AttentionDTA: prediction of drug–target binding affinity using attention model

Qichang Zhao
Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education
Xiangtan University
Xiangtan, Hunan, China
zqc137241304@gmail.com

Fen Xiao *
Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education
Xiangtan University
Xiangtan, Hunan, China
xiaof@xtu.edu.cn
* Corresponding author

Mengyun Yang
School of Computer Science and Engineering
Central South University
Changsha, Hunan, China
mengyunyang@csu.edu.cn

Yaohang Li
Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
yaohang@cs.odu.edu

Jianxin Wang
School of Computer Science and Engineering
Central South University
Changsha, Hunan, China
jxwang@mail.csu.edu.cn

***Abstract*—In bioinformatics, machine learning-based prediction of drug-target interaction (DTI) plays an important role in virtual screening of drug discovery. DTI prediction, which have been treated as a binary classification problem, depends on the concentration of two molecules, the interaction between two molecules, and other factors. The degree of affinity between a drug molecule (such as a drug compound) and a target molecule (such as a receptor or protein kinase) reflects how tightly the drug binds to a particular target and is quantified by the measurement which can reflect more detailed and specific information than binary relationship. In this study, we proposed an end-to-end model, named AttentionDTA, based on deep learning, which associates attention mechanism to predict the binding affinity of DTI. The novelty in this work is to use attentional mechanisms to consider which subsequences in a protein are more important for a drug and which subsequences in a drug are more important for a protein when predicting its affinity. So that the representational ability of the model is stronger. The model uses one-dimensional Convolution Neural Networks (1D-CNNs) to extract the abstract information of drug and protein, and makes the drug and protein representations mutually adapt through the attention mechanisms. We evaluate our model on two established drug–target affinity benchmark datasets, Davis and KIBA. The model outperforms DeepDTA, a state-of-the-art deep learning method for drug–target binding affinity prediction, with better Mean Squared Error (MSE), Concordance Index (CI), r_m^2 , and Area Under Precision Recall Curve (AUPR).Our results show that the attention-based model can effectively extract effective representations by calculating the weight of the representation between the drug and the protein. Finally, we visualize the attention weight. It proves our model can obtain the information of binding sites.**

Keywords—drug-target interaction, attention mechanism, deep learning

I. INTRODUCTION

The newly discovered drug–target interactions (DTIs) play an important role in drug discovery and drug repositioning. But wet lab experiments are inefficient, expensive and time-consuming [1,2]. This can be accelerated by providing the most effective DTI in silico prediction of drug-target interaction. Therefore, understanding the interaction between compounds and target proteins through computational methods is an important task in drug research. Recent studies [4] have shown that machine learning-based approaches can learn from limited interaction data, supplemented by information on the similarities between compounds and proteins, making it possible to predict interactions between compounds and proteins on a large scale.

The methods about predicting drug-protein relationships could be classified into two categories. One is binary classification method and the other predicts drug-target affinity (DTA) by regression methods. In binary classification-based DTI prediction studies, many researches have begun to use deep learning technologies to deal with DTI problem including restricted Boltzmann machines [15], deep neural networks [16,17,18], stacked auto-encoders [19,20], and deep belief networks [21]. In [22], the authors used a model called Wide-and-Deep to predict interactions by combining various drug and protein representations. Zheng X et al. proposed a deep-learning-based hybrid model, named DTI-RCNN, that integrated a long short-term memory (LSTM) networks with a convolutional neural network (CNN) to further improve DTIs prediction accuracy [23].

