ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

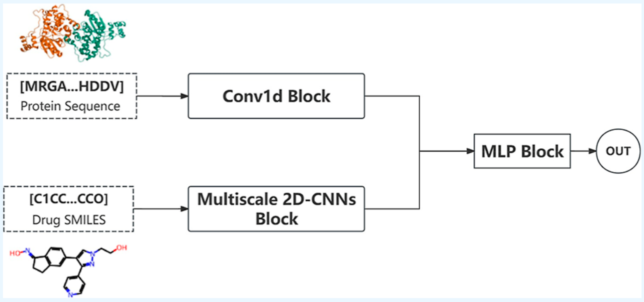
ImageDTA: A Simple Model for Drug–Target Binding Affinity Prediction

Li Han, Ling Kang, and Quan Guo*

Cite This: *ACS Omega* 2024, 9, 28485–28493

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Predicting the drug–target binding affinity (DTA) is crucial in drug discovery, and an increasing number of researchers are using artificial intelligence techniques to make such predictions. Many effective deep neural network prediction models have been proposed. However, current methods need improvement in accuracy, complexity, and efficiency. In this study, we propose a method based on a multiscale 2-dimensional convolutional neural network (CNN), namely ImageDTA. Many studies have shown that CNN achieves good learning effects with limited data. Therefore, we take a unique perspective by treating the word vector encoded with a simplified molecular input line entry system (SMILES) string as an "image" and processing it like handling images, fully leveraging the efficient processing capabilities of CNN for image data. Furthermore, we show that ImageDTA has higher training and inference efficiency than pretrained large models and outperforms attention-based graph neural network models in accuracy and interpretability. We also use visualization techniques to select appropriate convolutional kernel sizes, thereby increasing the network's interpretability.

1. INTRODUCTION

The process of discovering drugs through traditional experiments is both time-consuming and expensive.[1,2] Deep learning has been used to develop models predicting drug–target binding affinity (DTA). One such model is DeepDTA,[3] which uses convolutional neural networks (CNN) to learn drug–target feature representations from protein sequences and simplified molecular input line entry system (SMILES) strings representing drug molecules. WideDTA[4] is an extension of the DeepDTA method and uses additional features such as protein domains, motif information, and maximum common substructures (MCS) of molecules. However, these methods using a 1-dimensional CNN (1D-CNN) and pooling operations can lead to information loss, making it difficult to further improve the accuracy of predictions.

GraphDTA[5] is a model proposed by Nguyen et al. that encodes drugs as undirected graphs with a feature map and an adjacent matrix. It uses graph convolutional networks,[6] graph attention networks,[7] and graph isomorphic networks[8] to learn features from drug molecular strings. GraphCL-DTA[9] introduced a graph contrastive learning method that preserves the semantic information on molecular graphs. GraphscoreDTA[10] developed a novel graph neural network strategy that combines Vina distance optimization to predict DTA. However, the complex graph structures of these models make their interpretability a barrier for domain experts to adopt.[11] With unique ethical and regulatory requirements, the biomedical field's demand for interpretable deep models continues to grow.

Some other studies utilize large, pretrained models to learn features of protein sequences[12] and then use deep neural network models to predict DTA. Furthermore, this method relies on additional pretraining and fine-tuning stages for efficient protein sequence encoding, leading to low training and inference efficiency.

Some research has been conducted using attention-based models, which have also contributed to the prediction of DTA. AttentionDTA[13] performs an attention-like process on DeepDTA between the convolutional features of the drug and the target. DrugBAN[14] is a domain-adaptive deep bilinear attention network for drug–target binding affinity prediction based on molecular graphs and protein sequences. DrugVQA[15] proposes a question-answering model for drug–target interaction tasks, utilizing sequential attention mechanisms to capture the dependency relationships of dynamic CNN. Fang et al. proposed ColdDTA,[16] which uses data augmentation and attention-based feature fusion to improve the generalization ability to predict DTA. Although attention-based models can effectively build long-range dependency relationships, the

Received: March 12, 2024
Revised: June 12, 2024
Accepted: June 13, 2024
Published: June 24, 2024

relatively limited, well-annotated medical image data makes it difficult for such models to extract diverse global features, leading to attention collapse.[17]

DeepGS[18] encodes SMILES strings into a 100 × 100 matrix using Smi2Vec and then utilizes a 23 × 23 convolutional kernel for feature extraction. However, the use of a fixed convolutional kernel size results in cutting SMILES characters, which can damage their specific semantic information. Additionally, the fixed kernel size cannot effectively extract the structural information on different substructures (MCS) in drug molecules.

In this paper, we present a novel 2D-CNN model based on multiscale large convolutional kernels called ImageDTA. We treat protein sequences and drug SMILES strings as text containing biological language and perform word vector encoding. This encoding form is simple, easy to understand, and maximally preserves semantic information. We take a unique perspective of treating the word vector encoding small molecules of the drug as an "image" and use a multiscale 2D-CNN to perform feature learning on the "image", fully utilizing the ability of CNN to efficiently process image data. In situations with limited data, CNN models can achieve better learning effects than attention mechanism models.[17] In our model, we use an $h \times w$ convolutional kernel, where w is the dimension of the word vector encoding, which replaces the commonly used pooling operation in CNN, and the advantage is to avoid the loss of semantic information caused by pooling operations.

Some studies have shown that drugs are represented by the most common substructures, known as the ligand MCS.[4,19] Based on the research findings related to drug molecules, we have chosen convolutional kernel sizes. Additionally, we utilized visualization techniques to experimentally compare the impact of different convolutional kernel quantities and sizes, thereby demonstrating the superior interpretability of our model. To evaluate the effectiveness of ImageDTA's performance, we conducted comparison experiments on the Davis[20] and KIBA[21] data sets. We compared ImageDTA with state-of-the-art methods in terms of concordance index (CI)[22] and mean squared error (MSE). The results showed that ImageDTA achieves comparable or higher prediction accuracy, training, and inference efficiency than pretrained large models and better prediction accuracy and interpretability than models based on attention and graph neural networks.

