BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/10 pp1–10
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020005

FingerDTA: A Fingerprint-Embedding Framework for Drug-Target Binding Affinity Prediction

Xuekai Zhu, Juan Liu*, Jian Zhang, Zhihui Yang, Feng Yang, and Xiaolei Zhang

Abstract: Many efforts have been exerted toward screening potential drugs for targets, and conducting wet experiments remains a laborious and time-consuming approach. Artificial intelligence methods, such as Convolutional Neural Network (CNN), are widely used to facilitate new drug discovery. Owing to the structural limitations of CNN, features extracted from this method are local patterns that lack global information. However, global information extracted from the whole sequence and local patterns extracted from the special domain can influence the drug-target affinity. A fusion of global information and local patterns can construct neural network calculations closer to actual biological processes. This paper proposes a Fingerprint-embedding framework for Drug-Target binding Affinity prediction (FingerDTA), which uses CNN to extract local patterns and utilize fingerprints to characterize global information. These fingerprints are generated on the basis of the whole sequence of drugs or targets. Furthermore, FingerDTA achieves comparable performance on Davis and KIBA data sets. In the case study of screening potential drugs for the spike protein of the coronavirus disease 2019 (COVID-19), 7 of the top 10 drugs have been confirmed potential by literature. Ultimately, the docking experiment demonstrates that FingerDTA can find novel drug candidates for targets. All codes are available at http://lanproxy.biodwhu.cn:9099/mszjaas/FingerDTA.git.

Key words: drug-target binding affinity; fingerprint; new drug discovery

1 Introduction

In vivo drug discovery usually involves target-based screening, phenotypic screening, modification of natural substances, and biologic-based approaches[1]. These experiments are time-consuming, laborious, and costly. Virtual pre-screening of potential drug candidates can guide subsequent wet experiments. Compared to the traditional blind screening process, the pre-screening process uses high throughput techniques before the detailed experimental inspection. Virtual pre-screening can minimize costs and improve the success rate of drug discovery. Three kinds of strategies exist for virtual pre-screening as follows. (1) Strategies based on a high throughput assay system. Bao et al.[2] used green fluorescent protein to screen nuclear translocation inhibitors for cancer treatment. Wang et al.[3] developed a high throughput flow cytometry to support antibody discovery without tedious sample preparation. Keusgen[4] reviewed new drug discovery approaches of screening through biosensors. These methods allow for small, straightforward, and comparable wet experiments. (2) Strategies based on simulated molecular docking. Gupta et al.[5] explored inhibitors of Plasmodium falciparum's serine/threonine protein phosphatase through molecular docking. Rasool et al.[6] explored anti-viral molecules against dengue by simulating docking phytochemical against NS2B/NS3 proteases. Ghosh et al.[7] performed structural and physicochemical interpretation analysis on some small

• Xuekai Zhu, Juan Liu, Jian Zhang, Zhihui Yang, Feng Yang, and Xiaolei Zhang are with the School of Computer Science, Wuhan University, Wuhan 430072, China; E-mail: xuekaizhu@whu.edu.cn; liujuan@whu.edu.cn; moxi.zj@alibaba-inc.com; zhy@whu.edu.cn; feng.yang@whu.edu.cn; xlzhang@whu.edu.cn.

* To whom correspondence should be addressed.
Manuscript received: 2021-11-22; revised: 2021-12-21; accepted: 2022-02-16

molecule structures to the COVID-19 Mpro inhibition. These virtual pre-screening methods mimic drug-target binding processes and minimize the time spent on biological experiments. (3) Strategies based on drug-target affinity prediction models. Intuitively, a high affinity between the drug and the critical target means that the drug can be a candidate for treating the corresponding disease. Hakime et al.[8] conducted drug-target binding affinity prediction using Convolutional Neural Network (CNN). Rajpura and Ngom[9] used a support vector machine to predict drug and target interaction which can reposition known drugs to unknown targets. Zhou et al.[10] reviewed some drug-target interaction models and algorithms. Modern methods are mainly based on similarities of drug-drug, target-target, and drug-target. Although the first two strategies have been widely used to discover new candidate drugs, they warrant in-depth experimental design and verification, which is not suitable for gigantic scale screening. Given the accumulation of diverse omics data and the development of deep learning technology, the third strategy has great advantages in efficiency and cost. Recently, deep neural networks have been applied successfully to predict drug-target binding affinity rapidly[11]. Accordingly, we focus on developing a deep learning model for drug-target binding affinity prediction in this paper.

The CNN model is a widely used deep learning framework that can extract representative features (hereinafter called convolutional features) from training data. CNN models are more powerful than traditional models based on traditional machine learning methods[8]. However, elements in the convolutional are separated, indicating that convolutional features are local patterns. Resultantly, the processes of model calculation lack global information. In the processes of *in vivo* drug-target interaction, whole structure features and special domain features can influence affinity scores. In the actual biological compounds or proteins, some structures are distant in sequence but near in space, such as some transmembrane structures. Although this information is ignored by CNN models, it remains useful for predicting drug-target binding affinity. Furthermore, a fusion of global information and local patterns may help CNN models achieve better performance. This paper presents a Fingerprint-embedding framework for Drug-Target binding Affinity prediction (FingerDTA). Fingerprints of drugs and targets are calculated from the whole Simplified Molecular Input Line Entry System (SMILES) sequence and the whole amino acid sequence. Global information can be extracted by some Fully Connected (FC) layers from these fingerprints. We combine global information with baseline CNN models through an attention-like process that can guide the CNN model training.

