Bioinformatics, 2024, **40(1)**, btad778
https://doi.org/10.1093/bioinformatics/btad778
Advance Access Publication Date: 23 December 2023
Original Paper

OXFORD

Data and text mining

TEFDTA: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug–target affinities

Zongquan Li[1,2], Pengxuan Ren[2], Hao Yang [2], Jie Zheng [1], Fang Bai [1,2,3,*]

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai, 201210, China
[2]Shanghai Institute for Advanced Immunochemical Studies and School of Life Science and Technology, ShanghaiTech University, Shanghai, 201210, China
[3]Shanghai Clinical Research and Trial Center, Shanghai, 201210, China

*Corresponding author. Shanghai Institute for Advanced Immunochemical Studies and School of Life Science and Technology, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210, China. E-mail: baifang@shanghaitech.edu.cn

Associate Editor: Jonathan Wren

Abstract

Motivation: The prediction of binding affinity between drug and target is crucial in drug discovery. However, the accuracy of current methods still needs to be improved. On the other hand, most deep learning methods focus only on the prediction of non-covalent (non-bonded) binding molecular systems, but neglect the cases of covalent binding, which has gained increasing attention in the field of drug development.

Results: In this work, a new attention-based model, A Transformer Encoder and Fingerprint combined Prediction method for Drug–Target Affinity (TEFDTA) is proposed to predict the binding affinity for bonded and non-bonded drug–target interactions. To deal with such complicated problems, we used different representations for protein and drug molecules, respectively. In detail, an initial framework was built by training our model using the datasets of non-bonded protein–ligand interactions. For the widely used dataset Davis, an additional contribution of this study is that we provide a manually corrected Davis database. The model was subsequently fine-tuned on a smaller dataset of covalent interactions from the CovalentInDB database to optimize performance. The results demonstrate a significant improvement over existing approaches, with an average improvement of 7.6% in predicting non-covalent binding affinity and a remarkable average improvement of 62.9% in predicting covalent binding affinity compared to using BindingDB data alone. At the end, the potential ability of our model to identify activity cliffs was investigated through a case study. The prediction results indicate that our model is sensitive to discriminate the difference of binding affinities arising from small variances in the structures of compounds.

Availability and implementation: The codes and datasets of TEFDTA are available at https://github.com/lizongquan01/TEFDTA.

1 Introduction

In the realm of drug research and development, predicting drug–target interactions/affinities (DTI/DTA) is an indispensable component. In early stages, researchers determined these interactions through experiments, which is time-consuming and costly. Subsequently, with advancements of computer technology, researchers started utilizing computers to predict drug–target interactions and simulate the binding poses of drugs and targets using docking programs, such as GLIDE (Friesner *et al.* 2004), Molegro Virtual Docker (Bitencourt-Ferreira and de Azevedo 2019). However, this docking method also has corresponding limitations, i.e. the docking process also takes a long time to do the computation and requires the three-dimensional structures of proteins.

With the development of machine learning and deep learning, researchers have attempted to incorporate these fields to DTI. Kronrls (Pahikkala *et al.* 2015) and SimBoost (He *et al.* 2017) use similarity matrix to predict DTA. In recent deep learning approaches for prediction DTA, tasks are divided into two categories: binary classification and regression. The binary classification primarily aims to determine whether there is or a strong enough interaction between drugs and targets, while the regression task is more inclined to determine the strength of the combination between molecules and protein targets, which is undoubtedly more challenging.

At present, deep-learning based methods have been extensively used. An advantage of these methods is their capability to automatically extract features. However, the initial input data, particularly the data description of proteins and small molecules, significantly influences the model's performance. Pafnucy (Stepniewska-Dziubinska *et al.* 2018) constructs a $20 \times 20 \times 20$ Å^3 box centered as the geometric center of a binding ligand, and then discretized the positions of heavy atoms into a lattice with a bin size of 1 Å. RosENet (Hassan-Harrirou *et al.* 2020) combines voxelized molecular mechanics energies and molecular descriptors as input for an ensemble of three-dimensional (3D) convolutional neural networks (CNNs). OnionNet (Zheng *et al.* 2019) draws shells with each atom of small molecules as the center, incrementing by δ as the radius, resembling layers of onions. Subsequently, it gathers all pairs between the atoms of the central molecule

Received: 25 September 2023; **Revised:** 23 November 2023; **Editorial Decision:** 18 December 2023; **Accepted:** 22 December 2023

and the protein atoms within the shell, utilizing them as inputs. Kdeep (Jiménez *et al.* 2018) voxelizes the whole protein, while PointNet and PointTransformer (Wang *et al.* 2022) convert the whole protein into a point cloud and take the coordinates as the input. The majority of the aforementioned models leverage CNN architectures. Masif (Gainza *et al.* 2020) utilizes the geometric information from the protein surface as its input and uses the GNN to perform training. These approaches consider the static 3D structural information of the protein. However, the protein structure undergoes dynamic changes. Presently, our data mostly contain singular conformation of the protein, prompting the need to assess whether relying on a single pose impacts practical results. Furthermore, not all proteins have their structures determined to date.

