Front. Comput. Sci., 2023, 17(5): 175903
https://doi.org/10.1007/s11704-022-2163-9

RESEARCH ARTICLE

FragDPI: a novel drug-protein interaction prediction model based on fragment understanding and unified coding

Zhihui YANG, Juan LIU (✉), Xuekai ZHU, Feng YANG, Qiang ZHANG, Hayat Ali SHAH

Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan 430072, china

Abstract Prediction of drug-protein binding is critical for virtual drug screening. Many deep learning methods have been proposed to predict the drug-protein binding based on protein sequences and drug representation sequences. However, most existing methods extract features from protein and drug sequences separately. As a result, they can not learn the features characterizing the drug-protein interactions. In addition, the existing methods encode the protein (drug) sequence usually based on the assumption that each amino acid (atom) has the same contribution to the binding, ignoring different impacts of different amino acids (atoms) on the binding. However, the event of drug-protein binding usually occurs between conserved residue fragments in the protein sequence and atom fragments of the drug molecule. Therefore, a more comprehensive encoding strategy is required to extract information from the conserved fragments.

In this paper, we propose a novel model, named FragDPI, to predict the drug-protein binding affinity. Unlike other methods, we encode the sequences based on the conserved fragments and encode the protein and drug into a unified vector. Moreover, we adopt a novel two-step training strategy to train FragDPI. The pre-training step is to learn the interactions between different fragments using unsupervised learning. The fine-tuning step is for predicting the binding affinities using supervised learning. The experiment results have illustrated the superiority of FragDPI.

Keywords affinity score, drug-protein interaction, BERT, Bi-Transformer, virtual drug screening

1 Introduction

Now that the traditional drug discovery pipeline suffers from high cost, long cycle, and low success rate [1]. In the application of virtual drug screening, more and more pharmaceutical companies adopt virtual drug screening by developing computational methods to reduce cost, accelerate the process and improve the success rate of drug discovery.

Virtual drug screening searches for candidate drug molecules that can bind with target proteins by using computational methods, which guide subsequent wet experiments. There are two main kinds of virtual drug screening techniques, i.e., molecular docking [2] and drug-protein interaction (DPI) prediction. Molecular docking can visualize the drug-protein binding process and provide possible conformations of the drug molecules. However, such techniques require defining a scoring function to evaluate the drug-protein binding energy [3]. In contrast, defining a convincible function to characterize the binding energy is not easy. Moreover, only one pair of drugs and protein can be docked each time, resulting in an inefficient molecular docking technique for large-scale drug screening. As a result, molecular docking is not a widespread technique in large-scale virtual drug screening. Unlike molecular docking, the DPI prediction methods build a machine learning-based model to predict the drug-protein binding affinities. Such methods are data-driven and can automatically learn good models from known drug-protein binding data with limited prior knowledge. In addition, the DPI prediction methods can be used for large-scale drug screening purposes, which is very helpful for finding the most suitable drug candidates that can reduce unnecessary trials in the follow-up experiments. Therefore, we focus on the DPI prediction in this paper.

With the successful applications of deep learning in many fields, including bioinformatics, many deep learning-based DPI prediction methods have recently been proposed [4−7]. Generally, these end-to-end methods extract features directly from protein and drug sequences and predict drug-protein binding affinities accordingly. Therefore, encoding the protein (drug) sequence is critical for building the models. Most existing methods usually use a one-hot vector to represent an amino acid (atom) and then use the one-hot matrix to represent the protein (drug) sequence. Such strategy is implicitly based on the assumptions that the single residue (atom) is the functional unit for binding and every residue (atom) plays the same role in the drug-protein binding. However, in the biochemical process, the DPI usually occurs between conserved fragments of protein and drug sequences, rather than a single residue and atom [8]. Obviously, the one-hot matrix encoding method cannot reveal the details of the actual drug-protein binding. Furthermore, most of the existing

Received March 23, 2022; accepted July 11, 2022

E-mail: liujuan@whu.edu.cn

methods adopt the two-stream network module [5,6], where the features of drugs and proteins are extracted separately. Such feature extraction module fails to learn drug-protein interaction information, lacking a local investigation of the interactions between drugs and proteins.

In this paper, we propose a novel method, named FragDPI, for predicting drug-protein binding affinity. In order to model the biochemical process, we encode proteins and drugs by using conserved fragments instead of single amino acids or atoms. Moreover, we employ a unified encoding strategy to combine the fragments of drug and protein in order to extract the interaction features. We adopt a novel two-step training strategy to build the model. In the first sage, we focus on fragment understanding (FU) which enables the model to understand the relationship of fragments between drug and protein. For such purpose, we pre-train the model to obtain the unified coding for each drug-protein pair by unsupervised learning. In the second stage, we focus on predicting the binding affinities by fine-tuning the model with supervised learning. In order to evaluate our model, we conduct comparison experiments and an ablation study to illustrate its advantages and potential applications.

The highlights of this paper mainly includes the following three points:

We propose a deep learning model using the Bidirectional Transformers as the backbone, FragDPI, to predict the drug-protein binding affinity. Moreover, a unified coding strategy was introduced that tokenizes the sequence of a protein by identifying conserved fragments in the sequence. We encode drugs and proteins into a unified vector to more precisely describe the details of drug-protein interactions and simplify the encoding process.