As elaborated in [5], there are two defects in treating DTI as a binary classification problem: (1) true-negative interactions and unknown values are not differentiated, and (2) binary relationships are too simple while it is more informative to use a continuous value that quantifies how strongly a drug binds to a target. On the other hand, binding affinity reflects the strength of the interaction between drug-target (DT) pairs and is usually expressed in terms of such measures as dissociation constant (K_d), inhibition constant (K_i), or the half maximal inhibitory concentration ($IC50$), which is determined by the concentration of the drug and protein[3]. The advantages of treating the prediction problem of drug protein relationship as a regression problem are to avoid the influence of negative sample selection on the model and can provide more practical and useful information [5]. The prediction in KronRLS [5] is based on the similarity score of each drug-target pair, which is defined by the

978-1-7281-1867-3/19/$31.00 ©2019 IEEE
978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Kronecker product of drug-drug similarity matrix and target-target similarity matrix. The similarities between proteins is based on Smith-Waterman (S-W) alignment algorithm [29]. The PubChem structure clustering tool (http://pubchem.ncbi.nlm.nih.gov) were utilized to calculate the similarities between compounds. SimBoost [6], a gradient boosting machine-based method, is employed for the prediction of the binding affinity. which depends on feature engineering of compounds and proteins utilizing information such as similarity and network-inferred statistics. Using the drug SMILES and the amino acid sequence of the protein as inputs, Öztürk et al. [24] used two one-dimensional convolutional blocks to get their hidden features separately. The final features of the two CNN blocks were concatenated and fed into three fully-connected layers to process drug protein affinity prediction. But it is difficult to analyze these deep learning models due to their black-box characteristic.

In this study, we propose a model to predict the binding affinities of drug-protein interactions with attention model using only sequences (1D representations) of proteins and drugs. After CNN extracted the abstract matrix representation of drugs and proteins, we use the attention module to calculate the scores between drugs' and proteins' representations at different positions, which allows us to consider which subsequences in a protein are more important for subsequences in the drug when predicting its affinity score and vice versa. We combine these representations to feed into a fully connected layer block, so called AttentionDTA. We use the Davis Kinase binding affinity dataset and the KIBA large-scale kinase inhibitors bioactivity data to evaluate the effectiveness of our method. The attention mechanism to represent proteins and drugs used in AttentionDTA performs significantly better than DeepDTA algorithm on these two datasets. With our proposed algorithm, we also obtain the lowest Mean Squared Error (MSE) value and highest Concordance Index (CI), r_m^2 index and the Area Under Precision Recall Curve (AUPR) score on two datasets. Our results suggest representations of drugs and proteins combined with corresponding attention information are more effective. Furthermore, we demonstrate the superiority of our model by visualizing the comparison between the true binding sites and the predicted binding sites by our model.

II. DATASETS

We evaluate our proposed model on two different datasets, the Kinase datasets Davis [13] and KIBA datasets [14], which have been popularly used as benchmarks for binding affinity prediction assessments. Table 1 shows some details of these two datasets.

TABLE I. SUMMARY OF THE DATASETS

Datasets	***Number of Proteins***	***Number of Drugs***	***Number of Interactions***
Davis(K_d)	442	68	30,056
KIBA	229	2,111	118,254

A. KIBA dataset

KIBA dataset comprises interactions of 442 proteins and 68 drugs. The KIBA dataset is originated from an approach called KIBA, in which kinase inhibitor bioactivities from different sources such as K_i, K_d and IC50 were combined. KIBA scores were constructed to optimize the consistency between K_i, K_d and IC50 by utilizing the statistical information they contained. The affinity value ranges from 0.0 to 17.2. For the drugs of KIBA, the maximum length of a SMILES is 590, while the average length is equal to 58. The maximum protein sequence length in KIBA is 4,128 and the average length is 728.