Consequently, there are numerous advanced methods that do not directly incorporate the structure of the target protein into their inputs. DeepDTA (Öztürk *et al.* 2018) utilizes the FASTA sequence format to represent proteins, mapping each element to an amino acid, and uses the SMILES string format to characterize small molecules. DeepCDA (Abbasi *et al.* 2020) improves the feature extraction process of DeepDTA and utilizes LSTM to further extract features from proteins and drug molecules. However, these methods do not take into account the spatial structure of small molecules. GraphDTA (Nguyen *et al.* 2021) and DeepGS (Lin 2020) use the two-dimensional structures of small molecules, which contain the bonding information between atoms. MolTrans (Huang *et al.* 2021) and TransformerCPI (Chen *et al.* 2020) also use sequence data, using Transformer (Vaswani *et al.* 2017) (an excellent model that works well on language translation), to predict drug–target binding affinity. The authors of TransformerCPI have continually made notable progress in their research endeavors, leading to the development of TransformerCPI2.0 (Chen *et al.* 2023), which is a competitive sequence-based model compared to 3D structure-based methods like molecular docking. MGraphDTA (Yang *et al.* 2022) introduces a multi-scale graph neural network that not only preserves the ability to extract local features but also enables the capturing of global structural information in compounds. SAM-DTA (Hu *et al.* 2023) treats different proteins separately but trains them together in a multi-head strategy, which differs from the traditional models that process protein and ligand inputs separately and merge them later. DataDTA (Zhu *et al.* 2023) integrates multi-scale interaction information and achieves impressive predictive accuracy for binding affinity by using the dual interaction aggregation neural network strategy. However, it is worth noting that most of the aforementioned methods were primarily developed for predicting non-covalent binding affinity, and there is currently no deep learning method specifically designed for predicting covalent binding affinity.

In this paper, we present a novel model for predicting both covalent (bonded) and non-covalent (non-bonded) binding affinities in drug-protein interactions, called Fingerprint Encoder DTA (TEFDTA). Our model TEFDTA, draws inspiration from two existing models, DeepDTA (Öztürk *et al.* 2018) and TransformerCPI (Chen *et al.* 2020). DeepDTA offers a way to extract features from sequences using 1D-CNN (1-Dimensional Convolutional Neural Network). This model focuses on extracting local pattern features from sequence information to facilitate feature extraction. While Recurrent Neural Networks (RNN) can also process one-dimensional inputs and perform feature extraction, they suffer from certain limitations. CNN, on the other hand, has constraints in capturing global features effectively. RNN, despite being able to process the entire sequence with network propagation, encounters the problem of forgetting information over time. TransformerCPI demonstrates that Transformers can effectively address the issues present in both CNN and RNN models. Transformers are built upon an encoder and a decoder. In light of this, we utilize the Transformer as a feature extractor to distill complex molecular sequences. It is important to note that a single encoder is sufficient for this task, as more complex models would require significantly longer training times due to potential convergence difficulties, without necessarily enhancing information extraction. The main contributions of this paper are concluded as follows.

- We propose a module that can extract features of protein and drug molecules separately to improve the accuracy of the model in modeling protein molecular interaction problems. Resulting in near state-of-the-art performance on the public datasets Davis (Davis *et al.* 2011) and KIBA (Tang *et al.* 2014).
- We take the lead on predicting covalent binding affinity using deep learning methods, and experiments have shown that the prediction of covalent binding affinity still has certain accuracy.

2 Materials and methods

2.1 Datasets

Our DTI model was trained and tested with three non-covalent drug target binding benchmark databases (details as below), KIBA (Tang *et al.* 2014), Davis (Davis *et al.* 2011), and BindingDB (Liu *et al.* 2007), which are widely used for DTI tasks. We then fine-tuned the model with the covalent database CovalentInDB (Du *et al.* 2021) based on the model trained with BindingDB.

1) KIBA is the abbreviation for Kinase Inhibitor BioActivity, which combines three biochemical analysis indices K_i (inhibitory constant), K_d (dissociation constant) and IC_{50} (half maximal inhibitory concentration). The KIBA database contains 52 498 drugs and 467 targets, and there are 246 088 KIBA scores in total. However, in SimBoost, the database was filtered. The filtered KIBA database has a total of 2111 drugs and 229 target proteins, with a total of 118 254 scores for drug–target bioactivities.
2) The database proposed by Davis *et al.* (2011) is mainly contains the bioactivity data in the format of K_d, with a total of 442 proteins and 68 drugs, and a total of 30 056 values. In order to make our model comparable with the results of other models (Öztürk *et al.* 2018), we transform K_d value to pK_d [Equation (1)].

$$pK_d = -\log\left(\frac{K_d}{1e9}\right) \tag{1}$$

In addition, we found that the sequence information of some proteins was not correctly recorded in the Davis database, which is provided by DeepDTA and serves as a benchmark dataset for DTA tasks. This issue further affects the accuracy of models that utilize incorrect data.

In particular, mutant amino acid sequences were still represented with the wide type, which could also affect the accuracy of our prediction. Therefore, we corrected these sequences to ensure their accuracy.