To evaluate the effectiveness of the performance of FingerDTA, we conduct comparison experiments on Davis[12] and KIBA[13] datasets. Following previous works[8, 14], experiments illustrate that FingerDTA performs comparable results to state-of-the-art methods on Mean Squared Error (MSE), Concordance Index (CI), and r_m^2. We also apply FingerDTA to screen drugs for COVID-19. For the spike protein of COVID-19, 7 of the top 10 drugs have been confirmed potential by literature, which validates the effectiveness of FingerDTA. This finding confirms that FingerDTA can screen potential drugs for COVID-19 or other special targets.

2 Methodology

2.1 Overview

This section describes the details of FingerDTA. As shown in Fig. 1, the amino acid sequence of proteins and SMILES sequence of drugs are used as input. First, protein and drug are encoded into one-hot matrices. To characterize global information, the ECFP4 fingerprint[15] is used to represent drugs. Imitating ECFPs fingerprints algorithm, a target fingerprint algorithm is designed to represent proteins. These

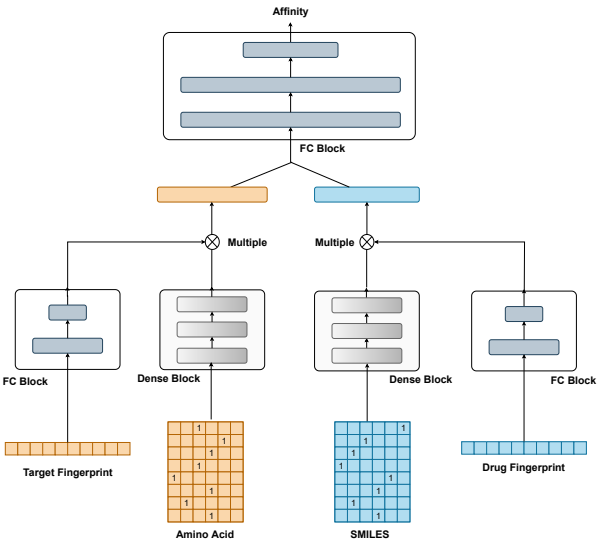

Fig. 1 Details of FingerDTA. Dense blocks extract local features from one-hot matrices, and FC layers compress fingerprints into global features. Then, FingerDTA utilizes these features to predict drug-target affinity scores.

fingerprints are calculated from the drug's SMILES[16] strings and target amino acid strings, which are based on enumerating the drugs' substructures and targets' amino acid arrangements. Second, dense blocks extract local features from one-hot matrices, and FC layers extract global features from fingerprints. To improve the performance of FingerDTA, a Dense Convolutional Block (DCB) is used to construct the CNN framework instead of a Normal Convolutional Layer (NCL). Third, FingerDTA fuses global and local features to predict drug-target affinity. In addition, fingerprints are calculated before training the affinity-predicting model. The data preprocessing is described in the next section. In summary, FingerDTA contains four main parts: drug fingerprint, target fingerprint, dense blocks, and fingerprint-embedding.

2.2 Drug fingerprint

Drug fingerprint is extracted as ECFPs[15] by RDKit[17]. The main objective is to list all environments formed by neighbor atoms and bonds. These environments can be interpreted as a drug's substructures. After some iterating epochs of calculation, substructures from the local scale to the global scale are all enumerated as patterns. Each pattern is converted to a one-hot vector which is defined as the fingerprint. The total process can be summarized into three main stages:

(1) Initialization: All atoms are given an integer tag except for hydrogen atoms within a drug according to the Daylight atom invariants rule[18].

(2) Iteration: All atoms are tagged and transformed into an array as $[n, tag, band_1, tag_1, band_2, tag_2, \dots]$, in which n means iterating epoch, tag means the integer tag of this atom, $band_i$ means bond type between neighbor atom, and tag_i means the tag of its neighbor atom. This array is hashed by iteratively applying the hash function for each item. Then atoms in the same environment (or same structure) are given the same tag by some rules during one iteration[15].

(3) Conversion: All unique tags produced are collected during each iteration. Each of them represents a unique substructure pattern. They are hashed into a one-hot 1024-dimensions vector, which is the fingerprint of this drug.