We build the model through a novel two-step training strategy. For the pre-training stage, we propose FU to explore the potential information between the protein sequence fragments and the drug molecule fragments. For the fine-tuning stage, we use the pre-trained FragDPI to train in the specific datasets and then complete the affinity prediction of the drug-protein pair.

We conduct extensive performance evaluations with state-of-the-art methods on four datasets. The results show that FragDPI achieves the comparable performance in RMSE and Pearson's r. Compared with same-task methods, FragDPI can explore more binding features from conserved fragments between protein sequence and drug molecule string through the attention module of the model.

2 Related work

2.1 Drug screening

Previous works have been used to predict DPI through the biological experiment method and computer-assisted methods.

The principal method of the biological experiment is to analyze the structure of the target protein and fluorescent protein labeling. Researchers leverage cryo-electron microscopy and other biological techniques to investigate the structure of the target protein or use fluorescent protein labeling to track key proteins [9,10]. However, these methods have various limitations, such as experimental resources, reagents, and economic conditions. Therefore, virtual drug screening is more advantageous in this situation.

2.2 DPI prediction

The computer-assisted DPI prediction methods are divided into three categories according to their ideas: similarity between target proteins, twin-tower framework, and mixed coding for different information. Kernel regression [11] and matrix factorization [12] utilize off-the-shelf DPI pairs to infer the new pairs using the similarity of proteins. The twin-tower framework extracts the features from protein and drug separately, like DeepDTA, AttentionDTA and GraphDTA. DeepDTA [13] employs convolutional neural networks to extract features in sequences of proteins and drugs. AttentionDTA [6] adds attention modules to the previous framework. GraphDTA [14] leverages a graph convolutional network to extract features closer to the data structure of drug representation. As for mixed coding, DeepDock [7] utilizes structural information and sequence information together and predicts DPI with only a simple fully-connected model.

In general, the above methods encode the protein and drug sequences usually based on the assumption that each amino acid and atom has the same contribution to the binding, resulting in the information of each amino acid (atom) contained being equally considered. This conflicts with the true biological DPI, so a specific method to portray the realistic drug-protein binding process is necessary. We choose residue fragments in the protein sequence and atom fragments of the drug SMILES as the processed input of our network. In protein or drugs, conserved fragments are the leading participant in DPI.

2.3 Fragment representation

MolTrans [5] applies a frequency-based mining method to mine the sub-sequence with high frequency. Although they modify the smallest sub-structure of the smallest granularity of the sequence to the fragment of the sequence, they only apply attention to the protein or drug embedding module without put attention on the interaction module instead of a CNN layer.

Therefore, they can learn a few features characterizing the drug-protein interaction. A better coding strategy to explore the DPI, not only a single protein or drug, is helpful in predicting the affinity score.

2.4 Pre-trained model

Previous studies [15,16] have demonstrated that pre-training can help the model on downstream tasks. The pre-training can help the model learn as many common data features as possible. These advanced methods with pre-training are shown below. GPT-2 [17] utilizes large-scale unsupervised data sets to generate texts. While applying the pre-trained model to downstream tasks can greatly improve the performance of downstream tasks. T5 [18] is based on the encoder-decoder pre-trained model and can achieve state-of-the-art results on multiple natural language tasks. From the above methods, pre-training is a decent strategy to learn features from large-scale unsupervised datasets. This is also one of the sources of inspiration for our method.

These strategies are mainly based on the extraction of

sequence features through self-attention. These methods have in common with the task of extracting features from protein sequence and drug representation string in biological processes. So the pre-training was also applied in FragDPI. The pre-training was used to mine the conserved fragments of the multi-interaction process between residue fragments in the protein sequence and atom fragments of the drug molecule. The fine-tuning stage is used to predict DPI affinity with the pre-trained model.

3 Methodology

3.1 Overview

The FragDPI is a deep neural network model to predict the drug-protein binding affinity. We apply the transformer encoder module to the sequence fragments, which helps us understand the interaction of conserved fragments from the protein sequence and drug molecule.

The structure of the model is shown in Fig. 1. We get the fragments' vocabulary with the descent frequency of the protein sequence and drug SMILES through the FCS algorithm. After obtaining conserved fragments of the drug-protein pairs, we tokenized the segmented fragments with vocabulary. Then we take the embedding operation to the tokenized fragment, including token embedding and position embedding to capture the sequence text information and position information. Following [19], we get the embedding vector to each fragment of the drug-protein pairs. The next step is the interaction module, which concatenates the drug embedding vector and protein embedding vector together to explore the interaction fragment of the pair. After the concatenation operation, we put the embedding vectors into the Bidireacion Transformer-Encoder.

A novel two-step training strategy was applied to the FragDPI, pre-training stage for fragments understanding and fine-tuning stage for predicting the binding affinity as shown in Fig. 1. The difference in the model between the two stages is that we add fully-connected layers to the model during the fine-tuning stage to output the affinity scores.

The overview of the training process is summarized in Algorithm 1.