3) BindingDB is a public, web-accessible database of measured binding affinities. It was launched on the web in 2000 and currently contains about 2.7 million binding data for 9000 protein targets and 1.2 million small molecules. After data cleaning, including the removal of data with unclear binding affinity data and duplicate data in the database, the remaining dataset contains 80 324 compound molecules and 5561 proteins, with a total of 1 254 402 interaction data. The summary of the non-covalent datasets, and how we divide it into training set, validation set and test set are shown in the Table 1.
4) CovalentInDB is a comprehensive database for covalent inhibitors and their targets. It contains not only the basic information of covalent inhibitors such as warheads, reaction mechanisms, and binding sites, but also the experimental data of covalent binding affinity, i.e. IC_{50}, K_d and K_i. The dataset contains 4511 covalent inhibitors (including 68 approved drugs) with 57 different reactive warheads for 280 protein targets. We performed a statistical analysis of our covalent data. The six most abundant targets are Kinase, Acyltransferase, Aminopeptidase, Hydrolase, Oxidoreductase, and Protease. The remaining types were categorized into a class called others, as illustrated in Fig. 1A. The ratio of each type of target was marked in the figure, with the largest proportion being Kinase (61%), followed by Protease, which accounts for about 13.2%. The types of the residue types involved in the covalent reaction have been clustered, and the results show that the observed residue types with the highest percentage are cysteine (66.5%), serine (18.7%), histidine (5.2%), and tyrosine (4.7%), as illustrated in Fig. 1B. The largest portion of the types of ligand warheads is michael acceptor (44.9%), followed by the types of carbonyl (5.2%), nitrile (5%), etc., as illustrated in Fig. 1C. We then performed fine-tuning based on each specific warhead type.

2.2 Model Architecture

To enhance the information extraction from drug and protein sequences, we have developed a novel framework called Fingerprint Encoder DTA, or TEFDTA for short. First, our model obtains FASTA sequence information of the protein and the SMILES of the drug molecule, and extracts the corresponding features. Proteins are performed label encoding and put into the embedding layer. Obtained features are fed into CNN blocks consisting of 1D-CNN layers. On the other hand, the drug molecules are converted from the SMILES format into a fingerprint type known as MACCS. These fingerprints are then fed into an embedding layer where position encoding is performed. Following this, the drug molecules' features are successfully extracted through the encoder module in the transformer. Once the features of the protein and drug molecule have been extracted, we concatenate the two representation vectors and input them into a fully connected layer. Finally, the model predicts the binding affinity score for a pair of ligand-protein. See Fig. 2 for the framework of the model.

2.3 Design of the input of the network architecture

2.3.1 Representation of compounds

The MACCS key is a binary fingerprint consisting of 166 bits. Each bit position represents the presence 1 or absence 0 of a predefined structural feature in the drug molecule. Through our research, we have discovered that using the MACCS key as input allows for better extraction of drug molecule features compared to using the SMILES representation. This improvement can be attributed to the MACCS fingerprint's ability to capture substructure information of compounds, thereby providing more informative input for the model.

Furthermore, using the MACCS fingerprint offers an additional advantage. The sequence lengths of most drug molecules are not uniform, which would require padding the sequences with zeros when using the SMILES representation. However, by utilizing the fingerprint, we can avoid this step altogether, as the fingerprint has a fixed length and does not require padding.

2.3.2 Representation of proteins

In our approach, we use integer encoding to represent amino acids in the protein sequences. We have defined 25 amino acids, and each amino acid is assigned a corresponding integer value. (e.g. "A": 1, "C": 2, "B": 3, "E": 4, etc.). The label encoding for an example of FASTA, "MAAVIL" is given below:

$$[\mathrm{M\ A\ V\ I\ L}] = [11\ 1\ 1\ 22\ 8\ 12]$$

After encoding the drug molecules and protein sequences, the resulting vectors are passed through an embedding layer to generate a matrix or feature map. For drug molecule vectors, the transformed matrix is denoted as $M_D \in R^{L_D \times E_D}$. Where L_D represents the sequence length of drug molecules and E_D represents the embedding dimension. For protein vectors, the transformed matrix is $M_P \in R^{L_P \times E_P}$. Where L_P stands for the sequence length of the proteins, with a cutoff set at 1000. E_D represents the embedding dimension. Specifically, the representation of drug molecules undergoes a step called positional encoding. Although Transformers can effectively address the limitations of CNN and RNN in processing sequential information, they are unable to distinguish the same character at different positions. To overcome this, we apply positional encoding. For each different position, there is a separate feature map $M_{Pos} \in R^{L_D \times E_D}$, which has the same dimension as the feature map of molecules. Finally, after the position encoding is completed by simple addition, it is entered into the transformer block $M_T = M_D + M_{Pos}$.

Table 1. Statistical analysis of benchmark datasets and the division of training, validation, and test sets.

Dataset	No. of compounds	No. of proteins	No. of interactions	Training set	Validation set	Test set
KIBA	2111	229	118 254	78 836	19 709	19 709
Davis	68	442	30 056	20 037	5009	5010
BindingDB	803 234	5561	1 254 402	1 172 682	81 720	20 001

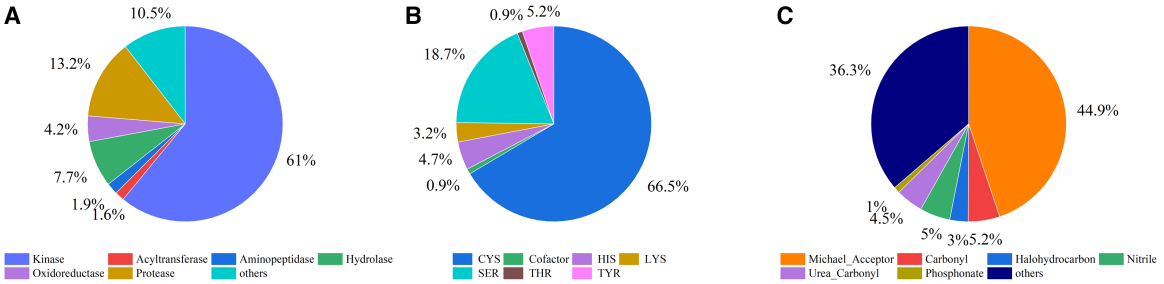


Figure 1. The pie charts present the distribution analysis of covalent data used for finetuning. (A) Proportion analysis of target types in the covalent dataset. (B) Proportion analysis of covalent residue types in the dataset. (C) Proportion analysis of covalent warhead types in the dataset.