2.3 Target fingerprint

A target sequence can be divided into many slices. Each slice represents an amino acid arrangement. The base conversion method is used to obtain this characteristic number, as shown in Fig. 2. This method conveniently

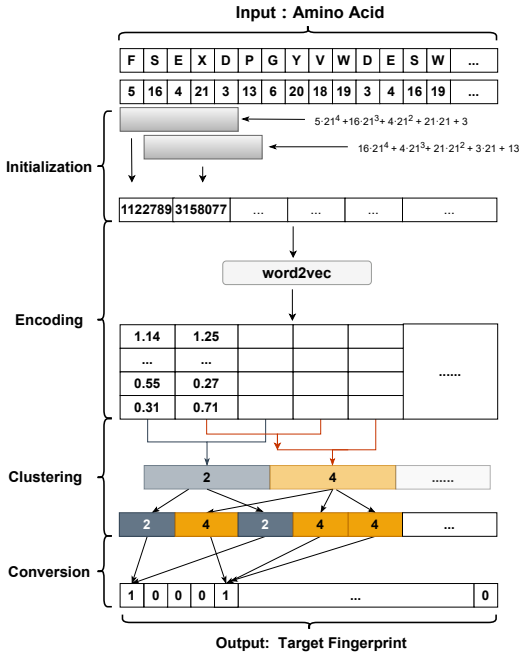

Fig. 2 Procedures for target fingerprint generation. The source amino acid sequence is encoded into the target fingerprint.

and intuitively converts a sequence of numbers into a feature number. Different sets of arrangements may have different kinds of long-range interaction. Given the considerable number of target slices, the target fingerprint is generated through a clustering-based process. This process concentrates many amino acid slices into a fixed number of critical patterns. Details are described in four main stages:

(1) Initialization: Each amino acid is tagged to a simple number from 1 to 20. Rare existing ones are tagged to the same integer of 21. Slices are sequentially generated for each target string by encoding every five numbers to a unique number (see Fig. 2). The step length of sliding is one. Thus, a tag array is generated for each target.

(2) Encoding: A word2vec model is trained[19] with the above-mentioned tag arrays through a skip-gram algorithm with a window size of five, and slices whose appearance count is less than three are ignored. This model is used to encode each slice of the target to a 64-dimension vector. Every two slices located in a similar environment will obtain less cosine distance between their vectors.

(3) Clustering: Slices are encoded into 1024 types by cosine distance between every two vectors of slices with a bottom-up hierarchical clustering method, which can be found in the Scikit-Learn python package[20].

(4) All appearing slice types in a target are collected

and mapped to a one-hot 1024-dimension vector, which is the fingerprint of this target.

2.4 Dense convolutional blocks

To improve the predicting performance of FingerDTA, the convolutional framework is optimized. Three DCBs are used to construct the convolutional framework of the CNN model. Each DCB contains four 1D convolutional layers. In this block, each result from a convolutional layer has the same number of channels which will be concatenated with all outputs of the former layers in the block through a channel dimension. All outputs of these four layers are concatenated in the channel dimension and activated by a Rectified Linear Unit (ReLu)-activating layer to generate the block's output. Kernel sizes for these four layers in the dense block are assigned to 1, 3, 5, and 7, respectively, and the padding sizes are respectively 0, 1, 2, and 3. The structure of the three DCBs is shown in Fig. 3. The output channels of these three DCBs are 128, 256, and 96, respectively.

2.5 Fingerprint embedding

Fingerprints are first fed into two FC layers to extract global features. As shown in Fig. 1, these two FC layers have 512 and 96 neutral nodes, and a dropout layer is between them. The output features contain global information of the whole length sequence. They are multiplied into each channel of the convolutional output to guide the CNN model training in the global background. This simple process introduces the global information into convolutional features.

Let the convolutional features be

$$x_{ci}^d, x_{cj}^t \in \mathbb{R}^{c\times l}, \quad i \in [1, 100], \quad j \in [1, 1200].$$

In the following contents, c means channel dimension, and i, j are the indexes in the sequence dimension (l). d and t represent drug and target.

Similarly, let the output features from two FC layers be $g_c^d, g_c^t \in \mathbb{R}^c$. The output of the attention-like process y_{ci}^d and y_{cj}^t is calculated as

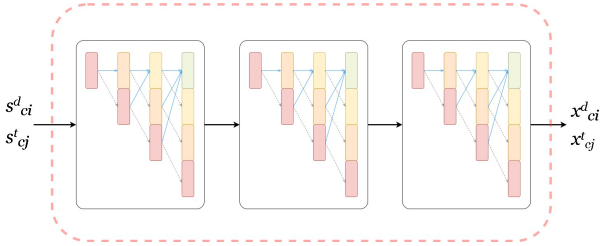

Fig. 3 Details and data calculation of the three DCBs. s_{ci}^d and s_{cj}^t are source one-hot matrics for drug and target, respectively.

$$y_{ci}^d = x_{ci}^d \times g_c^d \tag{1}$$

$$y_{cj}^t = x_{cj}^t \times g_c^t \tag{2}$$

The final feature in the global background y^d and y^t are calculated as

$$y_c^d = \max_{1\leqslant i\leqslant 100}(y_{ci}^d) \tag{3}$$

$$y_c^t = \max_{1\leqslant j\leqslant 1200}(y_{cj}^t) \tag{4}$$

Finally, these features are fed into three FC layers to fit the drug-target binding affinity. The neural node numbers for these three layers are 1024, 1024, and 512, respectively. Between each two neighbor layers are a dropout layer and a following ReLu-activating layer. The final FC layer following these three layers will give an affinity value between the drugs and the targets.