B. Davis dataset

There are binding affinities observed for all pairs of 68 drugs and 442 targets, measured by K_d value (kinase dissociation constant). The affinity value ranges from 5.0 to 10.8. For the compounds of the Davis dataset, the maximum length of a SMILES is 103, while the average length is equal to 64. The maximum protein sequence length in Davis is 2,549 and the average length is 788. We use the values transformed into logarithm space, pK_d, similar to [6], explained as follows:

$$pK_d = -\log 10\left(\frac{K_d}{1e^9}\right) \quad (1)$$

III. PROPOSED MODEL

In this study, we treat drug-protein affinity prediction as a regression problem aiming to predict the binding affinity scores. As mentioned earlier, we adopt a popular deep learning architecture, CNN, to extract useful information from the sequence. The goal of attention model is then to derive a drug representation that captures relevant information of protein and a protein representation that captures relevant information of drug. Because the high affinity sites of different drugs on protein molecules are located in different regions. Through the attention mechanism, we can obtain the attention scores between drug's and protein's subsequences. The vector representations of subsequences are strengthened or weakened according to the attention scores. Detailed model is described below. An overview of the proposed method, AttentionDTA, is shown in Fig. 1.

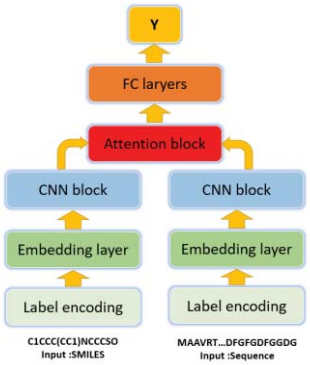

Fig. 1. AttentionDTA model with attention block to correlate drug and protein information.

A. Input representation

The inputs of the model are the amino acid sequences of the proteins and SMILES sequences of the drugs. We use integer/label encoding that uses integers for the categories to

represent the inputs. According to [24], the SMILES sequences are made up of 64 different characters. For protein sequences, they are 25 different categories. Here we construct two dictionaries for drug sequences and protein sequences, such as {'C': 1, 'N': 2, 'O': 3, '=': 4, '(': 5, ')': 6 etc.} for drug. The label encoding for the example SMILES, 'CC(C(=O)O)O', is represented below.

$$[CC(C(=O)O)O] = [11515436363]$$

Protein sequences are encoded in a similar way using label encodings by a different dictionary, such as {'A': 1, 'C': 2, 'D': 3, 'E': 4, 'F': 5, 'G': 6 etc.}. The label encoding for the protein sequence, 'AACGFED', is represented below.

$$[AACGFED] = [1126543]$$

The drugs and protein sequences present sequences of variable length. Hence, in order to create an effective representation form, similar to [24], we use a fixed maximum 100 characters length for SMILES and 1,200 for protein sequences for the two datasets. The sequences that are longer than the maximum length are truncated, whereas shorter sequences are 0-padded. Such a length setting can not only reduce computational complexity, but also retain sufficient valid information.

B. Filter function in CNN

Let $x_i \in \mathbb{R}^k$ be the k-dimensional vector corresponding to the i-th character in the sequence. A sequence of length n (padded where necessary) can be expressed as

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \tag{2}$$

where $\oplus$ is the concatenation operator. In general, let $x_{i:i+j}$ be a subsequence which refer to the concatenation of characters x_i, x_{i+1},…, x_{i+j}. A convolution operation involves a filter $w \in \mathbb{R}^{hk}$, which uses the h characters in the window to generate a new feature. For example, the i-th feature c_i is generated from the subsequence $x_{i:i+h-1}$ by

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{3}$$

Here $b \in \mathbb{R}^1$ is a bias term and $f()$ is a non-linear function such as the ReLU. This operation is applied to each possible window of characters in the sequence $\{x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}\}$, from left to right, to generate a new feature representation.

$$C = [c_1, c_2, \ldots, c_{n-h+1}] \tag{3}$$

C. Attention model

The CNN blocks generate new drug representation $D \in \mathbb{R}^{N \times F}$, where N is the length of the drug sequence and F is the depth and new protein representation $P \in \mathbb{R}^{M \times F}$, where M is the length of the drug sequence and F is the depth.