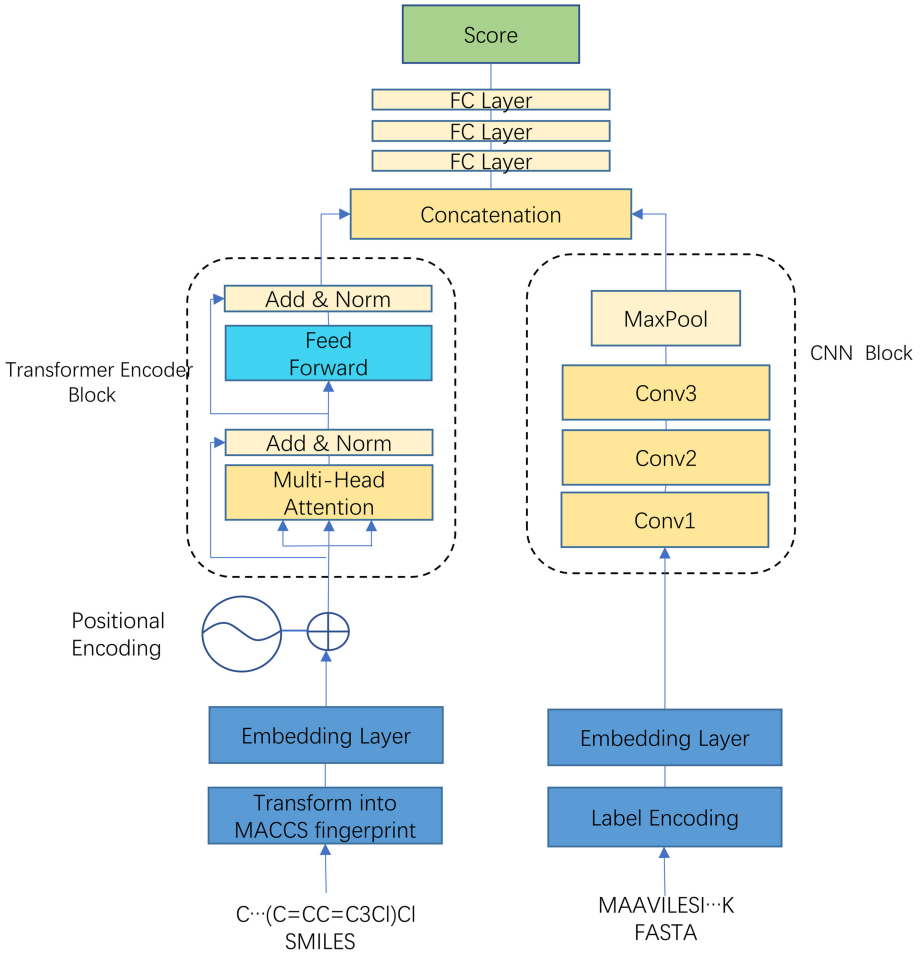


Figure 2. The framework of the designed module TEFDTA for drug–protein binding affinity prediction.

2.3.3 Transformer encoder block

For the encoder in the transformer, detailed description can be seen in Vaswani *et al.* (2017). Briefly, the molecular feature map M_T is fed into a module called the multi-head attention block, which is structured as follows: first, Q, K, and V are calculated:

$$Q = M_T W_Q,\ K = M_T W_K,\ V = M_T W_V \tag{2}$$

$Q,\ K,\ V \in R^{L_D \times E_D}$ are query, key, and value, and generated based on the projection of the original feature map. $W_Q,\ W_K, W_V \in R^{E_D \times E_D}$ are projection matrices that need to be used to compute. Then do self-attention operation on Q, K, and V matrices, it is formulated as:

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

For $M_T \in R^{L_D \times E_D}$, we split the dimension E_D. In this work, we use $h(\text{head}) = 8$ and $E_D = 256$, so that d_k is formulated as a dimension in each attention operation as $d_k = E_D/h = 32$:

$$\text{MultiHead}(Q,\ K,\ V) = \text{Concat}(\text{head}_1 \ldots, \text{head}_h)W^O \tag{4}$$

$$\text{head}_i = \text{Attention}(Q^i, K^i, V^i) \tag{5}$$

When we extract different features from different dimensions of the input feature map, and finally concatenate them together, we project them by W^O to the same dimension as the original one. Adding and normalizing stand for residual connection and layer normalization respectively, which are not shown here, and feed forward is a simple MLP. Finally, the output of the transformer-encoder block $M_O \in R^{L_D \times E_D}$ has the same dimension as the input. However, the features of the drug molecules were extracted.