3 Experiment

3.1 Data pre-processing

Davis and KIBA data sets are the benchmarks for drug-target binding affinity prediction. Table 1 summarizes both data sets. Data preprocessing is performed following Zhao et al.[14] A total of 64 types of characters exist in all SMILES sequences and 21 types of amino acid characters in all target sequences strings. The SMILES sequences of drugs are padded or truncated to a length of 100 and the amino acid sequences of targets to 1200. The matrix dimensions for both are 100×64 and 1200×21, respectively. Both fingerprints are pre-calculated into 1024-dimension vectors.

3.2 Slice similarity analysis

Different targets have different sets of slices. For example, different targets have different kinds of domains, and each domain involves a slice set. The target fingerprint is a 1024-dimension vector. Each bit of this vector indicates a kind of slice. To validate that those different domains contain different kinds of slices and generate different bit sets in a target fingerprint, the slice similarity in a domain and between different domains is judged. The first experiment makes a statistic of the slice similarity of domains which is calculated as Cosine Similarity (CosSim) between two 64-dimension slice vectors (see Eq. (5), in which two vectors are denoted as

Table 1 Summary of Davis and KIBA data sets.

Data set	Number of targets	Number of ligands	Number of interactions
Davis	442	68	30 056
KIBA	229	2111	118 254

a and b, and k is the dimension of a vector). The target domains and their corresponding amino acid sequences are determined by the Pfam database. The vectors are calculated using the word2vec model. The similarity of slices in a domain is called the Self-Similarity (SSim), and that between different domains is called the Cross-Similarity (CSim). Considering the considerable number of slices in a domain, the mean similarity of both types is calculated to determine whether slices in different domains are less similar than in the same domains. The mean SSim and mean CSim are calculated as Eqs. (6) and (7). k or l is the index of the slice in a domain. The numbers of slices in two different domains are denoted as n and m.

$$\mathrm{CosSim} = \frac{\sum_{k=1}^{n} a_k b_k}{\sqrt{\sum_{k=1}^{n} a_k}\sqrt{\sum_{k=1}^{n} b_k}} \tag{5}$$

$$\mathrm{SSim} = \frac{\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} \mathrm{CosSim}_{kl}}{\frac{1}{2}n(n-1)} \tag{6}$$

$$\mathrm{CSim} = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m} \mathrm{CosSim}_{kl}}{nm} \tag{7}$$

The second experiment compares CNN models using DCBs or NCLs to construct the convolutional framework. FingerDTA embeds global information extracted from a fingerprint, which is fused with the corresponding convolutional features similar to an attention-like process. Furthermore, two CNN models are compared with FingerDTA, DeepDTA[8] and AttentionDTA[14]. DeepDTA is a CNN model without an attention-like process that performs better than traditional machine learning models. AttentionDTA performs an attention-like process on DeepDTA between the convolutional features of drug and target, which can be interpreted as the interaction between the whole target and the whole drug. Regarding performance, FingerDTA is compared with the above two models[21]. These models are all deep learning models for predicting drug-target binding affinity.

3.3 Baselines

To verify the effectiveness of our model, comprehensive experiments were conducted on the following baselines.

DeepDTA builds the base framework, which only utilizes sequence information. It extracts features by CNN fully blocks[8].

AttentionDTA averages the attention mechanism to combine drug and protein information based on DeepDTA, which can be treated as the interaction between the entire target and the entire drug[14].

GraphaDTA mines drug data information through modules, such as graph convolution[21].

3.4 Model settings and metrics

The immediate fingerprint features are 96-dimension vectors to fit the dimension of convolutional features for drug and target. Adam optimizer[22] is used for model training with a learning rate of 0.0001. Batch sizes for Davis and KIBA are 64 and 128, as advised by Hakime et al.[8] Each model is trained for 300 epochs to gain a stable performance. The dropout rate is set to 0.5 according to conventions.

All models are evaluated by a fivefold cross validation following Öztürk et al.[8] The evaluating metrics are MSE, CI and r_m^2 index. MSE is used to measure the difference between the predicted value and the real affinity score. CI is utilized to measure whether the predicted affinity scores of two pairs are in the same order as their labels were. r_m^2 shows the predictive performance of the models. Furthermore, the lower value of MSE shows the predicted scores are closer to the real score. For CI and r_m^2, the higher values indicate better performance.

4 Result

4.1 Distribution of slice similarities

The mean cosine similarities of slices are shown in Fig. 4, which comprise slices in the same domain and between CSim. The distribution of SSim is higher than CSim. This indicates that slices in the same domain are more likely to be clustered into the same class and refer to the

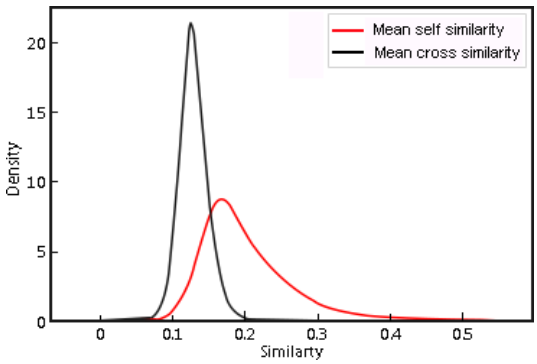

Fig. 4 Distribution of the mean similarities of slices. The curves show slices in the same domain belonging to the same class, which corresponds to the same bit in the target fingerprint.

same bit in the target fingerprint.