Algorithm 1 Two-stage training process

Satge 1: FU pre-training
Input: Protein Amino Acid Sequence and Drug SIMLES Sequence, $[S_p :: S_d]$
Output: Masked Fragments

Tokenize the $[S_p :: S_d]$
Mask randomly
Get Embedding $E = E_t + E_p$
Init randomly the FragDPI
while loss is not stable **do**
  $P(x_{mask}) = FragDPI(E)$
  $loss = CrossEntropy(x_{predict}, x_{mask})$
end while
return pre-trained model

Satge 2: DTI prediction
Input: Protein Amino Acid Sequence and Drug SIMLES Sequence, $[S_p :: S_d]$
Output: Affinity Scores, A

Tokenize the $[S_p :: S_d]$
Get Embedding $E = E_t + E_p$
Init FragDPI with pre-trained checkpoint
while loss is not stable **do**
  $affinity = FragDPI(E)$
  $loss = MSE(A_{predict}, A_{true})$
end while
return affinity score

3.1.1 Conserved fragments mining

Before tokenizing the drugs (proteins) sequence, we need to mine and identify the conserved fragments to build a fragments' vocabulary. Due to the lack of labeled subsequence data, we use FCS (Frequent Consecutive Subsequence mining algorithm [5]) to identify the conserved fragments in protein amino acid sequences and drug SMILES sequences.

The FCS algorithm identifies a similar set of fundamental biochemical conserved fragments based on the frequency of tokens in massive unlabelled data. The FCS algorithm scans through tokenized set W and identifies the most frequent consecutive tokens, then updates every token in the tokenized set W with the new token. This operation merges frequent sub-

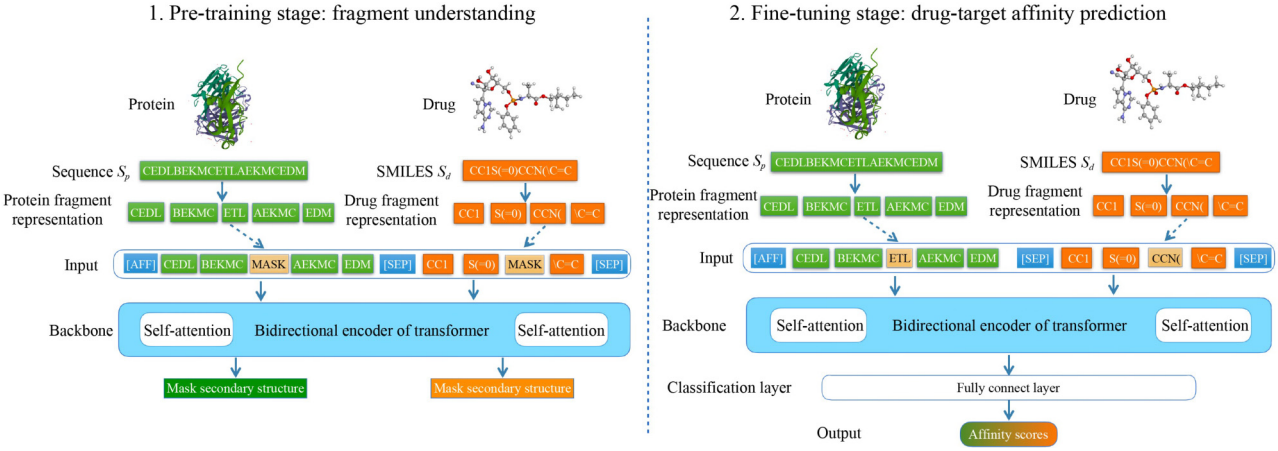

Fig. 1 Overview of model. The left is FU pre-training stage, which predicts masked fragments of the sequences. And the right is fine-tuning stage, which predicts the affinity scores

sequences into one token, and sub-sequences that are not frequent enough are decomposed into shorter tokens. The FCS algorithm provided the conserved fragments with high frequency in the massive protein datasets, helping the model explore the functional motif of binding position.

Using the FCS algorithm, we got a conserved fragments' vocabulary size of 23,614.

3.1.2 Unified coding to tokens

After the conserved fragments mining, the protein sequence and drug SMILES needed to be tokenized according to the conserved fragments' vocabulary.

In order to learn the relationship between conserved fragments of the drug-protein pair, we code it by an unified coding vector as input of the model. Therefore, we concatenate the drug and protein sequence and encode them in the meantime. As shown in Fig. 2, we split the concatenated sequence into conserved fragments. The tokenization of the conserved fragments corresponds to the vocabulary index we obtained in the previous step. Hence, we obtain the sequence tokenization, i.e., a one-dimensional vector containing the token number of each conserved fragment. Note that the conserved fragments vocabulary is sorted with long segments first and short ones last.

3.1.3 Embedding module

To capture the biological semantics of conserved fragments, we apply encoding methods from [19] to the token vector of the sequence.

