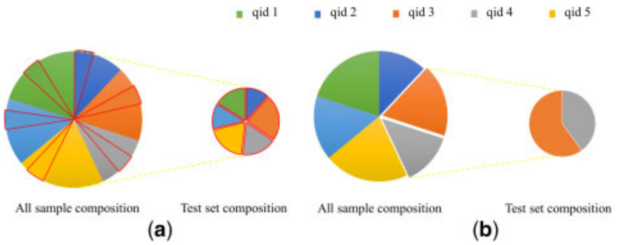

Fig. 3. The distribution of the query id and its corresponding samples under different categories

3 Experiments and results

3.1 Strategy setting of multi-scenario applications

In this study, multi-scenario applications are realized using different types of queries and different sample settings. According to the distribution of the query id and its corresponding samples in test and training set, these scenarios can be divided into two categories. (i) The scenarios of predicting drugs (or targets) that can bind to the known targets (or known drugs), including drug repurposing and protein new function prediction. (ii) The scenarios of predicting drugs (or targets) that can bind to the new targets (or new drugs), including new drug discovery and new target discovery. The division of training and test samples of category i is shown in Figure 3a, several samples under each query are extracted to form a test set, and the rest form a training set. The dataset division form of category ii is shown as Figure 3b, it extracts all samples corresponding to certain queries as the test set, instead of partially extracting the samples corresponding to each query. For example, in an experiment, the training set contains all samples corresponding to query q_1–q_n, and the test set consists of all samples corresponding to query q_{n+1}.

In this study, the drugs and targets that appear in the training set are called known drugs and known targets. Otherwise, called new drugs and new targets.

For drug repurposing, it aims to identify potential undiscovered targets for known drugs, therefore each drug can be set as a query. By observing the flattened indexes corresponding to the test samples and the training samples as well as the form of the matrix information on which the flattening depends (the rows and columns of the matrix represent drugs and proteins, respectively), it can be obtained that the purpose of many previous studies is to predict undiscovered proteins that can bind to known drugs. Therefore, we use the same sample partitioning settings as DeepDTA (Öztürk *et al.*, 2018) and GraphDTA (Nguyen *et al.*, 2021).

For protein new function prediction, each protein can be set as a query, we allocate samples according to the flattened indexes corresponding to the test samples and training samples in drug repurposing. However, please note that the flattened indexes are calculated based on the transpose of the matrix used in drug repurposing.

For the discovery of new drug and protein, drugs and proteins are set as queries, respectively. In this study, many experiments are conducted to determine the results of new drug and protein discovery. We extract five times from all queries (assuming *n*) and randomly extracted *n/5* kinds of queries each time. The extracted samples correspond to *n/5* types of queries form the test set, and the rest form the training set. Finally, five test sets and five training sets are formed. The union of these five test sets is the total samples, and the intersection is empty. The union of each test set and its corresponding training set is the total samples, and the intersection is empty.

3.2 Parameter setting of MART

The parameters and default parameter setting are listed in Table 3. 'tree' represents the number of trees, 'leaf' represents the number of leaves on each tree, 'shrinkage' is the learning rate, 'threshold candidates (tc)' is the number of candidate thresholds for tree splitting, and 'min leaf support (mls)' is the minimum number of samples each leaf must contain.

We set the final parameter value according to the performance of NerLTR-DTA in the scenario of protein new function prediction on the KIBA dataset. KIBA contains more samples and the number of queries corresponding to protein new function prediction is less, the parameters set under such conditions are more applicable. Moreover, in this study, each parameter is tuned by controlling variables and simple gradient descent idea. For example, when adjusting the learning rate, we first fix other parameters to their default values, and then set the learning rate value within a certain range to continuously fine-tune. Each ultimate result is the average performance score of the test set on five models trained with the same parameters on five different training sets.