2.3.4 CNN block

To address the memory overhead and increased parameter requirements associated with using the transformer's encoder for training due to the long length of protein's FASTA sequence, we opted for the same 1D-CNN structure as DeepDTA to extract protein features. For the input protein feature map $M_P \in R^{L_P \times E_P}$, like 3D CNN used in the image field, we have a one-dimensional convolution kernel with a fixed field of view h and a size of $k_1 \in R^{h \times E_P}$, which detects and extracts the features of h amino acids according to the model. After a convolution operation, we will get a new protein feature representation $M_p^{(2)} \in R^{(L_P-h+1)\times E_P}$. After repeating the convolution for three times, the final protein representation $M_P \in R^{(L_P-b_1-b_2-b_3+3)\times E_P}$ is obtained by a Max pooling layer. For simplicity, we consider it as $M_P \in R^{L_P' \times E_P}$.

After obtaining the feature map of the drug molecule $M_O \in R^{L_D \times E_D}$ and the feature map of the protein $M_P \in R^{L_P' \times E_P}$, we perform the MaxPooling operation to reduce the dimensions to one. Then do concatenation to merge the vectors $M_C \in R^{L_D+L_P'}$, put them into three fully linked layers and finally output the binding affinity score.

3 Experiments and results

3.1 Evaluation metrics

We use three metrics to evaluate our model: mean squared Error (MSE), concordance index (CI), and a type of correlation index for prediction and ground truth r_m^2.

MSE is the loss function:

$$\mathrm{MSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{6}$$

The CI is another evaluation indicator for assessing the consistency in ranking the strength of binding affinities for two randomly selected pairs of drug and target between predicted and actual values.

$$\mathrm{CI} = \frac{1}{N} \sum_{y_i > y_j} h(\hat{y}_i > \hat{y}_j) \tag{7}$$

N is a normalization constant equal to the number of data pairs. $h(x)$ is the step function and defined as:

$$h(x) = \begin{cases} 1, & x > 0 \\ 0.5, & x = 0 \\ 0, & x < 0 \end{cases} \tag{8}$$

r_m^2 is evaluating metrics for model external prediction performance.

Table 2. Performance comparison of different models on Davis.

Model	CI (SD)	MSE	r_m^2 (SD)
KronRLS	0.871 (0.001)	0.379	0.407 (0.005)
SimBoost	0.872 (0.002)	0.282	0.644 (0.006)
DeepDTA	0.878 (0.004)	0.261	0.630 (0.017)
DeepCDA	**0.891 (0.003)**	0.248	0.649 (0.009)
TEFDTA	0.890 (0.002)	**0.199**	**0.756 (0.008)**

Bold and underline corresponds to the best performance for each metric.

Table 3. Performance comparison of different models on KIBA.

Model	CI (SD)	MSE	r_m^2(SD)
KronRLS	0.782 (0.001)	0.411	0.342 (0.001)
SimBoost	0.836 (0.001)	0.222	0.629 (0.007)
DeepDTA	0.863 (0.002)	0.194	0.673 (0.009)
DeepCDA	**0.889 (0.002)**	**0.176**	0.682 (0.008)
TEFDTA	0.860 (0.001)	0.184	**0.731 (0.006)**

Bold and underline corresponds to the best performance for each metric.

$$r_m^2 = r^2 * \left(1 - \sqrt{r^2 - \gamma_0^2}\right) \tag{9}$$

The squared correlation coefficients, r^2 and γ_0^2 quantify the relationship between observed and predicted values, accounting for and excluding the intercept term, respectively. Only when the r_m^2 is greater than 0.5 proves the model reliable.

3.2 Comparison of the prediction efficiency

We first compared our model with the following benchmark models: KronRLS (Pahikkala *et al.* 2015), SimBoost (He *et al.* 2017), DeepDTA (Öztürk *et al.* 2018), and DeepCDA (Abbasi *et al.* 2020). Among them, KronRLS and SimBoost are similarity-based models. DeepDTA and DeepCDA are sequence-based models. The test results for these methods on the test set are summarized in Tables 2 and 3. As for the experimental results, we directly used the experimental data reported in the publications of these methods. We adopted this approach to avoid potential errors that may arise from implementation and training.

For the Davis dataset, TEFDTA and DeepCDA outperformed the other benchmark models. TEFDTA achieved superior performance compared to DeepCDA with a margin of 0.049 for mean squared error (MSE) and 0.107 for r_m^2. On the other hand, DeepCDA achieved the best CI performance among all the models evaluated.

For the KIBA dataset, TEFDTA and DeepCDA also achieved the best performance metrics. DeepCDA outperformed TEFDTA by 0.002 for CI and 0.015 for MSE. However, TEFDTA achieved the best r_m^2 among all the models evaluated.