In sequence modeling tasks, modern researches leverage token embedding to represent the content of sequence and position embedding to capture the position information in sequence. Note that token embedding and position embedding are built on a lookup dictionary with random initialization parameters. Protein parameter matrixes of token embedding and position embedding are denoted as W_p^t and W_p^p, and drug parameter matrix denotes as W_d^t and W_d^p. Final embedding is the sum of token embedding and position embedding. We could get drug representation (E_d) and protein representation (E_p) through following equations.

$$E_d = W_d^t * S_d + W_d^p * S_d, \tag{1}$$

$$E_p = W_p^t * S_p + W_p^p * S_p. \tag{2}$$

3.2 Interaction module

We choose the Bidirectional Transformers [19] as the backbone of our interaction module. Most modern works are encoding drugs and proteins separately, which thinks the process of encoding should be in the same domain. We concatenate E_d and E_p to feed into interaction module. Concretely, the input of the model, denoted as E, is characterized as $[E_d :: E_p]$ ([::] denotes the operation of concatenating) as shown "input" in Fig. 1.

$$E = [E_d :: E_p]. \tag{3}$$

Utilizing this form to input the model, FragDPI could model the relationship between conserved fragments between protein and drug. The Bidirectional Transformer Encoder leverages the self-attention mechanism to model the hidden state from the information of the other fragments.

3.3 Fragment understanding

In order to make the model enable to understand the correspondence of conserved fragments, we designed an unsupervised task named Fragment Understanding (FU). As shown in Fig. 1, we randomly mask fragments in the input of drug and protein. Then the goal of pre-training is to predict masked fragments. Formally speaking, the whole input sequence has n tokens$(x_0, x_1, ..., x_n)$. Furthermore, the mask position is noted x_i. h_i^{mask} is the final layer hidden state of masked position. In summary, the pre-training task can be

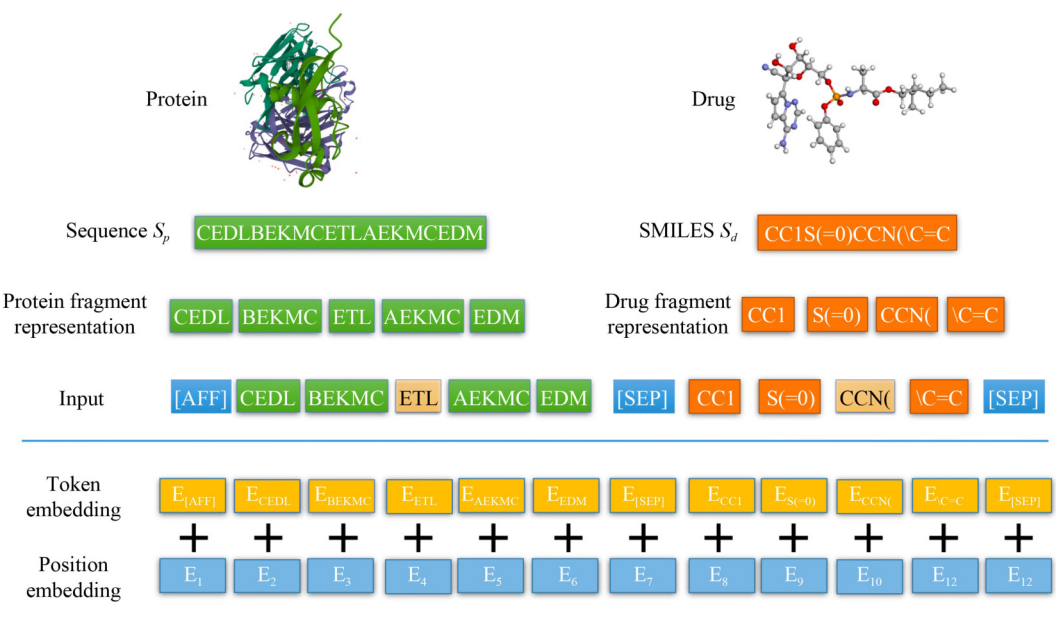

Fig. 2 Input of Model, which consists of token embedding and position embedding. Token embedding is used to represent the semantic of conserved fragments, and position embedding is used to provide position information of conserved fragments

formatted as the following equation:

$$P(x_i^{mask}|x_0,...,x_{i-1},x_{i+1},..,x_n) = softmax(h_i^{mask}). \quad (4)$$

Inspired from [15], we attempted to make model understand the correspondence of fragments. We mask randomly drug sequence and protein sequence separately. The strategy of mask is as followed: in 80% probability, we mask 15% of all fragments in each sequence at random; the masked token is replaced by a special token $[MASK]$ as shown in Fig. 1.

3.4 DPI prediction

The fine-tuning task is DPI prediction, which predicts the affinity score of input drug and protein. After pre-training about FU, we initialized the model with the pre-training parameter. Furthermore, we add a prediction head to accept the hidden of special token $[AFF]$. Specifically, each input will be added a special token as start token $[AFF]$, which token has no semantics. Furthermore, it could make attention to all tokens in input. Inspired by text classification in natural language processing, the hidden state of $[AFF]$, denoted as x_{aff}, is fed into the prediction head ($PreHead$), then outputs the score of corresponding drug and protein. In summary, this task can be described as the following formation:

$$Score = PreHead(x_{aff}), \quad (5)$$

In general, we divide the DPI prediction into a two-stage training task. For the first stage, the task of FU utilizes the information of other fragments to predict the masked fragments. For the second stage, the model makes DPI predictions based on the understanding of the fragment relationship.