Our paper's main contributions are as follows:

- We used multiple single-layer multiscale 2D-CNNs horizontally instead of stacking networks vertically, which significantly increased the interpretability of the network. This resulted in a competitive or even better performance than state-of-the-art on the public data sets Davis and KIBA.
- We treated the word vector-encoded SMILES sequences as "images" and processed them like handling images, which provided a unique perspective.
- We replaced traditional pooling operations in CNN with superlarge convolutional kernels, which preserved more semantic information.

2. MATERIAL AND METHODS

2.1. Data Sets. To evaluate the predictive performance of the proposed model, we utilized two public data sets for DTA prediction: Davis and KIBA. These data sets address the issue of data heterogeneity, and their specific statistical details are presented in Table 1 below.

Table 1. Statistical Analysis of Benchmark Datasets

data set	no. of proteins	no. of compounds	no. of interactions
Davis	442	68	30,056
KIBA	229	2111	118,254

2.1.1. Davis. The data set comprised 442 kinase proteins and 68 inhibitory drug small molecules, resulting in 30,056 binding affinity pairs. Most protein lengths were concentrated between 400 and 1500, with a peak distribution around 500 and a maximum length of 2549. The lengths of drug SMILES followed a Gaussian distribution, with most falling between 40 and 60 and a maximum length of 103. The strength of the interaction between target and drug molecules was calculated based on the logarithm of the kinase dissociation constant K_d; we transform K_d value to pK_d, as shown in the following eq 1.

$$\mathrm{p}K_\mathrm{d} = -\log_{10}\left(\frac{K_\mathrm{d}}{1 \times 10^{9}}\right) \tag{1}$$

2.1.2. KIBA. The data set used in this study was obtained through the KIBA method, which processed and screened 246,088 binding pairs made by 467 proteins and 52,498 ligands. Ultimately, 229 unique proteins, 2111 unique small-molecule drugs, and 118,254 complex macromolecules were identified. Protein sequence lengths were between 200 and 1500, with most around 700 and a maximum length of 4128. The lengths of the SMILES for drug molecules were mostly around 50, with a maximum length of 590.

According to Öztürk et al.,[3] 99% of protein pairs in the KIBA data set had a Smith−Waterman similarity of at most 60%, while 92% of protein pairs in the Davis data set had a target similarity of at most 60%, indicating that both data sets are nonredundant.

2.2. Model Architecture. The overall architecture of ImageDTA is shown in Figure 1. First, the amino acid sequence of proteins and the SMILES string of drugs were used as input. The drug molecule SMILES strings and protein sequences were input into the embedding layer. The drug molecules and proteins were encoded into 128-dimensional word vectors in this layer. Then, nine 2D-CNNs were extracted from the drug molecules, fusing these features with the First Concatenation layer. A combination network consisting of a three-layer 1D-CNN and a max-pooling layer extracted features from the protein sequences. To capture the local and global dependencies of the feature vectors, we applied a two-layer bidirectional long−short-term memory (BiLSTM) on the feature map from the Second Concatenation layer. Finally, after fusing the drug molecule features, protein features, and the output of the BiLSTM, they were input into the fully connected layer for prediction.

In our model, we used MSE as the loss function and Adam as the optimizer. The activation function for the fully connected layer was the rectified linear unit (ReLU). The ReLU and MSE are detailed in eqs 2 and 3:

$$g(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{3}$$

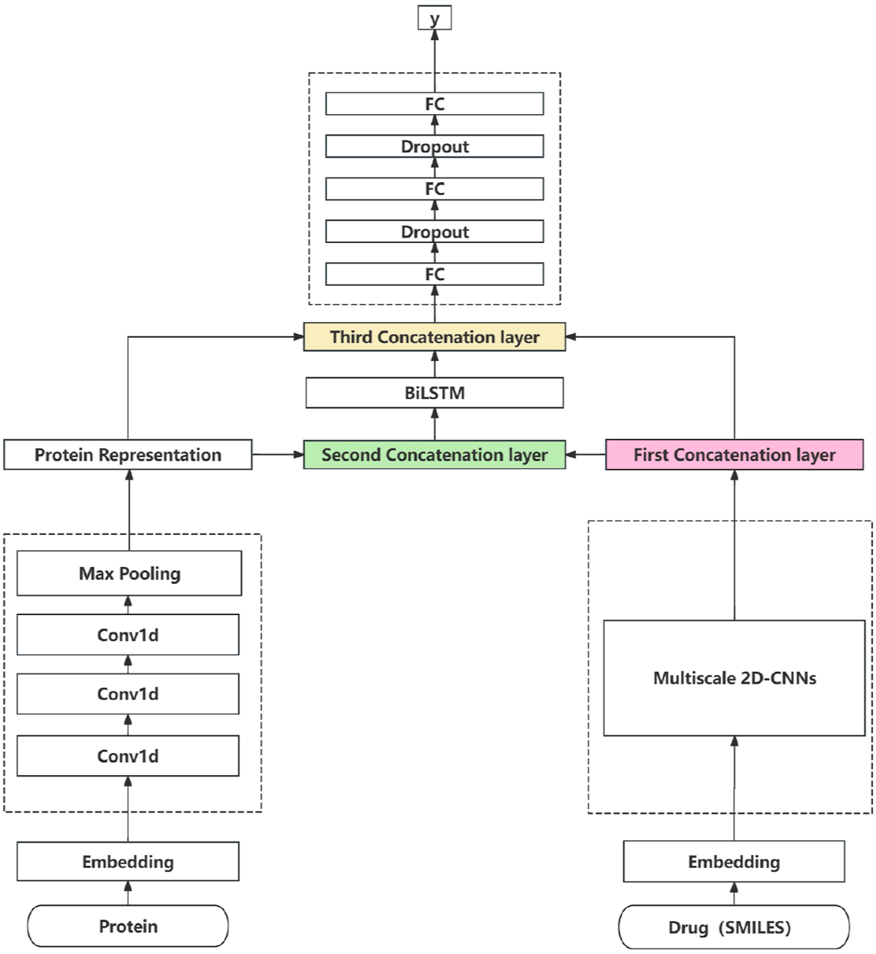


Figure 1. Architecture of the ImageDTA for drug–target binding affinity (DTA) prediction.

where n is the number of samples, y_i is the predicted value and $\hat{y}_i$ is the ground truth.