The results of parameter tuning are shown in Figure 4, it can be observed that the CI value is higher than 88.6% and the MSE value is lower than 14.6% under each parameter setting, which indicate that the MART algorithm is suitable for DTA prediction. The parameter combination that have better performance and faster calculation speed is selected as the final parameter setting for subsequent experiments. In this study, the selected parameter combination is shown in *Adjusted* column of Table 3.

3.3 Performance of NerLTR-DTA on predicting drugs (or targets) that can bind to the known targets (or known drugs)

In terms of protein new function prediction (S1), it can be observed from Table 4 that NerLTR-DTA achieves high predictive performance on Davis and KIBA. For Davis, the CI and MSE are 0.936 and 0.155, respectively. For KIBA, the CI and MSE are 0.893 and 0.134, respectively. In terms of drug repurposing (S2), four existing models are compared with NerLTR-DTA. As shown in Table 4, NerLTR-DTA is superior to the state-of-the-art models, showing excellent predictive ability. For Davis dataset, the CI value is 0.928 and the MSE value is 0.171, which is 3.5% higher than the highest CI value and 5.8% lower than the lowest MSE to the existing models, respectively. For KIBA dataset, NerLTR-DTA performs as well as GraphDTA on CI metric, and obtain the lowest MSE. Moreover, regardless of whether it is in the S1 or S2, $r^2_{m(\text{test})}$ values are always greater than 0.5, which show that our models on these two datasets are quite robust and predictive, is not obtained by chance.

Table 3. Parameters of MART algorithm

Parameters	Initial	Adjusted
Tree	1000	500
Leaf	300	300
Shrinkage	0.1	0.03
Threshold candidates(tc)	256	256
Min leaf support(mls)	1	5

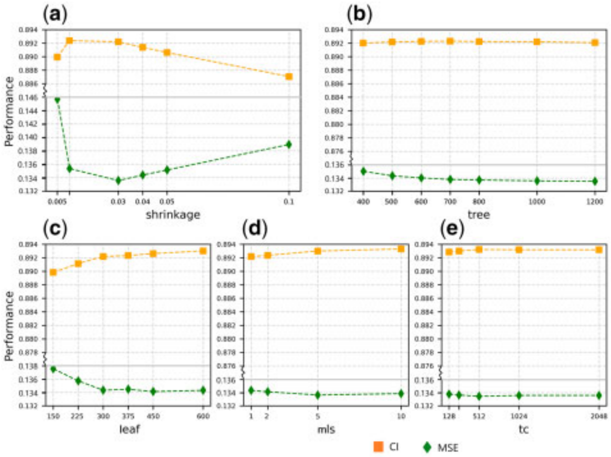

Fig. 4. The impact of each parameter in the MART algorithm

Table 4. Performance of NerLTR-DTA on S1 and S2

Method	Davis			KIBA		
	CI	MSE	$r^2_{m(\text{test})}$	CI	MSE	$r^2_{m(\text{test})}$
KronRLS (Pahikkala et al., 2015)	0.871	0.379	0.407	0.782	0.411	0.342
SimBoost (He et al., 2017)	0.872	0.282	0.644	0.836	0.222	0.629
DeepDTA (Öztürk et al., 2018)	0.878	0.261	0.630	0.863	0.194	0.673
GraphDTA (Nguyen et al., 2021)	0.893	0.229	—	0.891	0.139	—
NerLTR-DTA (S1)	0.936	0.155	0.792	0.893	0.134	0.800
NerLTR-DTA (S2)	0.928	0.171	0.766	0.891	0.135	0.794

Different models have their own preferences that produce different predicted affinity value for same drug–target pair. This stimulates us to assess the effectiveness of ensemble learning on our study, we integrate the five predicted affinity values for each sample and use the average of these five values as the final predicted affinity value. As shown in Table 5, it can be observed that our models are still reliable according to the value of $r^2_{m(\text{test})}$.