4 Experiments and results

In this section, we will describe the implementation details of the model and experiments.

4.1 Dataset

For measuring binding affinity, we utilize half-maximal Inhibitory Concentration (IC 50) as the main index, which is an important measure of potency for a given agent in pharmacology. The available experimental data is collected by the publicly released database [20]. The drug-protein binding data, drug SMILES information and protein amino-acid sequences are extracted separately from BindingDB [21], STITCH [22], and UniRef [23].

For the convenience of calculation, we use the logarithm form of the maximal inhibitory concentration (shown like the formula below) of label data.

$$pIC50 = -\log_{10} IC50. \quad (6)$$

To verify the applicability of FragDPI to different types of proteins, four distinct functional protein data were collected for further experiments: nuclear Estrogen Receptors (ER), Ion Channels (Ion-C), Receptor Tyrosine Kinases (RTK) and G-protein-coupled receptors (GPCR). The four protein datasets are designed to handle different biological processes and they are described below.

●nuclear Estrogen receptors

ER is a transcription factor, a member of the nuclear receptor superfamily, which regulates the transcription of hundreds of genes and ultimately leads to cell division, and has an essential role in mammary gland development and in the proliferative growth of cells that occurs during pregnancy.

● Ion channels

Ion-C are special proteins on the plasma membrane that provide a channel through which charged ions can pass along an electrochemical gradient across the plasma membrane.

● Receptor tyrosine kinases

RTK are high-affinity cell surface receptors for many peptide growth factors, cytokines and hormones. They act as signal transducers that mediate cell-to-cell communication by phosphorylating tyrosine residues on key intracellular substrate proteins.

● G-protein-coupled receptors

GPCR are the largest and most diverse group of membrane receptors in eukaryotes, which are integral membrane proteins that are used by cells to convert extracellular signals into intracellular responses.

As for data split strategy, starting with pIC50-labeled samples, it completely excluded four classes of proteins above from the training set. Then the rest has been randomly splited into the training set and the default test set without the aforementioned four classes of protein targets.

Note that the training set doesn't include the above four classes data. Table 1 shows the statistics information of datasets.

4.2 Implementation details

The model is implemented in Pytorch. For pre-training, we set the batch size and the epoch to 8 and 50, and the learning rate is 1e–5. For fine-tuning stage, the set is the same as pre-training.

As for tokenizing process, the max length we set is 512. We utilize the pre-trained results of FCS in [5], which generate a set of the hierarchy of frequent sub-sequences for sequences. FCS was trained on DrugBank for drugs and BindingDB for proteisn. The model is trained on a single NVIDIA RTX 2080Ti GPU with a memory capacity of 11GB.

4.3 Baselines and metrics

The root mean squared errors (RMSE) and Pearson correlation coefficient (r) are chosen to measure the performance of models. RMSE measures the deviation between the observed value and the true value, which is often used in fitting calculations. In the field of natural sciences, the Pearson's r is widely used to measure the degree of linear correlation between two variables. Loosely speaking, the lower RMSE

Table 1 Statistics informations of datasets about the number of drug-protein pairs and the average length of tokenized sequences

Data class	Train	Test	ER	Ion-C	RTK	GPCR
Number of drug-protein pairs	263584	113168	3374	14599	34318	60238
Average length	243	243	233	352	427	186

and the higher the Pearson's r is corresponding to better performance; vice versa.

In order to fully verify the effectiveness of our model, we choose shallow and deep model. The followings are our baselines:

● Ridge regression

Ridge regression is used to estimate the coefficients of multiple regression models when independent variables are highly correlated. It has been applied in many domains, including chemistry, econometrics, and engineering.

● Lasso regression

Lasso regression is a method of linear regression through using shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

● DeepAffinity

DeepAffinity is a deep model, which combines the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [20].

● DeepDTA

DeepDTA is a CNN-based prediction model that comprises two separate CNN blocks, each of which aims to learn representations from SMILES strings and protein sequences [13].

● AttentionDTA

AttentionDTA uses one-dimensional Convolution Neural Networks (1D-CNNs) to extract the abstract information of drug and protein and associates attention mechanism to predict the binding affinity of DTI [6].

● BACPI

BACPI employs graph attention network and convolutional neural network (CNN) to learn the representations of compounds and proteins and develop a bi-directional attention neural network model to integrate the representations [24].

4.4 Result analysis

In order to gain the most optimal model for this task, we first did parameter grid experiments on the test dataset, using RMSE and Pearson's r as metrics to demonstrate the performance of the model. It was found that FragDPI has the best performance when using a 6-layer attention block, with 12 heads per, and the RMSE is 0.84 and Pearson's r is 0.84. Also, the hidden size and embedding size of FragDPI is 384.

4.4.1 The necessity of pre-training

After establishing the best model, we further compared the results of model with FU pre-training and without FU pre-training to validate the necessity of the FU pre-training phase. The result in all datasets as shown in Figs. 3 and 4. As expected, on the test dataset, the MRSE error was decreased by 0.08 and Pearson's r increased by 0.05 through FU pre-training. On the other datasets, the FU pre-trained FragDPI achieved a better performance in ER, GPCR and RTK, implying that FU pre-training stage would improve the prediction of the model.