In this section, we design two calculation methods of attention, namely Attention1 and Attention2. We calculate the Attention1 by the following formula:

$$D_d = relu(DW + b) \tag{4}$$

$$P_p = relu(PW + b) \tag{5}$$

$$\alpha = \tanh(D_d P_p^T) \tag{6}$$

where W is the weight matrix, b is the bias, and T is the transpose operator. The weight value $\alpha \in \mathbb{R}^{N \times M}$, which is called attention scores, can be assumed to represent the interaction strength between the subsequences of drug and the subsequences of protein.

Next, the drug attention vector $\alpha_d \in \mathbb{R}^N$, which can be interpreted as attention scores of the expression of each position in the sequence of the drug to the whole protein, is calculated by applying row-wise sum operation to α. While protein attention vector $\alpha_p \in \mathbb{R}^M$ is calculated by applying column-wise sum operation to α. Each bit in the vector represents the extent to which the corresponding subsequence of protein is related to the drug representation.

$$\alpha_d = \sum_{j=1}^{M} \alpha_{i,j} \tag{7}$$

$$\alpha_p = \sum_{i=1}^{N} \alpha_{i,j} \tag{8}$$

Attention2 is designed as follows:

$$D_d = relu(DW_d + b) \tag{9}$$

$$P_p = relu(PW_p + b) \tag{10}$$

$$\alpha = \tanh(D_d P_p^T) \tag{11}$$

$$\alpha_d = \sum_{j=1}^{M} \alpha_{i,j} \tag{12}$$

$$\alpha_p = \sum_{i=1}^{N} \alpha_{i,j} \tag{13}$$

The difference between Attention1 and Attention2 is that different weights, W_d and W_p, are used to deal with drugs and proteins, respectively.

We repeat both α_d and α_p for F times, so that $\alpha_d \in \mathbb{R}^{N \times F}$ and $\alpha_p \in \mathbb{R}^{M \times F}$ becomes matrices. Then we update drug's and protein's representation:

$$D_\alpha = mutiply(\alpha_d, D) \tag{14}$$

$$P_\alpha = mutiply(\alpha_p, P) \tag{15}$$

where the mutiply () denotes element-wise product.

We then apply a max-pooling operation [28] over D_α and P_α to take the maximum value. The idea is to capture the most important feature—one with the highest value—for each feature map.

$$r_\alpha = maxpooling(D_d) \tag{16}$$

$$r_\alpha = maxpooling(P_p) \tag{17}$$

where $r_d \in \mathbb{R}^F$ and, $r_p \in \mathbb{R}^F$.

D. Output and training

The outputs from two max-pooling layers are concatenated and fed into multi-layer perceptron. Similar to [24], this multilayer perceptron consists of 5 layers of neural networks. The input layer receives output from max-pooling layers. We used 1024 nodes for the next two fully connected (FC) layers, each followed by a dropout layer of rate 0.5. Dropout [7] is a method to prevent the deep structure of artificial neural network from overfitting. In the learning process, it reduces the dependence among nodes, thereby regularization of the neural network and reduces its structural risk by randomly returning the weights or outputs of hidden layers to zero. The third FC layer consisted of 512 nodes and was followed by the output layer without Dropout.

The activation function that we used in FC layers is Leaky Rectified Linear Unit (Leaky ReLU)[8]. The aim for our model is to minimize the difference between the label and the prediction score during training. Since we work on a regression task, we use mean squared error (MSE) as the loss function, in which *P* is the prediction value of model, and *Y* corresponds to the actual label. n indicates the number of samples.

$$loss = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2 \tag{18}$$

The training details of these neural networks are described as follows. The learning is completed with 350 epochs and the batch-size are 64 for Davis and 256 for KIBA because KIBA is four times larger than Davis. Adam [9] was used as the optimization algorithm to train the networks with the default learning rate of 0.0001. We use embedding layer to represent characters with 128-dimensional dense vectors. The two CNN blocks consist of three 1D-convolutional layers with increasing number of filters. The number of filters is [32, 64, 96] for both drug and protein. The window sizes are different, [4, 6, 8] for drug and [4, 6, 12] for protein. The algorithm is coded based on Tensorflow. We use 5-fold cross-validation to assess the predictive ability of our model.