Compared with the average CI and MSE values of the test set on the five models trained by five different training sets, we observe that the CI value has increased and the MSE value has decreased. Specifically, the changes in MSE are more obvious, which demonstrate that ensemble learning can significantly narrow the difference between the predicted value and the real value. Figure 5 shows the predicted value against real value on Davis and KIBA, it can be observed that in the two scenarios for KIBA dataset, the fitting curves obtained based on linear regression basically coincide with the $y = x$ line. Similarly, for Davis dataset, the fitting curves are also close to the $y = x$ line.

3.4 Performance of NerLTR-DTA on predicting drugs (or targets) that can bind to the new targets (or new drugs)

NerLTR-DTA shows high predictability and acceptability on predicting drugs (or targets) that can bind to the new targets (or new drugs).

Table 5. The impact of ensemble learning on NerLTR-DTA performance

Method	Davis			KIBA		
	CI	MSE	$r^2_{m(\text{test})}$	CI	MSE	$r^2_{m(\text{test})}$
S1	0.940 (↑0.4%)	0.147 (↓0.8%)	0.790	0.896 (↑0.3%)	0.129 (↓0.5%)	0.799
S2	0.933 (↑0.5%)	0.162 (↓0.9%)	0.764	0.894 (↑0.3%)	0.131 (↓0.4%)	0.794

Note: The content in parentheses represents the increase ↑ and decrease ↓ of the numerical value under the ensemble results relative to the average CI and MSE scores of the test set trained on five different training sets.

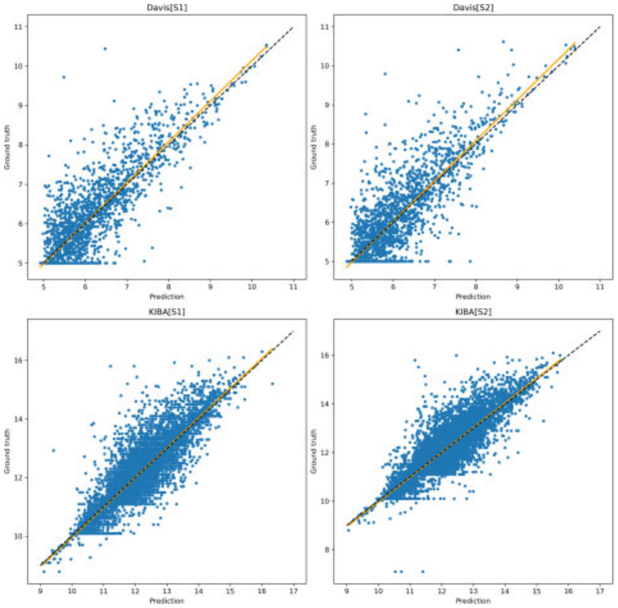

Fig. 5. Prediction against the ground truth on Davis and KIBA. The black dashed line is $y=x$, and the orange solid line is the linear regression fitting curve between the predicted value and the real value

Table 6. Performance of NerLTR-DTA on S3

Split	KIBA			Davis		
	CI	MSE	$r^2_{m(\text{test})}$	CI	MSE	$r^2_{m(\text{test})}$
1	0.877	0.138	0.728	0.913	0.191	0.690
2	0.863	0.161	0.700	0.930	0.171	0.728
3	0.879	0.128	0.769	0.929	0.166	0.794
4	0.873	0.318	0.687	0.901	0.214	0.638
5	0.914	0.232	0.866	0.919	0.229	0.719

Table 7. Performance of NerLTR-DTA on S4

Split	KIBA			Davis		
	CI	MSE	$r^2_{m(\text{test})}$	CI	MSE	$r^2_{m(\text{test})}$
1	0.873	0.170	0.760	0.922	0.204	0.715
2	0.872	0.193	0.783	0.912	0.223	0.751
3	0.859	0.185	0.715	0.913	0.238	0.695
4	0.884	0.151	0.757	0.915	0.194	0.600
5	0.849	0.172	0.663	0.897	0.235	0.652

In terms of new drug discovery (S3), as shown in Table 6. For KIBA dataset, CI values of five experiments are all above 0.860, and the highest CI value achieves 0.914. Moreover, three experiments have performed lower MSE values. For Davis dataset, NerLTR-DTA performs high accuracy, all CI values are greater than 0.900.