2.3. Drug and Protein Representation. The invention of SMILES was intended to help computers read molecular structures. It is a chemical notation that allows for efficient applications, including rapid retrieval and substructure searches. Similarly, protein sequences are encoded using label encodings. However, since drug molecules and protein sequences have varying lengths, network models often truncate or pad sequences, which may result in the loss of feature information or the addition of noise. For the Davis and KIBA data sets, we selected protein sequences and drug molecules of SMILES string lengths of 1000 and 100, respectively. If a protein sequence exceeds 1000 or a drug molecule SMILES string is over the set lengths, truncation is performed; otherwise, zero-padding is used to complete the sequences. Then, we utilized the Torch Embedding layer to represent characters with 128-dimensional word vectors. After converting the drug molecule SMILES strings into word vectors, we viewed them as “images” of the drug molecules, containing semantic and structural information, allowing us to use 2D-CNN for feature learning.

Because the length of the protein sequence is 1000, to improve the model’s efficiency, we used a method similar to that in DeepDTA, which involves using a three-layer 1D-CNN and a max-pooling layer for processing. The final features of the max-pooling layers were concatenated with the drug’s representations and fed into three FC layers.

2.4. Multiscale 2D-CNN Layer. We used the SMILES strings of drugs as input. In previous studies, 1D-CNN was used to extract structural features from drug strings. However, it is difficult to capture global feature information at different scales. Therefore, we used a set of multiscale 2D-CNN, which allowed the model to obtain richer local and global features from the SMILES strings. In the 2D-CNN, we used large convolutional kernels to replace the pooling operation, reducing feature loss. This part of the model consists of nine 2D-CNNs with different scale kernels arranged horizontally, as shown in Figure 2.

First, to obtain more semantic information about SMILES strings, we converted the input strings into a word vector matrix of size $length \times em\_dim$, where $em\_dim = 128$ is the word vector encoding dimension and length = 100 is the length of SMILES strings. In order to more accurately capture “chemical words” and enhance the model’s local sensitivity and global difference judgment ability, we introduced nine 2D-CNNs with different convolutional kernel sizes. Table 2 shows specific convolutional kernel information. The design of the size and number of convolutional kernels were based on a study (Woźniak et al., 2018), which compared about 2K molecules in pairs and showed that the patterns used by chemists to distinguish a set of molecules, such as the maximum common substructure (MCS), were the “chemical words”. The substructure size represented the characteristics of small drug molecules; the number of characters in the substructure (MCS) of drug molecules was between 1 and 29, most of which were 8–12 characters.

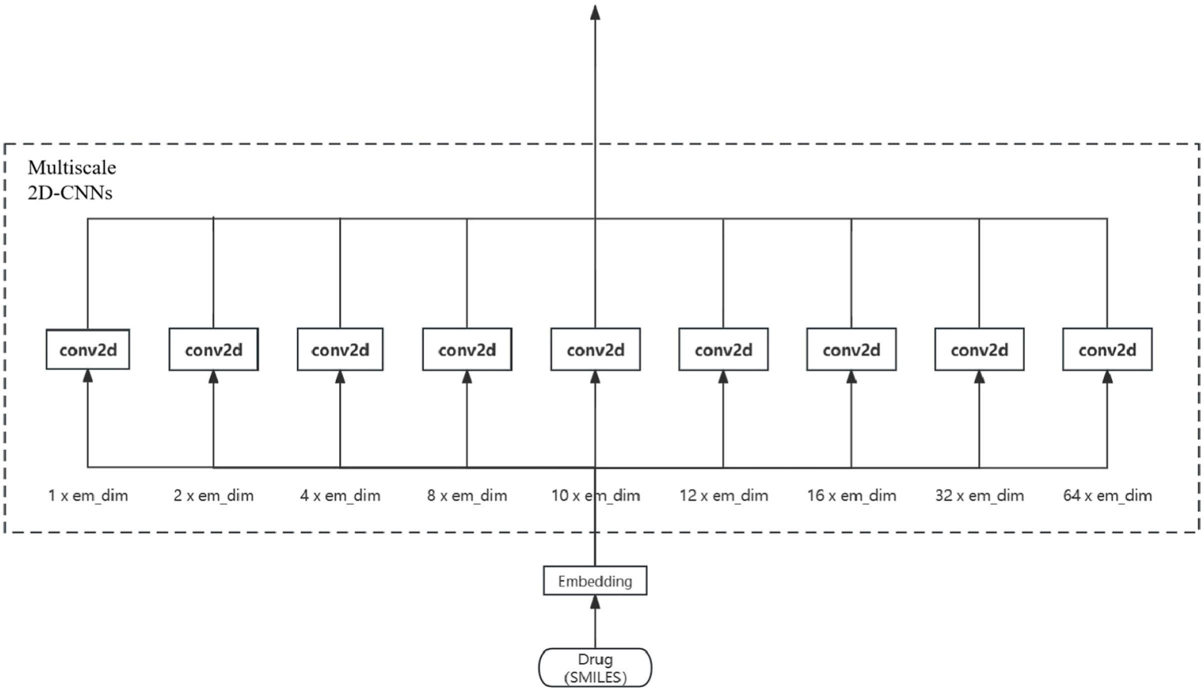


Figure 2. Framework of multiscale 2D-CNNs (em_dim = 128).

We developed a horizontally distributed CNN, allowing each convolutional layer to extract features from drug molecules within a specific receptive field (determined by the convolutional kernel size). To accommodate different sizes of the MSC, we designed nine horizontally distributed CNNs with convolutional kernels of varying sizes, ensuring that the model captures the features of drug molecules with different MSC sizes. This design approach was intended to address the limitations of traditional CNNs, such as ResNet, composed of multiple layers stacked deeply, making it difficult to control the receptive field of the CNN and also challenging to theoretically explain the use of convolutional kernels of different sizes to extract MSCs of varying sizes. Especially in the field of drug development, an unexplainable complex network cannot truly gain the trust and understanding of drug developers, and it is difficult to use in actual drug discovery work.[11] Therefore, we adopted a simpler horizontally distributed CNN structure model, reducing the level of the network and further increasing the interpretability of the model.