IV. Experiments and results

We propose a novel drug–protein affinity prediction method, which we named as AttentionDTA, based on the neural attention mechanism which only use sequence information of drugs and proteins. As mentioned in the previous section, we have defined two different methods of calculating attention. Correspondingly, our model has two different structures named Attention1DTA and Attention2DTA. We use MSE value, r_m^2 index and the AUPR score to measure the performance of the proposed model and compared it with the current state-of-art method, DeepDTA.

A. Baseline

DeepDTA [24] comprises two separate CNN blocks, each of which aims to learn representations from SMILES strings and protein sequences. For each CNN block, DeepDTA used three consecutive 1D-convolutional layers with increasing number of filters. The second layer had double and the third convolutional layer had triple the number of filters in the first one. The convolutional layers were then followed by the max-pooling layer. The final features of the max-pooling layers were concatenated and fed into three FC layers. DeepDTA used 1024 nodes in the first two FC layers, each followed by a dropout layer of rate 0.1. The third layer consisted of 512 nodes and was followed by the output layer.

B. Evaluation metrics

We evaluate the performance of the proposed model using the similarly evaluation metrics as [24]. Because we model DTA as regression problems, MSE and Concordance Index (CI) are the most commonly used evaluation indexes:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2 \tag{19}$$

$$CI = \frac{1}{Z}\sum_{y_i > y_j} h(p_i - p_j) \tag{20}$$

$$h(x) = \begin{cases} 1 & , \quad x > 0 \\ 0.5 & , \quad x = 0 \\ 0 & , \quad x < 0 \end{cases} \tag{21}$$

For MSE, p is the prediction value, y corresponds to the actual label and n is the number of samples. CI measures if the predicted binding affinity values of two random drug–target pairs were predicted in the same order as their true values were. In (20), sample i has a bigger label value than sample j.

In an effort to provide a better assessment of our model, we compare the performances of AttentionDTA and DeepDTA with two different metrics as well. r_m^2 index is used to evaluate the external predictive performance. The metric is described as follows:

$$r_m^2 = r^2 * (1 - \sqrt{r^2 - r_0^2}) \tag{22}$$

where r^2 and r_0^2 are the squared correlation coefficients with and without intercept, respectively. The value of r_m^2 (test) should be greater than 0.5 for an acceptable model. The details of the formulation are explained in [10,11]. The AUPR score is widely used by many studies in binary prediction. In order to calculate AUPR, we converted the two datasets into binary datasets by selecting appropriate binding affinity thresholds. For Davis dataset we used 7 as threshold similar to what is done in [6]. For KIBA dataset we used the suggested threshold value of 12.1 [6,12].

C. Results

1) Effectiveness

Table1 and Table2 show the average MSE, CI scores, r_m^2 values and AUPR for Davis and KIBA datasets. All of the data were averaged over 5-fold cross-validation results. We divided the dataset into 5 parts on average, and took one part as the test set and the remaining four parts as the training set. These five models were trained with the same

hyperparameters. The results were calculated on five completely different test sets.

In the KIBA dataset, our two attention-based models exceed baseline in all evaluation metrics, where Attention1DTA yields the best performance. Among them, MSE has the most obvious change, which decreases from 0.187 to 0.155. The change in CI is distinct. It also achieved a high value of 0.882, which is a 1.4% increase over the comparison method. Both r_m^2 and AUPR have improved significantly compared with DeepDTA, too. r_m^2 is 0.755, an 8.6% increase. Meanwhile, AUPR go up 3.8% to 0.829.