Since the pre-training introduced a unified coding strategy for protein and drug conserved fragments with unsupervised learning, the model learned the contextual information of some protein and drug molecule fragments through FU pre-training. Moreover, because of the mechanism of action of DPI, the FragDPI mines more information related to protein and drug subsequences during the interaction and improves the prediction score. This pre-training approach can therefore be extended to another task to improve the overall effectiveness of the model.

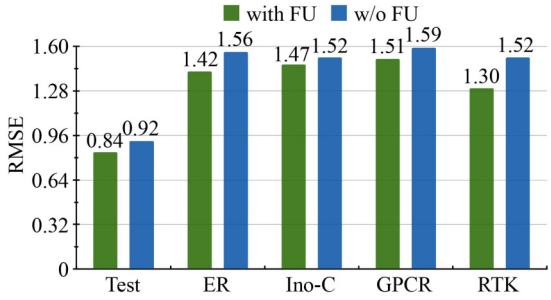

Fig. 3 RMSE results of FragDPI with FU pre-training and without FU pre-training in five different datasets. Phrase "with FU" denotes FragDPI with FU pre-training and characters "w/o FU" denotes without FU pre-training

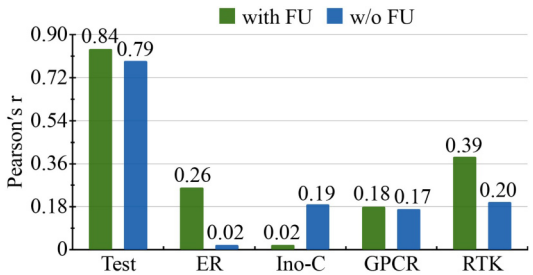

Fig. 4 Pearson's r results of FragDPI with FU pre-training and without FU pre-training

4.4.2 Compared with baselines

To verify the superiority of the model, we first compared FragDPI and baselines on the test dataset and the results are shown in the first column from Tables 2 and 3. Based on the results, it can be seen that the traditional shallow network approaches (like Ridge regression and Lasso regression) performs inferior to the deep network approaches. Among the deep network methods, FragDPI outperforms DeepDTA, attentionDAT in prediction, both in terms of RMSE and Pearson's metrics. The results show that FragDPI is very competent in predicting affinity scores.

Further comparison experiments were done to prove the generality of the model. The models were tested and compared on other datasets and the results are shown in Tables 2 and 3. For the four functional proteins, We compared the performance of different baselines and datasets on Pearson's r, and found that FragDPI had the same score as DeepAffinity on test, RTK, both of which were higher than the other methods, except BACPI. Specifically, FragDPI had the best performance on the ER dataset. However, the results of FragDPI on Ion-C or GPCR were not so good, implying that the generalizability of FragDPI needs further enhancement.

Interestingly, the Ridge regression method had the best predictions in the Ion-C type proteins, suggesting that comparable performance can also be achieved with shallow

Table 2 Results of main experiment based on RMSE. The **Bold** denotes the best performance

RMSE	Test	ER	Ion-C	GPCR	RTK
Ridge regression	1.23	1.46	**1.26**	**1.34**	1.51
Lasso regression	1.22	1.48	1.32	1.37	1.50
DeepAffinity	**0.78**	1.53	1.34	1.40	1.24
DeepDTA	0.98	1.48	1.45	1.40	1.25
AttentionDTA	1.18	1.97	1.72	1.85	1.75
BACPI	0.79	0.67	1.53	1.64	**0.37**
FragDPI(ours)	0.84	**1.42**	1.47	1.51	1.30

Table 3 Results of experiments based on Pearson's r

Pearson's r	Test	ER	Ion-C	GPCR	RTK
Ridge regression	0.54	0.18	**0.23**	0.20	0.10
Lasso regression	0.55	0.18	0.17	0.17	0.11
DeepAffinity	0.84	0.16	0.17	0.24	0.39
DeepDTA	0.75	0.17	0.11	**0.24**	0.34
AttentionDTA	0.69	0.13	0.04	0.19	0.23
BACPI	0.84	0.16	−0.01	0.26	**0.43**
FragDPI(ours)	**0.84**	**0.26**	0.02	0.18	0.39

models. In the ER protein category, the FragDPI method beat the other methods on Pearson's r. However, the results of FragDPI on Ion-C or GPCR were not so good, implying that the generalizability of FragDPI needs further enhancement.

Although the deep learning models described above achieve good performance, shallow network models also have the advantage of fast convergence and more straightforward implementation for some specific tasks. While deep learning models can achieve better results, it is also based on increasing the number of parameters and training time.

It is to be noted that the single-stream coding approach we use is more challenging to learn than the baseline models described above, as there are two types of data in the input and the model needs to distinguish and derive valid information from them. Our framework is more flexible and can also be applied to other task of the same type to extract sub-sequence features. In addition, we use an attention model to demonstrate the effectiveness of our conserved fragments coding.

4.4.3 Ablation studies

To investigate the performance of FragDPI further, we conducted additional ablation experiments on FragDPI regarding the encoding strategy of SSPro and FCS. As shown in Table 4, the unified coding of FragDPI with FCS achieved better performance in the test set, ER, Ion-C, and r in RTK, compared with SSPro.