In terms of new protein discovery (S4), as shown in Table 7. For KIBA dataset, the lowest CI value of five experiments is 0.849 and all experiments achieve lower MSE values, all of which are lower than 0.200. For Davis dataset, the lowest CI value is as high as 0.897, and the highest MSE value is only 0.238.

It can be observed from Tables 6 and 7 that performance of NerLTR-DTA on predicting drugs (or targets) that can bind to the new targets (or new drugs) is not as stable as performance of NerLTR-DTA on predicting drugs (or targets) that can bind to the known targets (or known drugs), this result is not surprising. Because the drug–target associations of new drugs (targets) are unknown and the data distribution between training and test set may have changed drastically. However, the $r^2_{m(\text{test})}$ values that obtained based on such diverse datasets are all greater than 0.5, which fully demonstrates the robustness of our method.

3.5 Performance of NerLTR-DTA on ranking the proteins (or drugs) that can be bind to each target drug (or protein)

LTR framework can be used to clearly and conveniently obtain the ranking list of related proteins (or drugs) that can bind to each query drug (or query protein). As shown in Figures 6 and 7, we summarize the distribution of CI values for all queries. It is worth noting that two types of special queries are not included. (i) The query which the affinity values of all drug–target pairs are equal. (ii) The query which only corresponds to one drug–target pair.

For Davis, 406 query proteins that meet the requirements are included in S1. According to Figure 6a and b, the CI values of most query points are densely distributed between 0.9 and 1. Four models can provide accurate ranking lists of related drugs for about 70% of queries, even though the worst-performing model can provide accurate ranking for 68% of queries. There are 66 query drugs that meet the requirements in S2. According to Figure 6a and c, our method has CI values between 0.8 and 1 on at least 90% of the queries.

For KIBA, there are 221 query proteins that meet the requirements in S1, and Figure 7a and b clearly show that the CI values of most query points are between 0.8 and 1. In S2, there are 1554 query drugs that meet the requirements. As shown in Figure 7a and c, compared with S1, a higher percentage of query points in S2 do not achieve good performances, but the CI values of most query points still reach 0.8. By observing the number of samples corresponding to query points with different CI values obtained in S2, most query points with lower CI values correspond to a relatively small number of samples. Therefore, the reason for poor performance may be that the number of samples corresponding to that query point is relatively small, and the model fails to learn the characteristics well. It can be concluded that as time increases, the effect of our model may be better with more samples.

3.6 Predictive power of various features

We explore the effectiveness of SAF and AAF in S1 and S2. It can be seen from Table 8 that the performance of AAF-based model is better than that of SAF-based model, meaning that AAF can more

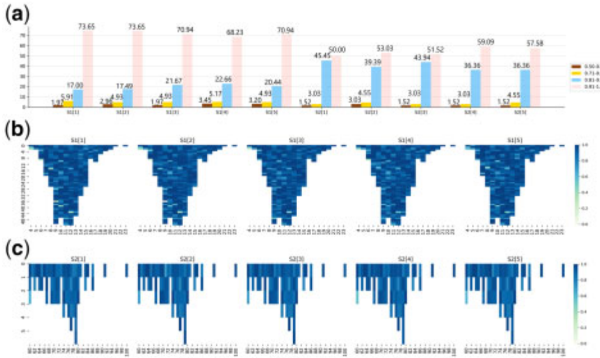

Fig. 6. CI value of each query object for Davis dataset. (**a**) Percentage of queries under different CI value ranges. (**b**) S1[1]–S1[5], respectively, represent the CI value distribution of the query proteins in the five models for scenario S1. (**c**) S2[1]–S2[5], respectively, represent the CI value distribution of the query drugs in the five models for scenario S2. In (b) and (c), the horizontal axis represents the number of samples corresponding to the query, and the vertical axis represents the number of queries that meet the conditions of the horizontal axis. Each rectangle in the figure represents the CI value of a query