For Davis dataset, our two methods also achieve better results than the baseline. Attention1DTA get the best results. MSE decreases 14.3% to 0.215. The result of CI is 0.893, better than DeepDTA, which increase 1.4%. Meanwhile, a slight improvement in AUPR measure is gained by the proposed methods, at 0.776. In terms of r_m^2, Attention1DTA gets the best performance, at 0.677.

2) Case Study for Interpretability

The key advantage of our model over baseline is that the attention mechanism not only improves predictive power but also makes the model interpretable. To demonstrate this, we map the regions with high attention weight values calculated from the attention model onto a known 3D protein structure. Here, Fig. 2 shows the complex of imatinib and Tyrosine-protein kinase SYK (PDB ID: 1XBB), and Fig. 3 shows the complex of aspirin and Phospholipase A2 (PDB ID: 1TGM). The regions in protein, which are not binding sites but predicted to be binding sites, are highlighted in blue. The regions marked in red are true binding sites but not given high attention scores. The black regions indicate that these regions are true binding sites with high attention scores. In the case of 1XBB(Fig. 2), there are three binding sites getting high attention scores. Fig. 3 shows that all binding sites are marked by attention mechanisms.

V. Conclusion

In this paper, we have proposed an end-to-end deep-learning based model, AttentionDTA, which combines attention mechanism to find the weight relationships between drug subsequences and protein subsequences to obtain more effective drug and protein representations. We use two separate 3-layer convolution blocks to learn representations for drug and protein sequences, respectively. An attention block is then used to correlate the drug representation with the corresponding protein representation, improving the expression of the model. Finally, the affinity prediction task is completed by a fully connected network. Experimental results show that our model is better than the baseline model. Our improvement is significant when we have more data, since the KIBA dataset is four times larger than the Davis dataset. With more data, our model can better learn the representations of drugs and proteins and the relationship between these representations. In addition, the attention mechanism provides good visualization, allowing us to analyze the model well.

It is worth noting that the SMILES sequence of drug represents the 2D structure. The use of graph network models to process drugs has been successful in [25,26,27]. In future work, we will explore the relationship between graph network-based drug representation and protein representation.

Fig. 2. The complex of imatinib and Tyrosine-protein kinase SYK (PDB ID: 1XBB).

TABLE II. The average MSE, CI scores, r_m^2 values and AUPR of the test set trained on five different training sets for the KIBA dataset

KIBA	*MSE (std)*	*CI (std)*	*R_m^2 (std)*	*AUPR (std)*
DeepDTA	0.187(0.008)	0.869(0.001)	0.695(0.026)	0.798(0.003)
Attention1DTA	**0.155**(0.003)	**0.882**(0.004)	**0.755**(0.017)	**0.829**(0.005)
Attention2DTA	0.162(0.004)	0.880(0.002)	0.738(0.016)	0.811(0.005)

TABLE III. The average MSE, CI scores, r_m^2 values and AUPR of the test set trained on five different training sets for the Davis dataset

Davis	*MSE (std)*	*CI (std)*	*R_m^2 (std)*	*AUPR (std)*
DeepDTA	0.251(0.014)	0.881(0.005)	0.672(0.012)	0.766(0.023)
Attention1DTA	**0.216**(0.019)	**0.893**(0.005)	**0.677(0.024)**	**0.776**(0.024)
Attention2DTA	0.222(0.017)	0.886(0.005)	0.676(0.025)	0.775(0.019)

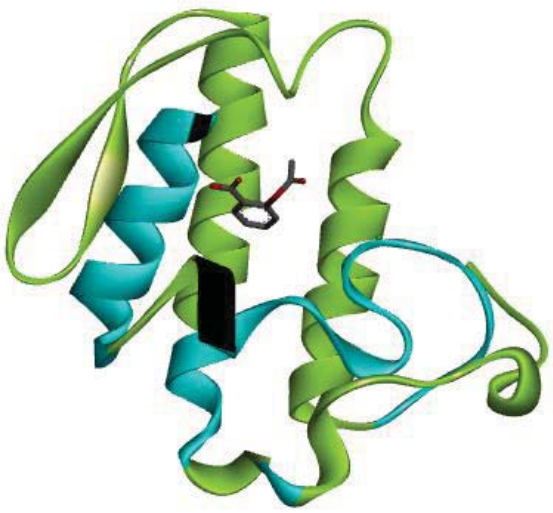

Fig. 3. The complex of aspirin and Phospholipase A2 (PDB ID: 1TGM).