In order to have a clearer understanding of the test results, Fig. 3 illustrates the performance evaluation of our model in predicting binding affinities for two datasets: Davis and KIBA. The figure showcases the predicted values plotted against the measured values, emphasizing the effectiveness of our model. For the Davis dataset, the overall prediction results demonstrate a satisfactory agreement with the measured values. However, it is noteworthy that a cluster of points is concentrated around the value of five. This anomaly arises from the approximation made during data organization, where affinity values below 10 000 nM were

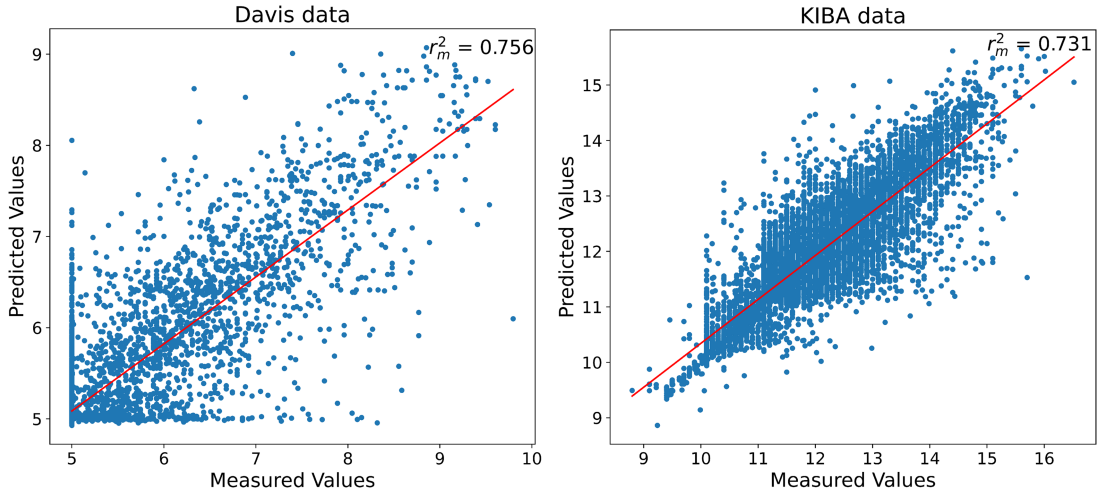


Figure 3. Comparison of correlation between the predicted and measured values for Davis and KIBA Data.

Table 4. Ablation experiments on feature extraction methods and model framework modifications in the Davis dataset.

Different method	CI (SD)	MSE	r_m^2 (SD)
Only Fingerprint	0.877 (0.002)	0.235	0.679 (0.008)
Fingerprint with CNN	0.879 (0.002)	0.232	0.719 (0.003)
SMILES with encoder	0.874 (0.001)	0.228	0.724 (0.004)
Proposed model	**0.890 (0.002)**	**0.199**	**0.756 (0.008)**

Bold and underline corresponds to the best performance for each metric.

Table 5. Different models' prediction performance on the dataset of BindingDB.

Model	CI (SD)	MSE	r_m^2 (SD)
DeepDTA	0.795	0.812	0.618
DeepCDA	0.811	0.832	0.628
TEFDTA	**0.814**	**0.701**	**0.631**

Bold and underline corresponds to the best performance for each metric.

approximated to 10 000 nM, introducing a certain degree of bias. On the other hand, the KIBA dataset exhibits a closer correspondence between the predicted and measured values, highlighting the model's capability to accurately predict binding affinities.

4 Ablation experiments

The use of fingerprint transformation and Transformer encoder in our method significantly contributes to the overall performance by enhancing the understanding of molecular representation. To assess the specific impact of these components, we conducted ablation experiments where we removed or modified parts of the module. The results of these experiments are presented in Table 4.

In the first modification, we removed the Transformer encoder module from the architecture. This experiment helps us evaluate the contribution of the Transformer encoder in extracting high-level drug information.

In the second modification, we replaced the Transformer encoder module with a 1D-CNN module. This alteration allows us to compare the performance of the Transformer encoder against the 1D-CNN in terms of drug representation extraction.

In the third modification, we directly used SMILES as the input for the encoder, bypassing the fingerprint transformation step. This experiment helps us understand the impact of using the fingerprint transformation in capturing important substructure information of the drug molecules.

Through these ablation experiments, we can observe the advantages of using fingerprint transformation and Transformer encoder as the drug representation extraction modules. These components enable the model to capture and utilize high-level drug information effectively, leading to improving overall performance in drug–target interaction prediction.

5 Binding affinity prediction for covalent ligands

After confirming the accurate prediction capability of our model for non-covalent binding affinity, we conducted training and testing on a larger dataset, BindingDB, and compared its performance with DeepDTA and DeepCDA (refer to Table 5). Due to the lack of detailed hyperparameter information in DeepCDA's publication, we used their experimental results directly. For DeepDTA, we conducted training on the dataset while ensuring consistency with our model.

Subsequently, we gathered data containing covalent binding affinities from the CovalentInDB database and used the model trained on the BindingDB database. However, the experimental results revealed poor prediction performance, indicating that the model did not effectively capture the patterns associated with covalent binding.

To address this limitation, we classified the data in CovalentInDB and identified six commonly occurring warheads. We performed fine-tuning on the model specifically for each warhead category and utilized it to predict the covalent binding data for individual warhead types. The comparison between the pre-fine-tuned and post-fine-tuned data is presented in Fig. 4. Evaluation metrics such as mean squared error (MSE) and Pearson's correlation coefficient were used.

The results demonstrate that prior to fine-tuning, the model trained solely on the BindingDB dataset exhibited inadequate performance in predicting covalent binding affinity.