Different strategies for protein conserved fragments exploration may lead to different results. And the results of ablation show that FCS has better performance. However, SSPro is more fine-grained in its concentration of protein information. In the process, it may consider many properties of a protein. Intuitively, it condenses over 1000 amino acids into about 50, which loses too much information. In contrast, FCS is based on pre-trained data and the frequency of conserved fragments. There is still a gap with the real conserved fragment. However, MolTrans [5] also verified the correspondence between FCS and biological conserved fragment. This ablation experiment proves that sub-sequence coding is an effective method, but this method still deserves to be explored and studied in depth.

4.4.4 Protein drug binding site analysis

To validate the model's ability to mine conserved fragments of drug-protein interactions, we picked specific drug-protein pairs, fed them into the model, and visualized the attention scores of the drug and protein.

In this paper, we have selected protein kinase C β type from the BindingDB database, an essential protein kinase in the human life process and is involved in various life processes such as transcription and apoptosis. The drug BDBM2591, a maleimide derivative, was also selected based on the corresponding data in the BindingDB. Detailed drug and protein information can be found in the BindingDB database.

To validate the model's ability to mine conserved fragments of drug-protein interactions, we picked specific drug-protein pairs, fed them into the model, and visualized the attention scores of the drug and protein.

In this paper, we have selected protein kinase C β type from the BindingDB database, an essential protein kinase in the human life process involved in various life processes such as transcription and apoptosis. The drug BDBM2591, a maleimide derivative, was also selected based on the corresponding data in the BindingDB. Detailed drug and protein information can be found in the BindingDB database.

The amino acid sequence of the protein kinase and the SMILES sequence of the drug BDBM2591 were concatenated into the model. The attention scores associated with the 103-116 token bits of the protein and the corresponding drug were output are shown in Fig. 5.

In Fig. 5, the vertical axis is the drug fragment and the horizontal axis is the protein fragment at 103-116 token bites. We can get the key information that the protein conserved residue fragment "IWD" is the region with the higher attention score of the drug, indicating that the fragment "IWD" contributes more to the drug-protein interaction. In addition, the fragment was also validated in the literature and is consistent with the binding site data provided in the BindingDB database [25,26].

From the above experiments, the multilayer attention modules proposed in this paper will focus more on the possible binding region of the drug to the protein during the model calculation. The information about this region will contribute more to the binding affinity of the model.

Table 4 Results of ablation based on RMSE (and Pearson's r), which compares different sequence fragments identifying strategy on FragDPI. SSPro means another strategy of identifying fragments

RMSE(r)	Test	ER	Ion-C	GPCR	RTK
FragDPI(SSPro)	0.98(0.76)	1.46(0.26)	1.47(0.065)	**1.36(0.24)**	**1.29**(0.35)
FragDPI(FCS)	**0.84(0.84)**	**1.42(0.26)**	**1.47**(0.02)	1.51(0.18)	1.30(**0.39**)

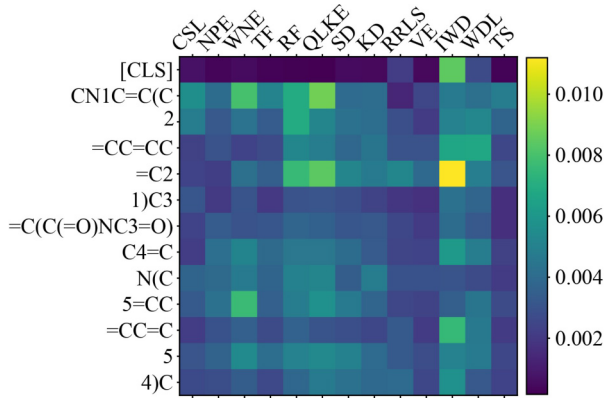

Fig. 5 The fragment attention score of the protein kinase C beta in 103-116 token bites and BDBM2591

4.4.5 Case study

As we developed FragDPI to discover potential drugs through protein sequences, we applied FragDPI to a specific task to illustrate the effectiveness of FragDPI for drug screening, so we performed a case study to demonstrate its predictive function. We chose the spike protein of COVID-19 (rSARS-CoV-2) [27] which is an essential receptor for COVID-19, as a specific target. In this case, we tried to use FragDPI in the dataset to find some potential molecules that would inactivate COVID-19 by binding to the critical spike target. We obtained the spike protein sequence (GenBank: QIG55857.1) from NCBI and applied the trained FragDPI to find the candidate drugs in the test dataset we mentioned in the section of Experiments and Results.

The top 10 candidate drugs with high affinity values are shown in Table 5. We have performed literature validation of the screened drug candidates to determine if they have the potential to bind Spike proteins.

According to the Table 5, the targets of the ten drugs identified include PI3-kinase subunit delta (PI3K-DELTA), Histone Deacetylase, Dipeptidyl peptidase (DPP) and Monoamine oxidase. We have compiled a list of each drug target by reviewing the literature. PI3K-DELTA is mainly responsible for phosphorylating specific signaling molecules that regulate cell growth and division [28]. DPP-4 inhibitors are a class of drugs prescribed to control hyperglycemia in adults with type 2 diabetes. Monoamine oxidase inhibitors (MAOIs) are treating different forms of depression and other nervous system disorders [30]. According to the literature, they are not strongly related to COVID-19.