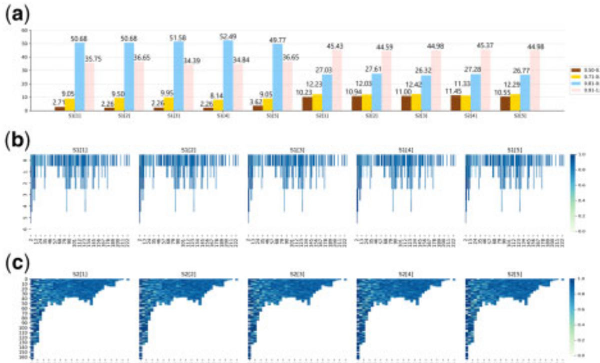

Fig. 7. CI value of each query object for KIBA dataset. (**a**) Percentage of queries under different CI value ranges. (**b**) S1[1]–S1[5], respectively, represent the CI value distribution of the query proteins in the five models for scenario S1. (**c**) S2[1]–S2[5], respectively, represent the CI value distribution of the query drugs in the five models for scenario S2. In (b) and (c), the horizontal axis represents the number of samples corresponding to the query, and the vertical axis represents the number of queries that meet the conditions of the horizontal axis. Each rectangle in the figure represents the CI value of a query

effectively characterize the properties of drugs and targets. As for the two types of AAF, $AAF_{similarity}$ and $AAF_{sharing}$, they perform differently on different datasets in terms of CI and MSE. But in terms of $r^2_{m(\text{test})}$, $AAF_{similarity}$ is obviously more effective and robust than $AAF_{sharing}$. Furthermore, SAF and AAF features are complementary, the performance of models built using these features alone is not as well as that based on all features.

4 Discussion

In this study, we propose a new method for DTA prediction. Our method performs high predictability in the scenarios of S1, S2, S3 and S4. Our experimental results clearly demonstrate that our method outperforms some state-of-the-art competing methods in S1. Moreover, the CI and MSE values obtained in S2 are only slightly varying to those obtained in S1, indicating that the performance of our method in predicting drugs (or targets) that can bind to the known targets (or known drugs) is hardly affected by sample differences, it has strong stability and strong generalization ability. As for S3 and S4, the performance of the models is not stable, however, even the worst-performing model has acceptable performance. Our method is designed based on the LTR algorithm. Obviously, LTR is a powerful tool to DTA prediction and should be widely used to any research that can be regarded as a search ranking task.

Table 8. Predictive power of various features

Dataset	Feature	S1			S2		
		CI	MSE	$r^2_{m(\text{test})}$	CI	MSE	$r^2_{m(\text{test})}$
Davis	SAF	0.866	0.281	0.652	0.852	0.323	0.584
	$AAF_{similarity}$	0.912	0.184	0.763	0.904	0.202	0.734
	$AAF_{sharing}$	0.924	0.209	0.734	0.916	0.232	0.706
	AAF	0.935	0.160	0.783	0.928	0.174	0.760
KIBA	SAF	0.844	0.232	0.664	0.842	0.236	0.648
	$AAF_{similarity}$	0.879	0.155	0.768	0.878	0.153	0.766
	$AAF_{sharing}$	0.854	0.210	0.685	0.850	0.213	0.679
	AAF	0.886	0.145	0.780	0.884	0.147	0.774

We observe that ensemble learning can improve the performance. Specifically, it can significantly narrow the difference between the predicted value and the real value. Therefore, we will make efforts to learn and apply exhaustive ensemble learning.