For example, the cross-similarity between the Pkinase_Tyr (kinase associated-1, KA1) domain in target Q7KZI7 and FGFR3_TM domain in target P11362 is lower than the SSim of those two (see Fig. 5). However, the cross-similarity between the Pkinase_Tyr domain in target Q7KZI7 and target P11362 is similar to their SSim (see Fig. 6). The above analysis indicates that similar slices tend to be clustered into the same class. Different targets with different domains contain different slice classes and generate different fingerprints. These fingerprints reflect all kinds of amino acid arrangements in a target, which is the general information of the target.

Although some bits of different domains fall into some same bits, this confliction can be reduced by enlarging the dimension of the fingerprint vector as the authors proposed the ECFPs in Ref. [15].

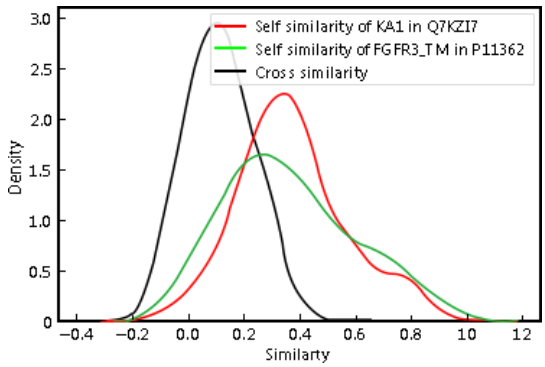

Fig. 5 Example of slice similarity distribution of different domains.

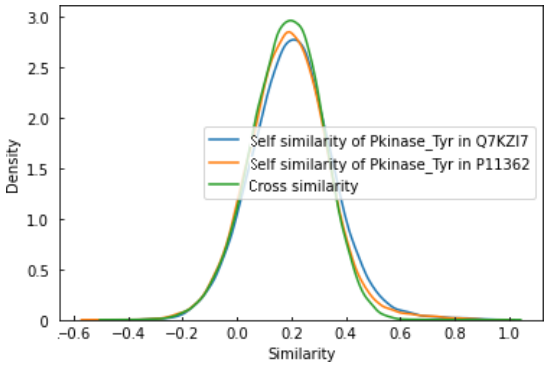

Fig. 6 Example of slice similarity distribution of the same domain.

4.2 Models utilizing attention-like process

Evaluating metrics on the testing data of Davis and KIBA data sets are shown in Tables 2 and 3, which show the main result comparison with the state-of-the-art models. Each metric is shown as the mean value of the fivefold cross validation result with the standard deviation in brackets. As Tables 2 and 3 demonstrate, the FingerDTA achieves the best performance in both data sets. Tables 4 and 5 present that the models utilizing attention-like processes (the AttentionDTA and the FingerDTA) exceed those (the DeepDTA) without any processing. Compared to building CNN models with NCLs, the origin DeepDTA utilizing our proposed attention-like process (the FingerDTA) performs best in both data sets with the lowest MSE value and the highest

Table 2 Comparison with state-of-the-art models on the Davis data set.

Model	MSE ↓	CI ↑	r_m^2 ↑
DeepDTA	0.242 (0.009)	0.883 (0.005)	0.674 (0.011)
AttentionDTA	0.241 (0.007)	0.885 (0.006)	0.668 (0.014)
GraphDTA	0.276 (0.003)	0.689 (0.001)	0.386 (0.022)
FingerDTA	**0.234 (0.003)**	**0.895 (0.002)**	**0.678 (0.008)**

Table 3 Comparison with state-of-the-art models on the KIBA data set.

Model	MSE ↓	CI ↑	r_m^2 ↑
DeepDTA	0.186 (0.003)	0.854 (0.002)	0.677 (0.005)
AttentionDTA	0.174 (0.002)	0.861 (0.002)	0.697 (0.004)
GraphDTA	0.157 (0.006)	0.859 (0.004)	0.505 (0.030)
FingerDTA	**0.150 (0.001)**	**0.885 (0.001)**	**0.750 (0.005)**

Table 4 Ablation results on the testing data of the Davis data set.

Model	Using NCLs			Using DCBs		
	MSE ↓	CI ↑	r_m^2 ↑	MSE ↓	CI ↑	r_m^2 ↑
DeepDTA	0.242 (0.009)	0.883 (0.005)	**0.674 (0.011)**	**0.234 (0.004)**	**0.897 (0.002)**	0.667 (0.010)
AttentionDTA	0.241 (0.007)	0.885 (0.006)	0.668 (0.014)	0.234 (0.007)	0.897 (0.002)	**0.688 (0.006)**
FingerDTA	**0.236 (0.001)**	**0.886 (0.002)**	0.671 (0.007)	0.234 (0.003)	0.895 (0.002)	0.678 (0.008)

Table 5 Ablation results on the testing data of the KIBA data set.