REFERENCES

[1] Ezzat, A. , Wu, M. , Li, X. L. , & Kwoh, C. K. . (2018). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Briefings in Bioinformatics.

[2] Chen, X. , Yan, C. C. , Zhang, X. , Zhang, X. , Dai, F. , & Yin, J. , et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. Briefings in Bioinformatics, 17(4), 696.

[3] Cer, R. Z. , Mudunuri, U. , Stephens, R. , & Lebeda, F. J. . (2009). Ic50-to-ki: a web-based tool for converting ic50 to ki values for inhibitors of enzyme activity and ligand binding. Nucleic Acids Research, 37(Web Server), W441-W445.

[4] Ding, H. , Takigawa, I. , Mamitsuka, H. , & Zhu, S. . (2014). Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Briefings in Bioinformatics, 15(5), 734-747.

[5] Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2014). Toward more realistic drug–target interaction predictions. Briefings in bioinformatics, 16(2), 325-337.

[6] He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of cheminformatics, 9(1), 24.

[7] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

[9] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization: Computer Science.

[10] Pratim Roy, P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. Molecules, 14(5), 1660-1701.

[11] Roy, K. , Chakraborty, P. , Mitra, I. , Ojha, P. K. , Kar, S. , & Das, R. N. . (2013). Some case studies on application of "\r, r\r, m\r, 2\r, " metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data. Journal of Computational Chemistry, 34(12), 1071-1082.

[12] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., & Aittokallio, T. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling, 54(3), 735-743.

[13] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., ... & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nature biotechnology, 29(11), 1046.

[14] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., & Aittokallio, T. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling, 54(3), 735-743.

[15] Wang, Y., & Zeng, J. (2013). Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics, 29(13), i126-i134.

[16] Wang, C., Liu, J., Luo, F., Tan, Y., Deng, Z., & Hu, Q. N. (2014, November). Pairwise input neural network for target-ligand interaction prediction. In 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 67-70). IEEE.

[17] Tian, K., Shao, M., Wang, Y., Guan, J., & Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. Methods, 110, 64-72.

[18] Wan, F., & Zeng, J. (2016). Deep learning with feature embedding for compound-protein interaction prediction. bioRxiv, 086033. Unpublished.

[19] Chan, K. C., & You, Z. H. (2016, July). Large-scale prediction of drug-target interactions from deep representations. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 1236-1243). IEEE.

[20] Wang, L., You, Z. H., Chen, X., Xia, S. X., Liu, F., Yan, X., ... & Song, K. J. (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. Journal of Computational Biology, 25(3), 361-373.

[21] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug–target interaction prediction. Journal of proteome research, 16(4), 1401-1409.

[22] Du, Y., Wang, J., Wang, X., Chen, J., & Chang, H. (2018, March). Predicting Drug-target Interaction via Wide and Deep Learning. In Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology (pp. 128-132). ACM.

[23] Zheng, X., He, S., Song, X., Zhang, Z., & Bo, X. (2018, October). DTI-RCNN: New Efficient Hybrid Neural Network Model to Predict Drug–Target Interactions. In International Conference on Artificial Neural Networks (pp. 104-114). Springer, Cham.

[24] Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821-i829.

[25] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. Journal of computer-aided molecular design, 30(8), 595-608.

[26] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In Advances in neural information processing systems (pp. 2224-2232).

[27] Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS central science, 3(4), 283-293.

[28] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(Aug), 2493-2537.

[29] Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.