Following the two conjectures of ‘similar drugs tend to target similar proteins’ and ‘drugs with a larger number of shared targets may share a new target at the same time,’ we use neighboring drug (or protein) information that has high similarity and high sharing with the target drug (or protein) to extract AAF features. Compared with SAF features, AAF shows stronger predictive power, meaning that it is feasible to extract features based on neighboring information. Therefore, diverse neighboring information, such as the proximity relationship between drugs and diseases, proteins and diseases, drugs and drugs, drugs and side effects, can be integrated to future research.

5 Conclusion

Compared with other methods, NerLTR-DTA has the following advantages. (i) Currently, data on drug–target pairs with known interaction relationships are not sufficient, and the small amount of data makes model unable to learn the essential characteristics. Relying on neighboring information can compensate for the lack of existing data to a certain extent. Compared to the way in which drugs and targets are regarded as independent individuals, this kind of feature based on neighbors can more effectively describe drug–target pair information, which is more conducive to the construction of high-performance models. (ii) Multi-scenario application of the model is realized by setting different types of queries and dividing different types of training sets and test sets, and the models achieve excellent performance in all these scenarios. Particularly, the characteristics of the ranking framework are suitable for the discovery of new drugs and targets, which are ignored in most existing studies. (iii) NerLTR-DTA, which is implemented based on the MART algorithm, is a ranking framework with regression properties. It can not only accurately predict specific affinity values, but also provide priority ranking of proteins (or drugs) related to drugs (or proteins). Overall, we believe that NerLTR-DTA can contribute to a certain extent in drug development and drug repurposing.

Funding

This work was supported by the National Natural Science Foundation of China [62131004, 61922020], the Sichuan Provincial Science Fund for Distinguished Young Scholars [2021JDJQ0025], the Special Science Foundation of Quzhou [2021D004], New Energy and Industrial Technology Development Organization 265 (NEDO) and JST COI-NEXT, JST SPRING [JPMJSP2124].

Conflict of Interest: none declared.

References

Bahuguna,A. and Rawat,D.S. (2020) An overview of new antitubercular drugs, drug candidates, and their targets. *Med. Res. Rev.*, **40**, 263–292.

Benson,M.L. *et al.* (2007) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Res.*, **36**, D674–D678.

Bleakley,K. and Yamanishi,Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, **25**, 2397–2403.

Burges,C. *et al.* (2005) Learning to rank using gradient descent. In: *Proceedings of the 22nd International Conference on Machine Learning*,at the *University of Bonn,Bonn,* pp. 89–96.

Cao,Z. *et al.* (2007) Learning to rank: from pairwise approach to listwise approach. In: *Proceedings of the 24th International Conference on Machine Learning*,at the *Oregon State University, Corvalis,* pp. 129–136.

Cao,D.S. *et al.* (2014) Computational prediction of drug–target interactions using chemical, biological, and network features. *Mol. Inf.*, **33**, 669–681.

Chen,X. *et al.* (2016) Drug–target interaction prediction: databases, web servers and computational models. *Brief Bioinform.*, **17**, 696–712.

Chen,J. *et al.* (2017) ProtDec-LTR2. 0: an improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank. *Bioinformatics*, **33**, 3473–3476.

Corsello,S.M. *et al.* (2017) The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.*, **23**, 405–408.

Davis,M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.

Ding,Y. *et al.* (2017) Identification of drug–target interactions via multiple information integration. *Inf. Sci.*, **418-419**, 546–560.

Ezzat,A. *et al.* (2019) Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform.*, **20**, 1337–1357.

Gönen,M. (2012) Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, **28**, 2304–2310.

Gönen,M. and Heller,G. (2005) Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**, 965–970.

He,T. *et al.* (2017) SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminf.*, **9**, 1–14.

Jin,X. *et al.* (2021) SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection. *Bioinformatics*, **37**, 913–920.

Kim,S. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.

Köksoy,O. (2006) Multiresponse robust design: mean square error (MSE) criterion. *Appl. Math. Comput.*, **175**, 1716–1729.