Model	Using NCLs			Using DCBs		
	MSE ↓	CI ↑	r_m^2 ↑	MSE ↓	CI ↑	r_m^2 ↑
DeepDTA	0.186 (0.003)	0.854 (0.002)	0.677 (0.005)	0.162 (0.005)	0.874 (0.006)	0.738 (0.019)
AttentionDTA	0.174 (0.002)	0.861 (0.002)	0.697 (0.004)	0.158 (0.001)	0.879 (0.001)	0.723 (0.003)
FingerDTA	**0.162 (0.001)**	**0.876 (0.002)**	**0.715 (0.012)**	**0.150 (0.001)**	**0.885 (0.001)**	**0.750 (0.005)**

CI and r_m^2 values. These results indicate that fingerprint embedded global information promotes the capability of the models in predicting drug-target binding affinity. These results indicate that fingerprint-embedded global information promotes models' capability of predicting drug-target binding affinity.

When using DCBs to construct these three CNN models, all of them exhibit better performance than those using NCLs. Three models show similar metrics on the Davis data set. This may be due to the small data size of the experimental interaction, and convolutional models using DCBs are strong enough to predict drug-target binding affinity without an attention-like process. When the models use DCBs, a significant difference between them emerges on the KIBA data set. Moreover, the size of the experimental interaction is three times larger than on the Davis data set. FingerDTA performs the best on the KIBA data set. DCB is more suitable for assembling CNN models on drug-target binding affinity prediction. These results also reveal that the data size is a restriction for the efficiency of the fingerprint.

4.3 Comparison experiments

In comparing FingerDTA with two models, FingerDTA performs the lowest MSE value and highest CI and r_m^2 value on Davis and KIBA data sets. This result indicates that global information can be extracted from fingerprints by two FC layers. This information helps FingerDTA to be more powerful in predicting the drug-target binding affinity.

4.4 Case study

As FingerDTA aims to discover potential drugs, we ground FingerDTA on a special target. A case study is performed to show its power. We choose the spike protein of COVID-19 as the special target, which is an important receptor for COVID-19. In this case, FingerDTA tries to find some potential molecules that can deactivate COVID-19 by combining them with the critical spike target. Given that the number of interactions in the Davis data set is smaller than that in the KIBA data set, FingerDTA is trained and screens candidate molecules in the KIBA data set. Interestingly, targets in the KIBA data set are all from humans but not viruses. Nevertheless, targets in viruses and humans share characteristic similarities in the gene[23]. However, collecting experimental affinity values between drugs and virus targets for training is difficult. Given the few studies of drug-target binding affinity information between these molecules and spike targets, we collect and examine some virus-inhibiting information related to these molecules or their derivatives to show their potentiality as drugs.

The top 10 candidate drugs with high affinity values are shown in Table 6. Three of them are unknown small molecules. There exist three Alternariol-related compounds or derivatives. Alternariol 5-O-Methyl Ether is known to have HIV-inhibiting activity[25], and HIV-1 shares similar mode of actions with SARS for viral entry[26]. SCH-47112 is a staurosporine derivative, and staurosporine can inhibit Rous sarcoma virus

Table 6 Top 10 drugs with high affinity score to spike.

Drug	Affinity value	ChEMBL synonyms	Deactivating virus activity information
CHEMBL554993	14.91	Alterporriol G/H	A tetrahydroanthraquinone derivative, Some tetrahydroanthraquinone has anti-viral activity[24]
CHEMBL483526	14.94	Alternariol 5-O-Methyl Ether	Inhibitor of HIV-1[25] HIV-1 share similar mode of actions with SARS for viral entry[26]
CHEMBL512054	14.94	Altersolanol A	An anthraquinone derivative, Anthraquinones can inactivate enveloped viruses[27] A potential inhibitor for COVID-19 by docking study[28]
CHEMBL291126	14.97	SCH-47112	A staurosporine derivative, Staurosporine can inhibit Rous sarcoma virus transformation[29]
CHEMBL520144	14.98	Unknown	Unknown
CHEMBL495727	**15.01**	**AT9283**	**It has COVID-19 deactivating activity[30]**
CHEMBL83790	15.09	Unknown	Unknown
CHEMBL519982	15.09	Alternariol	Its derivative Alternariol 5-O-Methyl Ether can inhibit HIV-1[25] HIV-1 shares similar mode of actions with SARS for viral entry[26]
CHEMBL1684800	15.15	Unknown	Unknown
CHEMBL483525	15.20	Alternariol 5-O-Sulfate	Its derivative Alternariol 5-O-Methyl Ether can inhibit HIV-1[25] HIV-1 shares similar mode of actions with SARS for viral entry[26]

transforming[29]. No obvious evidence points out its potentiality in deactivating COVID-19. Altersolanol A is an anthraquinone derivative. Anthraquinones can inactivate enveloped viruses[27], and it is a potential inhibitor for COVID-19 according to a docking study[28]. Alterporriol G/H is a tetrahydroanthraquinone derivative and some tetrahydroanthraquinones have anti-viral activity[24]. For example, Altersolanol A is also a kind of tetrahydro anthraquinone.

After our investigation for these drugs, none of the molecules have drug phase data, except for the 6th CHEMBL495727 on Phase II. We examined it on the ChEMBL website[31]. It is a multi-target kinase inhibitor and is a candidate drug for deactivating COVID-19 according to Ellinger et al.[30]. It gains an affinity score of 15.0, whereas the total mean score and the standard deviation are 12.11 and 0.80. This analysis confirms the benefits of FingerDTA in new drug discovery.