2.5. Concatenation Layer.

First Concatenation layer: The drug molecule features extracted by 2D-CNNs at different scales were merged. Assuming that the feature vector of the drug is $F^i = \{f^i_1, f^i_2, ..., f^i_n\}$, where $i \in [1,9]$, the feature fusion of the nine convolutional layers is

$$H_{\text{first}} = \text{concat}(F^1, ..., F^9) = \{f^1_1, f^1_2, ..., f^1_n, f^2_1, ..., f^2_{\text{t}}, ..., f^9_1, f^9_2, ..., f^9_m\} \quad (4)$$

where $n, t, m \in N^+$.

Second Concatenation layer: Similar to the algorithm of the First Concatenation layer, we fused the output of the drug string feature fusion layer with protein features as the input for the BiLSTM layer.

Assuming the protein feature vector is $P = \{p_1, p_2, \cdots, p_n\}$, the output of the fusion layer is.

$$H_{\text{second}} = \text{concat}(H_{\text{first}}, P) \quad (5)$$

Third Concatenation layer: The outputs of the First Concatenation layer, Protein Representation, and the Second Concatenation layer were fused as the input for the fully connected layer.

$$H_{\text{third}} = \text{concat}(H_{\text{first}}, P, H_{\text{second}}) \quad (6)$$

Table 2. Informations of Convolutional Kernel Sizes[a]

input	operator
length × *em_dim*	Conv2d, 1 × *em_dim*
length × *em_dim*	Conv2d, 2 × *em_dim*
length × *em_dim*	Conv2d, 4 × *em_dim*
length × *em_dim*	Conv2d, 8 × *em_dim*
length × *em_dim*	Conv2d, 10 × *em_dim*
length × *em_dim*	Conv2d, 12 × *em_dim*
length × *em_dim*	Conv2d, 16 × *em_dim*
length × *em_dim*	Conv2d, 32 × *em_dim*
length × *em_dim*	Conv2d, 64 × *em_dim*

[a] Length = 100, em_dim = 128.

3. RESULTS AND DISCUSSION

3.1. Evaluation Metrics.

We used CI[22] to evaluate the performance of ImageDTA.

Table 3. Hyperparameters of ImageDTA

parameters	range
batch size	512
embedding dimensional	128
dropout	0.1
optimizer	Adam
learning rate	0.0005

Table 4. Performance Comparison of Different Models on Davis and KIBA

	CI		MSE	
method	Davis	KIBA	Davis	KIBA
DeepDTA	0.878	0.863	0.261	0.194
DeepGS	0.882	0.863	0.252	0.194
AttentionDTA	0.885	0.861	0.241	0.174
WideDTA	0.886	0.875	0.262	0.179
DeepCDA	0.891	0.889	0.248	0.176
GraphDTA	0.893	0.891	0.229	0.139
FingerDTA	0.895	0.885	0.234	0.150
FusionDTA	0.913	0.906	0.208	0.130
TEFDTA	0.890	0.860	0.199	0.184
ImageDTA	0.901	0.886	0.215	0.147

Table 5. Comparing the ImageDTA against the Alternative Methods, i.e., GraphDTA and FusionDTA, in Terms of the Training and Inference Times on Graphics Processing Units (s/epoch)

model	runtime of training for Davis data set (s/epoch)		runtime of inference for KIBA data set (s/epoch)	
	Davis	KIBA	Davis	KIBA
GraphDTA	6	45	2	9
FusionDTA	335	1255	53	212
ImageDTA	87	300	7	26

$$\mathrm{CI} = \frac{1}{Z}\sum_{\delta_j > \delta_i} h(b_i - b_j) \tag{7}$$

where b_i is the prediction value for δ_i, b_j is the prediction value for δ_j, $h(x)$ is the step function, and Z is the normalized hyperparameter. Commonly, the step function $h(x)$ is defined as follows

$$h(x) = \begin{cases} 0.0, \ x < 0 \\ 0.5, \ x = 0 \\ 1.0, \ x > 0 \end{cases} \tag{8}$$

MSE is a statistical measure that evaluates the error directly. See eq 3.

3.2. Comparison of the Prediction Efficiency.

We proposed ImageDTA model to learn features for drugs and targets based on their word embedding. In this section, the Davis and KIBA data sets were utilized to evaluate the performance of the model. In ImageDTA, the hyperparameters used in these two data sets are shown in Table 3.

We compared our model with the following benchmark models: DeepDTA, DeepGS, WideDTA, GraphDTA, FusionDTA, AttentionDTA, DeepCDA,[23] FingerDTA,[24] and TEFDTA.[25]

In Table 4, we listed the performance of all the aforementioned models evaluated on the Davis and KIBA data sets. As shown, ImageDTA outperforms most of the models. Specifically, compared to the baseline model, ImageDTA improves the CI index by 0.006 and reduces MSE by 0.014 on the Davis data set, except for FusionDTA; on the KIBA data set, our ImageDTA can achieve competitive or even better performance in terms of CI and MSE compared to the aforementioned baseline models.

From the training and inference efficiency perspective, we compared our model with GraphDTA and FusionDTA using a graphics processing unit (RTX5000 16G). The results are shown in Table 5.

As can be seen from the above table, the ImageDTA model is faster in training and inference time. Although our model was slightly slower in inference and training time compared to GraphDTA, it performed better in terms of predictive ability, especially on the Davis data set.

3.3. Performance Comparison of the Predicted and Real Values.

In this section, we compared the predicted and real values for the Davis and KIBA data sets. As shown in Figure 3, the results confirm that ImageDTA predicts DTAs very close to the real values for the Davis and KIBA data sets.

3.4. Model Analysis.

In order to select the most suitable convolution kernel size for extracting drug features, we used the heat map visualization technique Grad-CAM[26] of the CNN to understand the decision-making process of the deep learning model in image classification tasks. Grad-CAM generated a heat map by calculating the gradient of each feature map of the CNN, showing image areas crucial for the model to make correct decisions. The color of the Grad-CAM heat map ranged from purple to yellow; the closer the color on the heatmap is to yellow, the more attention the model devotes to that region.