Liu,H. *et al.* (2015) Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, **31**, i221–i229.

Liu,Y. *et al.* (2016) Neighborhood regularized logistic matrix factorization for drug–target interaction prediction. *PLoS Comput. Biol.*, **12**, e1004760.

Liu,L. *et al.* (2020) HPOLabeler: improving prediction of human protein–phenotype associations by learning to rank. *Bioinformatics*, **36**, 4180–4188.

Liu,R. *et al.* (2021) A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat. Mach. Intell.*, **3**, 68–75.

Marmolin,H. (1986) Subjective MSE measures. *IEEE Trans. Syst. Man Cybern.*, **16**, 486–489.

Mousavian,Z. and Masoudi-Nejad,A. (2014) Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin. Drug Metab. Toxicol.*, **10**, 1273–1287.

Nguyen,T. *et al.* (2021) GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, **37**, 1140–1147.

O'Meara,M.J. *et al.* (2020) A SARS-CoV-2-human protein–protein interaction map reveals drug targets and potential drug-repurposing. *BioRxiv*.

Öztürk,H. *et al.* (2016) A comparative study of SMILES-based compound similarity functions for drug–target interaction prediction. *BMC Bioinformatics*, **17**, 128–111.

Öztürk,H. *et al.* (2018) DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, **34**, i821–i829.

Öztürk,H. *et al.* (2019) WideDTA: prediction of drug–target binding affinity. arXiv preprint arXiv:1902.04166.

Pahikkala,T. *et al.* (2015) Toward more realistic drug–target interaction predictions. *Brief Bioinform.*, **16**, 325–337.

Pratim Roy,P. *et al.* (2009) On two novel parameters for validation of predictive QSAR models. *Molecules*, **14**, 1660–1701.

Pushpakom,S. *et al.* (2019) Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, **18**, 41–58.

Ragoza,M. *et al.* (2017) Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, **57**, 942–957.

Rayhan,F. *et al.* (2017) iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci. Rep.*, **7**, 1–18.

Steck,H. *et al.* (2008) On ranking in survival analysis: bounds on the concordance index. In: *Advances in Neural Information Processing Systems*, at the *Hyatt Regency Vancouver, Vancouver*, pp. 1209–1216.

Tang,J. *et al.* (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.*, **54**, 735–743.

Theodoris,C.V. *et al.* (2021) Network-based screen in iPSC-derived cells reveals therapeutic candidate for heart valve disease. *Science*, **371**, eabd0724.

Trotman,A. (2005) Learning to rank. *Inf. Retrieval*, **8**, 359–381.

Tsubaki,M. *et al.* (2019) Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, **35**, 309–318.

Wang,L. *et al.* (2018) Rfdt: a rotation forest-based predictor for predicting drug–target interactions using drug structure and protein sequence information. *Curr. Protein Peptide Sci.*, **19**, 445–454.

Wei,H. *et al.* (2021) iCircDA-LTR: identification of circRNA–disease associations based on learning to rank. *Bioinformatics*, **37**, 3302–3310.

Xia,F. *et al.* (2008) Listwise approach to learning to rank: theory and algorithm. In: *Proceedings of the 25th International Conference on Machine Learning*,at the *University of Helsinki, Helsinki,* pp. 1192–1199.

Xia,Z. *et al.* (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In, BMC systems biology. *BioMed Central*, **4**, 1–16.

Yamanishi,Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yang,Z. *et al.* (2021) ML-DTI: mutual learning mechanism for interpretable drug–target interaction prediction. *J. Phys. Chem. Lett.*, **12**, 4247–4261.

Yuan,Q. *et al.* (2016) DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, **32**, i18–i27.

Zheng,L. *et al.* (2019) Onionnet: a multiple-layer intermolecular-contact--based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega*, **4**, 15956–15965.

Zhou,D. *et al.* (2021) MultiDTI: drug–target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. *Bioinformatics*, **37**, 4485–4492.