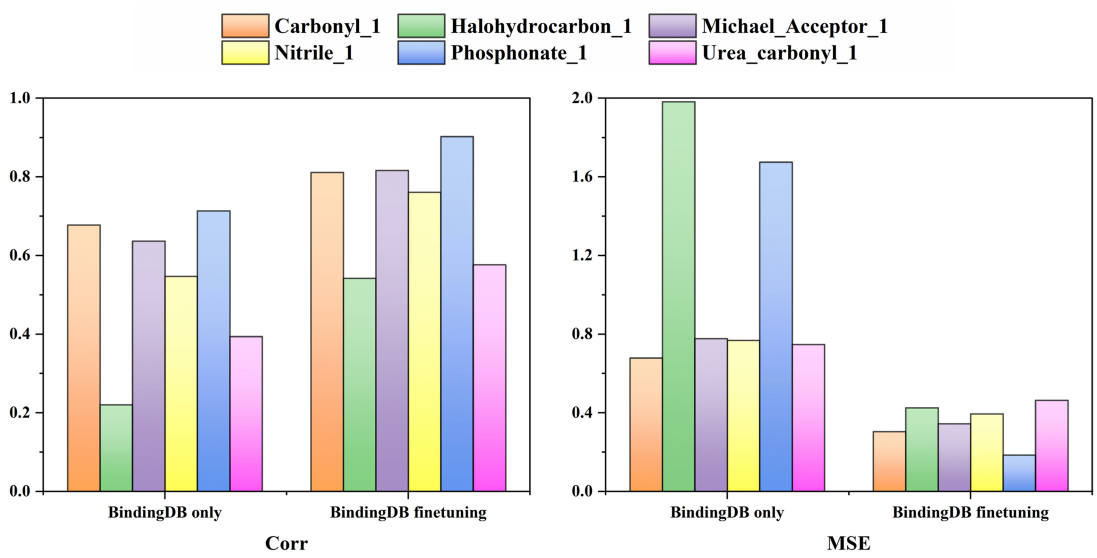
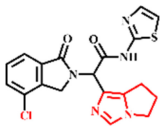
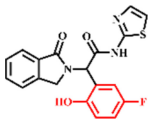
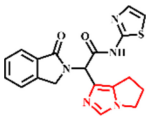
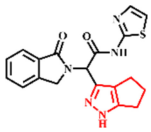
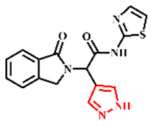


Figure 4. Comparison of covalent binding affinity predictions before and after fine-tuning on six common warheads.

Table 6. Results for investigating our model's sensitivity to discriminating the difference of binding affinities (pKd) arising from the minor structure variance.

Compound structure					
Measured value	8.39	8.22	7.79	7.63	6.02
Predicted value	6.16	6.41	5.89	5.89	5.74

However, after undergoing the fine-tuning process, the model became more proficient in recognizing the covalent warhead characteristics and achieved relatively accurate predictions for the corresponding covalent warheads. Fine-tuning effectively enhanced the model's ability to significantly improve the accuracy of covalent binding affinity prediction.

By undertaking fine-tuning on specific covalent warheads, we observed a considerable improvement in the model's prediction accuracy for covalent binding affinity. This underscores the importance of specialized training and fine-tuning when dealing with covalent binding data.

6 Case study

We conducted a case study to demonstrate the application of our proposed TEFDTA model in the field of drug design and, in particular, to investigate the ability of our method to discriminate molecules' activity cliffs. For this purpose, a set of novel heterodimeric inhibitors of epidermal growth factor receptor (EGFR)L858R, which were not included in either our training or test set, was selected them to perform the prediction (Obst-Sander *et al.* 2022).

These molecules have the same scaffold but different substituents in the same position, as shown in Table 6. It is worth mentioning that some of these molecules possess identical sequences due to the shared backbone structure. The aim of this case study was to investigate whether our model could accurately capture subtle variations in substituents for molecules with similar sequences.

As shown in Table 6, results from our experiments indicate that our model can accurately rank the strength of these analogues, despite not being able to guarantee absolute accurate values for the binding affinities. This consistency between the model's prediction and the actual measurements indicates the model's sensitivity to local structural changes within molecules. Moreover, it suggests a certain level of robustness and the model's capability to approximate the strength of binding affinity.

7 Conclusion

In this study, we propose TEFDTA, an approach for accurate prediction of drug–target interaction. Our method incorporates fingerprint transformation and Transformer encoder modules to enhance molecular representation understanding. To evaluate TEFDTA's performance, we conduct experiments on the Davis, KIBA, and BindingDB dataset, and compare the results with other binding affinity prediction models, i.e. DeepDTA and DeepCDA. The results confirmed TEFDTA's performance in binding affinity prediction. In addition, this model was further optimized by fine-tuning over a dataset of bonded protein–ligand interactions from the database CovalentInDB. Covalent binding data is classified based on common warheads, and individual fine-tuning is performed for each warhead category. The results demonstrate that fine-tuning process significantly improves the model's prediction accuracy for covalent binding affinity, emphasizing the importance of specialized training. Furthermore, we conduct a

case study on predicting the binding affinity of drug molecules targeting the EGFR. Our results indicate that while the model may not precisely predict the exact binding affinity values for molecules with identical backbone structures but different substituents, it is able to capture the trend of affinity variance introduced by a different substituent on the molecule. This suggests the model's potential sensitivity to local structural changes and its capability to approximate binding affinity strength, which needs to be confirmed with a larger size dataset for further evaluation or training.