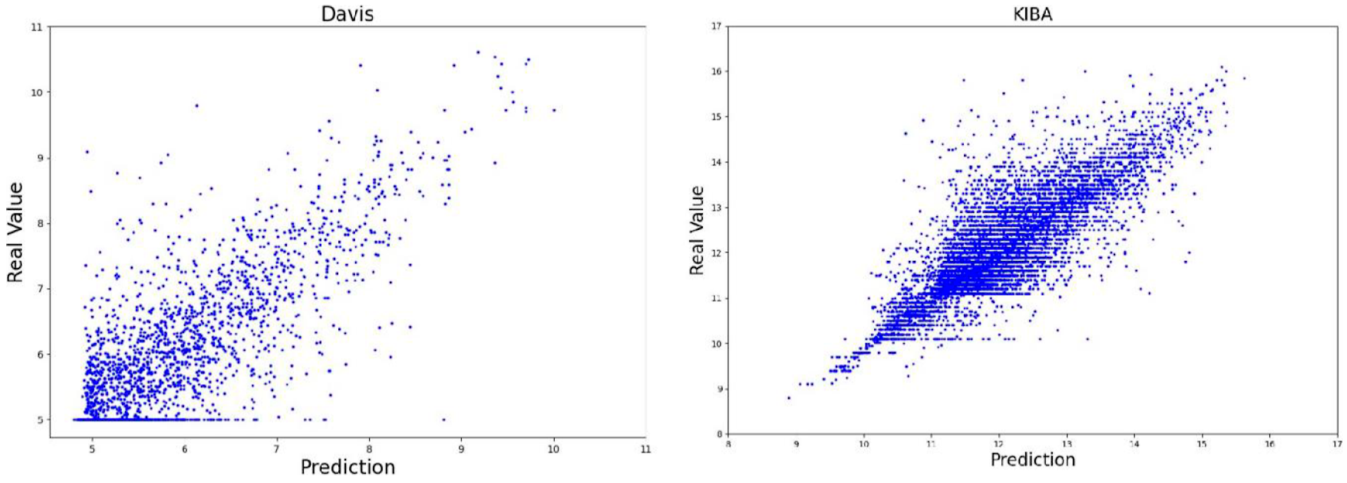


Figure 3. Comparison of the correlation between the predicted and real values for Davis and KIBA Data.

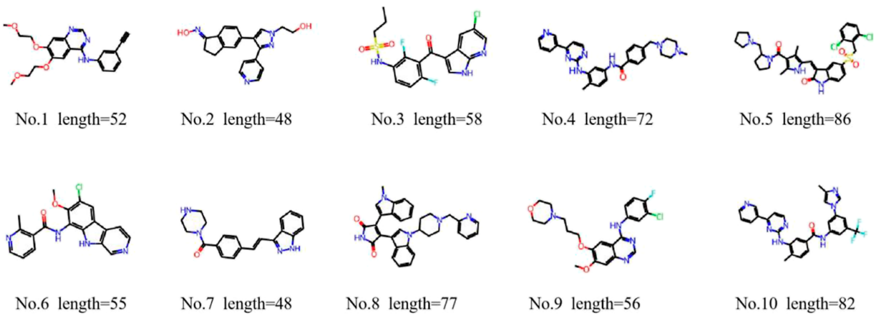


Figure 4. Drug molecules (length = SMILES string's size).