Further analysis of the structures of the three unknown drugs and the seven known drugs shows that they all have a large, conjugated structure between the polyaromatic ring and carbon-oxygen. Oxygen in these conjugated structures may be the binding atom to spike, such as the example in Fig. 7 (the binding site of CHEMBL483525 to spike predicted by AutoDock software). Wet experiments are still necessary to confirm it.

5 Conclusion

Drug-target binding affinity prediction is a promising way to discover new drugs for inhibiting viruses from attaching to their targets. It can provide precise guidance to save substantial manpower and material resources.

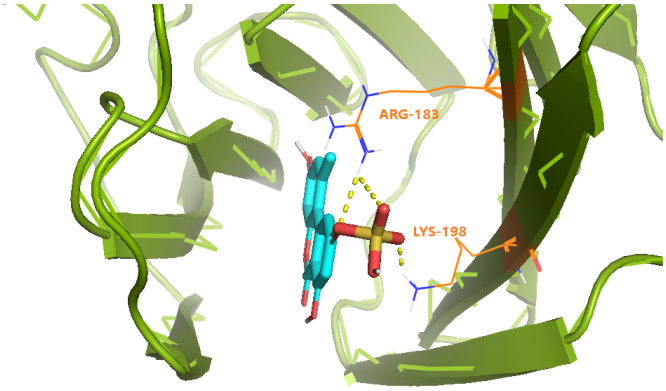

Fig. 7 Docking of CHEMBL483525 to spike. This experiment demonstrates that CHEMBL483525 may be a potential drug for COVID-19, which also indicates that FingerDTA can discover new drugs for unseen targets.

Moreover, it can accelerate the development of new drug research and provide a further reference for disease diagnosis. CNN models are powerful given their high speed and accuracy. This study first proposed a new algorithm to generate a target fingerprint. With the help of fingerprint representation, we built a fingerprint embedded framework (FingerDTA). It involves general information from the fingerprint of a drug or a target into a CNN model and promotes its performance in predicting drug-target binding affinity. The FingerDTA can help discover some potential drugs for deactivating COVID-19 by binding to the spike target. One of the top 10 predicted candidates is confirmed as a potential drug according to the ChEMBL website. FingerDTA is a powerful model for discovering new drugs. However, there remains a limit of experimental affinity data to train this model for screening drugs among gigantic scale molecules.

Future work must investigate the influence of different network structures on compound representation, such as graph neural networks. Moreover, unsupervised training methods must be introduced to utilize existing unlabeled data.

Acknowledgment

This work was funded by the China National Key Research and Development Program (No. 2019YFA0904300). The authors thank the editors and the anonymous reviewers for their helpful comments and suggestions on the quality improvement of our present paper.

References

[1] D. C. Swinney and J. Anthony, How were new medicines discovered? *Nat. Rev. Drug Discov.*, vol. 10, no. 7, pp. 507–519, 2011.

[2] Y. Bao, K. Nakagawa, Z. Yang, M. Ikeda, K. Withanage, M. Ishigami-Yuasa, Y. Okuno, S. Hata, H. Nishina, and Y. Hata, A cell-based assay to screen stimulators of the hippo pathway reveals the inhibitory effect of dobutamine on the yap-dependent gene transcription, *J. Biochem.*, vol. 150, no. 2, pp. 199–208, 2011.

[3] Y. Wang, T. Yoshihara, S. King, T. Le, P. Leroy, X. Zhao, C. K. Chan, Z. H. Yan, and S. Menon, Automated high-throughput flow cytometry for high-content screening in antibody development, *SLAS DISCOVERY Adv. Sci. Drug Discov.*, vol. 23, no. 7, pp. 656–666, 2018.

[4] M. Keusgen, Biosensors: New approaches in drug discovery, *Naturwissenschaften*, vol. 89, no. 10, pp. 433–444, 2002.

[5] S. Gupta, A. Jadaun, H. Kumar, U. Raj, P. K. Varadwaj, and A. R. Rao, Exploration of new drug-like inhibitors for serine/threonine protein phosphatase 5 of *Plasmodium falciparum*: A docking and simulation study, *J. Biomol. Struct. Dyn.*, vol. 33, no. 11, pp. 2421–2441, 2015.

[6] N. Rasool, A. Ashraf, M. Waseem, W. Hussain, and S. Mahmood, Computational exploration of antiviral activity of phytochemicals against NS2B/NS3 proteases from dengue virus, *Turk. J. Biochem.*, vol. 44, no. 3, pp. 261–277, 2019.

[7] K. Ghosh, S. A. Amin, S. Gayen, and T. Jha, Chemical-informatics approach to COVID-19 drug discovery: Exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors, *J. Mol. Struct.*, vol. 1224, p. 129026, 2021.