In conclusion, TEFDTA, with its incorporation of fingerprint transformation and Transformer encoder modules, provides an improved approach for accurate prediction of drug–target interactions. However, it is important to acknowledge the limitations of our model. While the model may successfully capture the effect of minor changes in the molecular sequence on affinity, it is not as sensitive to mutations in the protein segment, including single or few amino acid changes. For virtual screening tasks, it is very valuable to detect observable changes in affinity when mutations occur. It is also possible for us to achieve this goal. The direct extraction of features from the FASTA sequence of the protein makes it difficult to achieve this aim because single amino acid mutations are imperceptible in the embedding of the entire protein. However, with the advent of large language models, it has become possible to extract representations of proteins with unsupervised learning by pre-training these models on a large number of protein sequences. By fine-tuning the model with downstream tasks such as data having mutations, the model becomes sensitive to key amino acids. For example, in our current covalent database CovalentInDB, most of the covalent drug targets are cysteines, and the model is likely to be very sensitive to the amino acid cysteine. As soon as a change occurs, the embedding of the protein may undergo significant changes as well. Of course, this is only a hypothetical scenario that needs to be tested experimentally while the large size of data is available. In the future, we will also attempt to use large language models for the extraction of protein representation. Moreover, current prediction of covalent binding affinity requires priori knowledge of the type of covalent bond for a ligand and target pair, which may be a limitation for broad and proper application, especially not friendly for non-chemists. These aspects warrant further investigation and consideration in future research endeavors.

Acknowledgements

We are grateful for the support from HPC Platform of ShanghaiTech University.

Conflict of interest

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China [82003654, 82341093], the National Key R&D Program of China [2022YFC3400501, 2022YFC3400500], Shanghai Science and Technology Development Funds [20QA1406400, 22ZR1441400], Lingang Laboratory [LG202102-01-03], start-up package from ShanghaiTech University, and Shanghai Frontiers Science Center for Biomacromolecules and Precision Medicine at ShanghaiTech University.

Data availability

The codes and datasets used in this article are available on GitHub (https://github.com/lizongquan01/TEFDTA).

References

Abbasi K, Razzaghi P, Poso A *et al.* DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics* 2020;**36**:4633–42.

Bitencourt-Ferreira G, de Azevedo WF. Molegro virtual docker for docking. *Methods Mol Biol* 2019;**2053**:149–67.

Chen L, Fan Z, Chang J *et al.* Sequence-based drug design as a concept in computational drug design. *Nat Commun* 2023;**14**:4217.

Chen L, Tan X, Wang D *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 2020;**36**:4406–14.

Davis MI, Hunt JP, Herrgard S *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51.

Du H, Gao J, Weng G *et al.* CovalentInDB: a comprehensive database facilitating the discovery of covalent inhibitors. *Nucleic Acids Res* 2021;**49**:D1122–9.

Friesner RA, Banks JL, Murphy RB *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 2004;**47**:1739–49.

Gainza P, Sverrisson F, Monti F *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92.

Hassan-Harrirou H, Zhang C, Lemmin T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J Chem Inf Model* 2020;**60**:2791–802.

He T, Heidemeyer M, Ban F *et al.* SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J Cheminform* 2017;**9**:24.

Hu Z, Liu W, Zhang C *et al.* SAM-DTA: a sequence-agnostic model for drug–target binding affinity prediction. *Brief Bioinform* 2023;**24**: bbac533.

Huang K *et al.* MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* 2021;**37**:830–6.

Jiménez J, Škalič M, Martínez-Rosell G *et al.* KDEEP: protein–ligand absolute binding affinity pre diction via 3D-convolutional neural networks. *J Chem Inform Model* 2018;**58**:287–96.

Lin X. DeepGS: Deep Representation Learning of Graphs and Sequences for Drug–Target Binding Affinity Prediction. In: *24th European Conference on Artificial Intelligence (ECAI), Santiago de Compostela, Spain. European Assoc Artificial Intelligence*, ELECTR NETWORK, 2020, 1301–8.

Liu T, Lin Y, Wen X *et al.* BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.

Nguyen T, Le H, Quinn TP *et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 2021; **37**:1140–7.

Obst-Sander U, Ricci A, Kuhn B *et al.* Discovery of novel allosteric EGFR L858R inhibitors for the treatment of non-small-cell lung cancer as a single agent or in combination with osimertinib. *J Med Chem* 2022;**65**:13052–73.

Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 2018;**34**:i821–i829.

Pahikkala T, Airola A, Pietilä S *et al.* Toward more realistic drug-target interaction predictions. *Brief Bioinform* 2015;**16**:325–37.

Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 2018;**34**:3666–74.

Tang J, Szwajda A, Shakyawar S *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inform Model* 2014;**54**:735–43.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* 2017;**30**: 6000–10.

Wang Y, Wu S, Duan Y *et al.* A point cloud-based deep learning strategy for protein–ligand binding affinity prediction. *Brief Bioinform* 2022; **23**:bbab474.

Yang Z, Zhong W, Zhao L *et al.* MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chem Sci* 2022;**13**:816–33.

Zheng L, Fan J, Mu Y. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega* 2019;**4**:15956–65.

Zhu Y, Zhao L, Wen N *et al.* DataDTA: a multi-feature and dual-interaction aggregation framework for drug–target binding affinity prediction. *Bioinformatics* 2023;**39**:btad560.