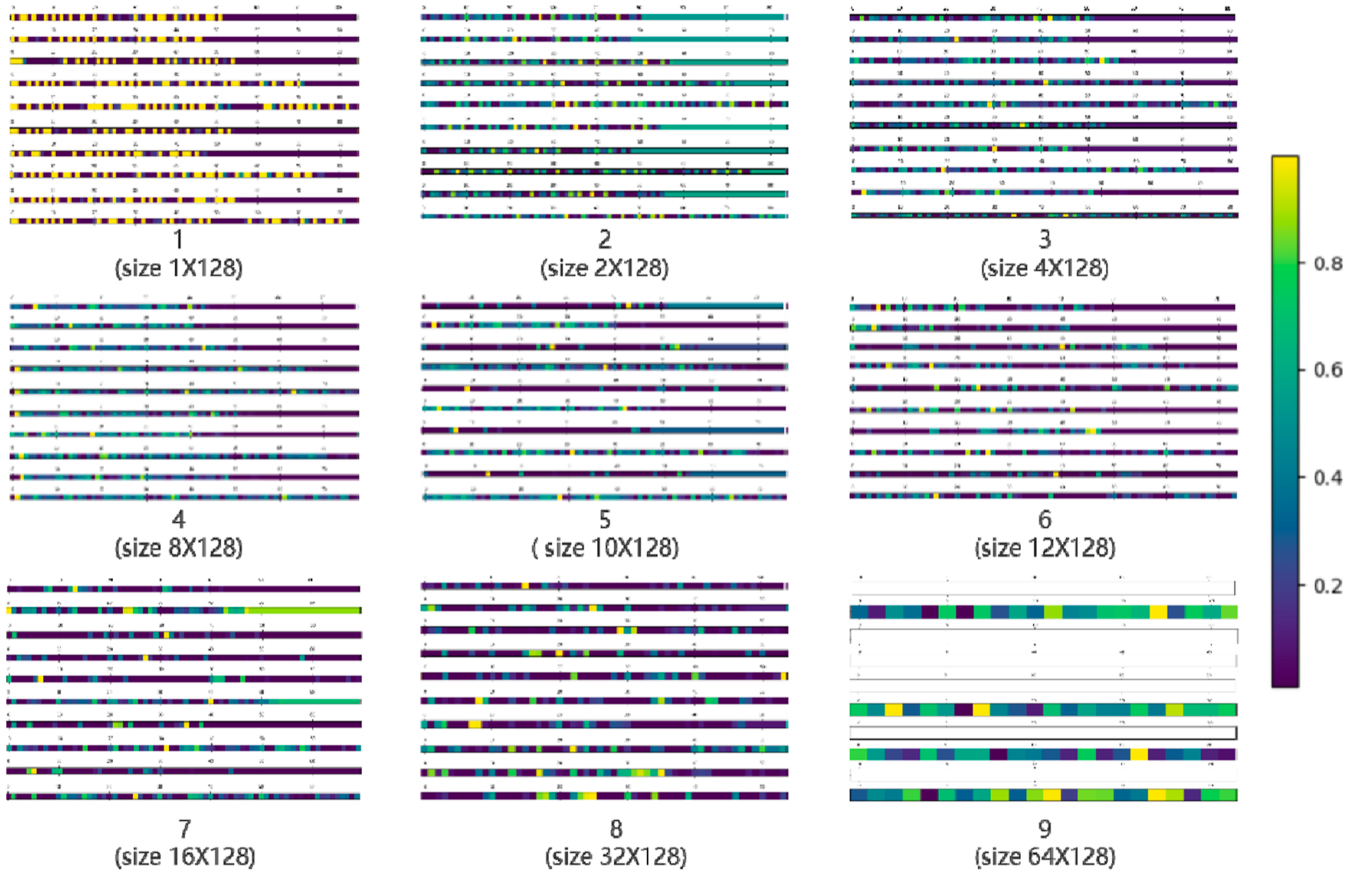


Figure 5. Feature heat maps obtained from 10 drug molecules under 9 different sizes of convolution kernels. In each heatmap, the arrangement order of the drug molecules from top to bottom is no. 1−10 (Figure 4).

Table 6. Drug SMILES String and Target Sequence

type	string/sequence	length
drug	C1CC(=NO)C2=C1C=C(C=C2)C3=CN(N=C3C4=CC=NC=C4)CCO	48
protein	MRGARGAWDFLCVLLLLLRVQTGSSQPSVSPGEPSPPSIHPGKSDLIVRVGDEIRLLCTDPGFVKWTFEILDETNENKQNEWITEKAEATNTGKYTCTNKHGLSNSIYVFVRDPAKLFLVDRSLYGKEDNDTLVRCPLTDPEVTNYSLKGCQGKPLPKDLRFIPDPKAGIMIKSVKRAYHRLCLHCSVDQEGKSVLSEKFILKVRPAFKAVPVVSVSKASYLLREGEEFTVTCTIKDVSSSVYSTWKRENSQTKLQEKYNSWHHGDFNYERQATLTISSARVNDSGVFMCYANNTFGSANVTTTLEVVDKGFINIFPMINTTVFVNDGENVDLIVEYEAFPKPEHQQWIYMNRTFTDKWEDYPKSENESNIRYVSELHLTRLKGTEGGTYTFLVSNSDVNAAIAFNVYVNTKPEILTYDRLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSASVLPVDVQTLNSSGPPFGKLVVQSSIDSSAFKHNGTVECKAYNDVGKTSAYFNFAFKGNNKEQIHPHTLFTPLLIGFVIVAGMMCIIVMILTYKYLQKPMYEVQWKVVEEINGNNYVYIDPTQLPYDHKWEFPRNRLSFGKTLGAGAFGKVVEATAYGLIKSDAAMTVAVKMLKPSAHLTEREALMSELKVLSYLGNHMNIVNLLGACTIGGPTLVITEYCCYGDLLNFLRRKRDSFICSKQEDHAEAALYKNLLHSKESSCSDSTNEYMDMKPGVSYVVPTKADKRRSVRIGSYIERDVTPAIMEDDELALDLEDLLSFSYQVAKGMAFLASKNCIHRDLAARNILLTHGRITKICDFGLARDIKNDSNYVVKGNARLPVKWMAPESIFNCVYTFESDVWSYGIFLWELFSLGSSPYPGMPVDSKFYKMIKEGFRMLSPEHAPAEMYDIMKTCWDADPLKRPTFKQIVQLIEKQISESTNHIYSNLANCSPNRQKPVVDHSVRINSVGSTASSSQPLLVHDDV	976

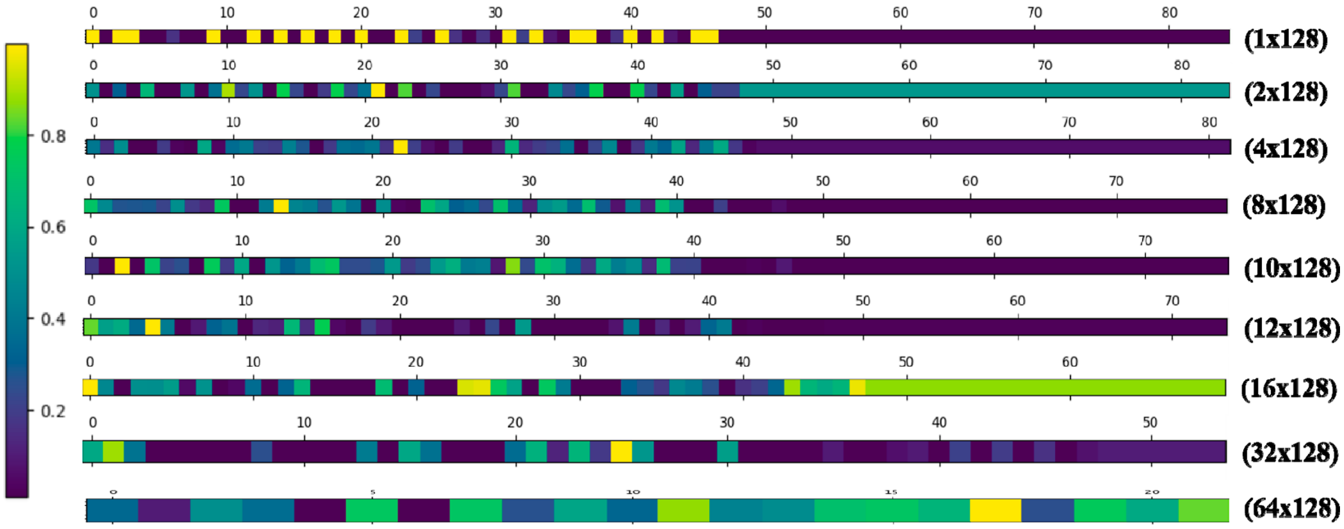


Figure 6. Heatmap of the feature map extracted by multiscale 2D-CNNs from the drug SMILES string.