[8] H. Öztürk, A. Özgür, and E. Ozkirimli, DeepDTA: Deep drug-target binding affinity prediction, *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[9] H. R. Rajpura and A. Ngom, Drug target interaction predictions using PU-leaming under different experimental setting for four formulations namely known drug target pair prediction, drug prediction, target prediction and unknown drug target pair prediction, in *Proc. 2018 IEEE Conf. Computational Intelligence in Bioinformatics and Computational Biology*, St. Louis, MO, USA, 2018, pp. 1–7.

[10] L. Zhou, Z. Li, J. Yang, G. Tian, F. Liu, H. Wen, L. Peng, M. Chen, J. Xiang, and L. Peng, Revealing drug-target interactions with computational models and algorithms, *Molecules*, vol. 24, no. 9, p. 1714, 2019.

[11] F. Hu, J. Jiang, and P. Yin, Interpretable prediction of protein-ligand interaction by convolutional neural network, in *Proc. 2019 IEEE Int. Conf. Bioinformatics and Biomedicine*, San Diego, CA, USA, pp. 656–659, 2019.

[12] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, 2011.

[13] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio, Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis, *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, 2014.

[14] Q. Zhao, F. Xiao, M. Yang, Y. Li, and J. Wang, AttentionDTA: Prediction of drug–target binding affinity using attention model, in *Proc. 2019 IEEE Int. Conf. Bioinformatics and Biomedicine*, San Diego, CA, USA, 2019, pp. 64–69.

[15] D. Rogers and M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[16] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[17] RDKiT, https://www.rdkit.org/, 2021.

[18] D. Weininger, A. Weininger, and J. L. Weininger, Smiles. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 2, pp. 97–101, 1989.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, in *Proc.1st Int. Conf. Learning Representations*, Scottsdale, AZ, USA, 2013, p. 5043.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[21] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, GraphDTA: Predicting drug–target binding affinity with graph neural networks, *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2021.

[22] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proc. 3^{rd} Int. Conf. Learning Representations*, San Diego, CA, USA, 2015.

[23] V. Nada and M. Ljiljana, Gene similarity between hepatitis C virus and human proteins: A blood transfusion problem, *Med. Pregl.*, vol. 58, nos. 11&12, pp. 582–586, 2005.

[24] S. Feng and W. Wang, Bioactivities and structure–activity relationships of natural tetrahydroanthraquinone compounds: A review, *Front. Pharmacol.*, vol. 11, p. 799, 2020.

[25] J. Ding, J. Zhao, Z. Yang, L. Ma, Z. Mi, Y. Wu, J. Guo, J. Zhou, X. Li, Y. Guo, et al., Microbial natural product alternariol 5-O-methyl ether inhibits hiv-1 integration by blocking nuclear import of the pre-integration complex, *Viruses*, vol. 9, no. 5, p. 105, 2017.

[26] Y. Kliger and E. Y. Levanon, Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy, *BMC Microbiol.*, vol. 3, p. 20, 2003.

[27] R. J. Sydiskis, D. G. Owen, J. L. Lohr, K. H. Rosler, and R. N. Blomster, Inactivation of enveloped viruses by anthraquinones extracted from plants, *Antimicrob. Agents Chemother.*, vol. 35, no. 12, pp. 2463–2466, 1991.

[28] S. Das and A. Singha Roy, Naturally occurring anthraquinones as potential inhibitors of SARS-Cov-2 main protease: A molecular docking study, ChemRxiv preprint ChemRxiv: 12245270.v1, 2020.

[29] H. Nakano, E. Kobayashi, I. Takahashi, T. Tamaoki, Y. Kuzuu, and H. Iba, Staurosporine inhibits tyrosine-specific protein kinase activity of Rous sarcoma virus transforming protein p60, *J. Antibiot.*, vol. 40, no. 5, pp. 706–708, 1987.

[30] B. Ellinger, D. Bojkova, A. Zaliani, J. Cinatl, C. Claussen, S. Westhaus, J. Reinshagen, M. Kuzikov, M. Wolf, G. Geisslinger, et al., Identification of inhibitors of SARS-Cov-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection, ResearchSquare preprint ResearchSquare: rs.3.rs-23951/v1, 2020.

[31] ChEMBL, https://www.ebi.ac.uk/chembl/, 2018.

Xuekai Zhu is a master student in the School of Computer Science, Wuhan University. His current research interests are in artificial intelligence methods for bioinformatics.

Juan Liu obtained the BS and PhD degrees in computer science from Wuhan University, in 1991 and 1996 respectively. She is now a professor in the School of Computer Science at Wuhan University. Her research interests include machine learning, bioinformatics, pattern recognition, and artificial intelligence methods for medicine.

Jian Zhang is a master student in the School of Computer Science, Wuhan University. His current research interests are in artificial intelligence methods for bioinformatics.

Zhihui Yang is a PhD candidate in the School of Computer Science, Wuhan University. His current research interests include synthetic biology, machine learning, metabolic pathway reconstruction, and metabolic flux analysis.

Xiaolei Zhang is a master student in the School of Computer Science, Wuhan University. Her current research interests are in artificial intelligence methods for bioinformatics, including metabolic pathway optimization and bioretrosynthesis.

Feng Yang is a PhD candidate in the School of Computer Science, Wuhan University. His current research interests include synthetic biology, bio retrosynthesis evolutionary computation, and machine learning.