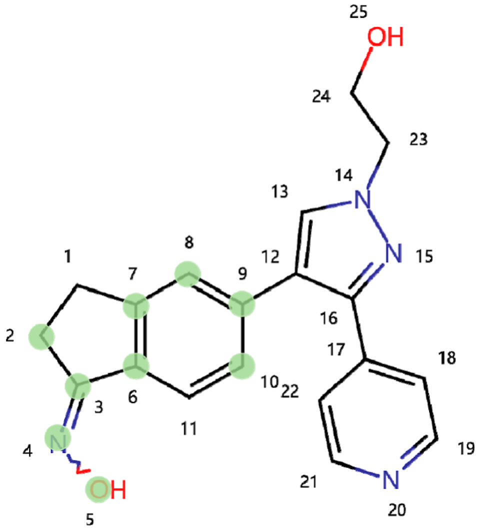


Figure 7. Molecular graph corresponds to the SMILES string for the drug, where the numbers in the graph represent the positional indices of the atoms in the string. The light green circles highlight the drug atoms that the three (8 × 128), (10 × 128), and (12 × 128) sized convolutional kernels focus on.

We randomly selected 10 drug molecules from the Davis data set, as shown in Figure 4. Additionally, we used the Grad-CAM visualization technique to present the heat maps of each SMILES string under different sizes of convolution kernels, as shown in Figure 5. The feature heat maps of the CNN were used to directly view the contributions of different sizes of convolution kernels to the predicted features.

In the analysis of drug molecule sequences, Figure 5 illustrates the focal points of different convolutional kernels when processing sequences. The closer the color on the heatmap is to yellow, the more attention the model devotes to that region. Specifically, a (1 × 128) convolutional kernel was depicted as a smaller highlighted area on the heatmap, indicating that the network devotes more attention to these local regions. Conversely, larger convolutional kernels appeared as larger highlighted areas on the heatmap, signifying that the network is capturing broader sequence features. By analyzing the heatmap, we gained a better understanding of how the model processed information through convolutional kernels of varying sizes. This understanding can aid in designing more effective network architectures, such as deciding which parts of the network to employ convolutional kernels with a local receptive field and which parts to use a wider receptive field to capture global information.

In Figure 5, the ninth convolution kernel contains a significant amount of white space. This may be attributed to the following reasons: First, when the molecular sequence length provided was <100, substantial padding of zeros was applied to compensate for the deficiency. When extracting these features using the large (64 × 128) convolutional kernel, an excessive amount of zero values impacted the model's learning ability, leading to an inability to extract sufficiently accurate features. Additionally, this white space indicated that certain structural drug molecules were unsuitable for feature extraction using an oversized convolutional kernel. These observations prompted us to prioritize them as a focal point for our next research endeavors.

We selected a drug−target pair as shown in Table 6. The heatmap of the feature map extracted by multiscale 2D-CNNs from the drug SMILES string is shown in Figure 6.

From Figure 6, it can be seen that all nine convolutional kernels can ignore the padded zeros and effectively extract the local relationships of the drug molecules we are interested in. The three sizes of convolutional kernels (8 × 128), (10 × 128), and (12 × 128) were designed based on the MCS. The heatmap generated by the three convolutional kernels above indicates that they predominantly focus on characters 3 to 20 in the SMILES string, corresponding to the atoms of the drug, as highlighted by the light green circles in Figure 7, can be considered potential binding sites with the protein shown in Table 6.

4. CONCLUSIONS

We proposed a novel model based on multiscale 2D-CNNs to predict DTA using protein sequences and drug SMILES strings and learn drug features. Additionally, we used three layers of 1D-CNN to learn features from the protein sequences, along with three fully connected layers in the affinity prediction task. We conducted our experiments on the Davis and KIBA data sets. The results showed that ImageDTA significantly improved performance compared to baseline methodologies for the Davis and KIBA data sets. Furthermore, our model, made to learn drug features from SMILES strings, outperformed methods based on pretrained large models, attention mechanisms, and graph neural network structures in terms of computational efficiency, accuracy, and interpretability. The major contribution of this study is in the following three aspects: First, by using multiple single-layer multiscale 2D-CNNs horizontally instead of stacking networks vertically, the interpretability of the network was significantly increased, resulting in a competitive or even better performance compared with the state-of-the-art on the public data sets of Davis and KIBA. Second, we took a unique perspective by treating the word vector encoded with SMILES strings as an "image" and processing them like handling an image. Third, we replaced traditional pooling operations in CNN with superlarge convolutional kernels, preserving more semantic information. Furthermore, in the future, we will focus on developing a method based on multiscale 2D-CNN to learn efficient representations of protein sequences.

■ ASSOCIATED CONTENT

Data Availability Statement

The codes and data sets used in this article are available on GitHub (https://github.com/neuhanli/ImageDTA).

■ AUTHOR INFORMATION

Corresponding Author

Quan Guo − *Neusoft Research Institute, Dalian Neusoft University of Information, Dalian, Liaoning 116023, China;* orcid.org/0009-0003-3598-3422; Email: guoquan@neusoft.edu.cn

Authors

Li Han − *Software and Big Data Technology Department, Dalian Neusoft University of Information, Dalian, Liaoning 116023, China;* orcid.org/0009-0008-3430-5386

Ling Kang − *Neusoft Research Institute, Dalian Neusoft University of Information, Dalian, Liaoning 116023, China;* orcid.org/0009-0002-6654-9875

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.4c02308

Author Contributions

The manuscript was written with contributions from all authors, and all authors have approved the final version. **Li Han (first author):** conceptualization, methodology, software, investigation, formal analysis, and writing-original draft; **Ling Kang:** data curation, writing-original draft, resources, supervision, validation, and writing-review and editing; **Quan Guo (corresponding author):** conceptualization, funding acquisition, resources, supervision, and writing-review and editing.

Funding

This work was supported by the Dalian Science and Technology Innovation Fund Program [2022JJ12GX017], the Foundation of Liaoning Education Ministry [LJKZ1414], the United Foundation for Medicoengineering Cooperation from Dalian Neusoft University of Information and the Second Hospital of Dalian Medical University [LH-JSRZ-202201], and the Technology Innovation Project of Dalian Neusoft University of Information [TIFP202302].

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful for the support from the Neusoft Research Institute of Dalian Neusoft University of Information.

■ ABBREVIATIONS

DTA	drug−target binding affinity
CNN	convolutional neural networks
SMIL	simplified molecular input line entry system
MCS	maximum common substructures

■ REFERENCES

(1) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770−803.

(2) Takebe, T.; Imai, R.; Ono, S. The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. *Clin. Transl. Sci.* **2018**, *11* (6), 597−606.

(3) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug-Target Binding Affinity Prediction. *Bioinformatics* **2018**, *34* (17), i821−i829.

(4) Öztürk, H.; Ozkirimli, E.; Özgür, A. WideDTA: Prediction of Drug-Target Binding Affinity, 2019. arXiv:1902.04166. http://arxiv.org/abs/1902.04166 (accessed 2023-12-07).

(5) Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; Venkatesh, S. GraphDTA: Predicting Drug−Target Binding Affinity with Graph Neural Networks. *Bioinformatics* **2021**, *37* (8), 1140−1147.

(6) Kipf, T. N.; Welling, M. Semi-supervised Classification with Graph Convolutional Networks, 2017. arXiv:1609.02907. http://arxiv.org/abs/1609.02907 (accessed 2024-01-27).

(7) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks, 2018. arXiv:1710.10903. http://arxiv.org/abs/1710.10903 (accessed 2024-01-27).

(8) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks?, 2019. arXiv:1810.00826. http://arxiv.org/abs/1810.00826 (accessed 2024-01-27).

(9) Yang, X.; Yang, G.; Chu, J.; Graph, C. L. D. T. A. A Graph Contrastive Learning with Molecular Semantics for Drug-Target Binding Affinity Prediction, 2023. arXiv:2307.08989. http://arxiv.org/abs/2307.08989 (accessed 2023-12-07).

(10) Wang, K.; Zhou, R.; Tang, J.; Li, M. GraphscoreDTA: Optimized Graph Neural Network for Protein−Ligand Binding Affinity Prediction. *Bioinformatics* **2023**, *39* (6), btad340.

(11) Preuer, K.; Klambauer, G.; Rippmann, F.; Hochreiter, S.; Unterthiner, T. Interpretable Deep Learning in Drug Discovery. In *Interpretable Deep Learning in Drug Discovery. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K. R., Eds.; Springer International Publishing: Cham, 2019; Vol. *11700*, pp 331−345..

(12) Yuan, W.; Chen, G.; Chen, C. Y. C. FusionDTA: Attention-Based Feature Polymerizer and Knowledge Distillation for Drug-Target Binding Affinity Prediction. *Briefings Bioinf.* **2022**, *23* (1), bbab506.

(13) Zhao, Q.; Xiao, F.; Yang, M.; Li, Y.; Wang, J. AttentionDTA: Prediction of Drug−Target Binding Affinity Using Attention Model 2019. In *In IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE: San Diego, CA, USA, 2019; pp 64−69..

(14) Bai, P.; Miljković, F.; John, B.; Lu, H. Interpretable Bilinear Attention Network with Domain Adaptation Improves Drug-Target Prediction. *Nat. Mach. Intell.* **2023**, *5*, 126−136.

(15) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Publisher Correction: Predicting Drug−Protein Interaction Using Quasi-Visual Question Answering System. *Nat. Mach. Intell.* **2020**, *2* (9), 551.

(16) Fang, K.; Zhang, Y.; Du, S.; He, J. ColdDTA: Utilizing Data Augmentation and Attention-Based Feature Fusion for Drug-Target Binding Affinity Prediction. *Comput. Biol. Med.* **2023**, *164*, 107372.

(17) Lin, X.; Yan, Z.; Deng, X.; Zheng, C.; Yu, L. ConvFormer: Plug-and-Play CNN-Style Transformers for Improving Medical Image

Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023; pp 642, 651, .

(18) Lin, X. DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction, 2020. arXiv:2003.13902. http://arxiv.org/abs/2003.13902 (accessed 2023-12-08).

(19) Woźniak, M.; Wołos, A.; Modrzyk, U.; Górski, R. L.; Winkowski, J.; Bajczyk, M.; Szymkuć, S.; Grzybowski, B. A.; Eder, M. Linguistic Measures of Chemical Diversity and the "Keywords" of Molecular Collections. *Sci. Rep.* **2018**, *8* (1), 7598.

(20) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29* (11), 1046−1051.

(21) Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; Aittokallio, T. Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis. *J. Chem. Inf. Model.* **2014**, *54* (3), 735−743.

(22) Gönen, M.; Heller, G. Concordance Probability and Discriminatory Power in Proportional Hazards Regression. *Biometrika* **2005**, *92* (4), 965−970.

(23) Abbasi, K.; Razzaghi, P.; Poso, A.; Amanlou, M.; Ghasemi, J. B.; Masoudi-Nejad, A. DeepCDA: Deep Cross-Domain Compound−Protein Affinity Prediction through LSTM and Convolutional Neural Networks. *Bioinformatics* **2020**, *36* (17), 4633−4642.

(24) Zhu, X.; Liu, J.; Zhang, J.; Yang, Z.; Yang, F.; Zhang, X. FingerDTA: A Fingerprint-Embedding Framework for Drug-Target Binding Affinity Prediction. *Big Data Min. Anal.* **2023**, *6* (1), 1−10.

(25) Li, Z.; Ren, P.; Yang, H.; Zheng, J.; Bai, F. TEFDTA: A Transformer Encoder and Fingerprint Representation Combined Prediction Method for Bonded and Nonbonded Drug−Target Affinities. *Bioinformatics* **2024**, *40* (1), 1−8.

(26) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM.: Visual Explanations from Deep Networks via Gradient-Based Localization. *arXiv* **2016**, arXiv:1610.02391.