Moreover, we found Histone Deacetylase (HDAC) enzyme activity signified the importance of HDAC isoform in tumorigenesis, cancer, neuronal disorders, parasitic/viral infections and other epigenetic regulations, so we have conducted an in-depth study of Histone deacetylase inhibitors. Histone deacetylase inhibitors suppress ACE2 and ABO simultaneously, which can prevent COVID-19 [29]. Furthermore, the corresponding drug of it, CHEMBL331781 and CHEMBL3317818, has a high-affinity score of 9.5085.

Screening candidate drugs through FragDPI can provide helpful predictive results, suggesting that FragDPI is a valuable tool for discovering potential drugs for target proteins.

5 Conclusion

Drug-protein prediction is a promising way to virtual screen drugs for the target. It could provide not only clinical guidance but also save resources and time. Modern researches encode protein or drug differently, which is separate and time-consuming. Furthermore, FragDPI models the protein and drug together with the attention mechanism. To the best of our knowledge, no method has been proposed to unify the coding of proteins and drugs until now. Therefore, we propose a model using a unified coding strategy, FragDPI, designed to extract conserved fragment features of proteins and drugs to make features more bio-interpretable.

The FragDPI employs Bidirectional Transformers as the backbone and uses a unified coding strategy for both proteins and drugs to explore the interaction information of DPIs at the conserved fragment level. The experimental results show that FragDPI achieves good results in Pearson's correlation coefficient r and RMSE.

And in the future, we will work on follow ing direction:

- In the unified coding strategy, we plan to find a novel way of coding to take into account the three-dimensional information of proteins and drugs as they interact in a three-dimensional structure, and the use of conserved fragment information may be insufficient for the critical information provided by the model.
- Due to the superiority of our model in uniform encoding and fragments extraction, we consider to improve the generality of the model and apply the model to other interaction tasks in the next step.
- In the FU pre-training phase, FragDPI requires a large

Table 5 Top 10 drugs with high affinity score to spike

Number	Candidates drug	Target name	Affinity score	Reference
1	BDBM198018::US9221795, 14	PI3-kinase subunit delta	9.5151	Cell growth and division [28]
2	**CHEMBL3317818**	**Histone Deacetylase 2 (HDAC2)**	**9.5085**	**Prevention or treatment of COVID-19 [29]**
3	**CHEMBL3317818**	**Histone deacetylase 8**	**9.5085**	**Prevention or treatment of COVID-19 [29]**
4	BDBM198096::US9221795, 91	PI3-kinase subunit delta	9.4446	Cell growth and division [28]
5	US9255098, Ex. 1::US9255098, Ex. 4	Dipeptidyl peptidase 4 (DPP4)	9.3964	Chronic hyperglycemia [28]
6	CHEMBL3605370	Monoamine oxidase	9.3290	Depression [30]
7	CHEMBL3605370	Monoamine oxidase	9.3290	Depression [30]
8	US9499523, 6	PI3-kinase subunit beta	9.3247	DNA replication and repair [31]
9	US9221795, 87	PI3-kinase subunit delta	9.3219	Cell growth and division [28]
10	US9169243, 41	AKT/p21CIP1	9.3119	Unknown

parameter space, so the model size is large. In the next step we try to optimise the model to make it lighter and faster.

Acknowledgements This work was supported by the National Key R&D Program of China (2019YFA0904303).

References

1. Swinney D C, Anthony J. How were new medicines discovered? Nature Reviews Drug Discovery, 2011, 10(7): 507–519
2. Gupta S, Jadaun A, Kumar H, Raj U, Varadwaj P K, Rao A R. Exploration of new drug-like inhibitors for serine/threonine protein phosphatase 5 of *Plasmodium falciparum*: a docking and simulation study. Journal of Biomolecular Structure and Dynamics, 2015, 33(11): 2421–2441
3. Yuriev E, Agostino M, Ramsland P A. Challenges and advances in computational docking: 2009 in review. Journal of Molecular Recognition, 2011, 24(2): 149–164
4. Huang K, Fu T, Glass L M, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug-target interaction prediction. Bioinformatics, 2020, 36(22−23): 5545−5547
5. Huang K, Xiao C, Glass L M, Sun J. MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics, 2021, 37(6): 830–836
6. Zhao Q, Xiao F, Yang M, Li Y, Wang J. AttentionDTA: prediction of drug−target binding affinity using attention model. In: Proceedings of 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019, 64−69
7. Liao Z, You R, Huang X, Yao X, Huang T, Zhu S. DeepDock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information. In: Proceedings of 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019, 311−317
8. Bai F, Morcos F, Cheng R R, Jiang H, Onuchic J N. Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(50): E8051–E8058
9. Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, Zhang J, Weng T, Zhang Z, Wu Z, Cheng L, Shi D, Lu X, Lei J, Crispin M, Shi Y, Li L, Li S. Molecular architecture of the SARS-CoV-2 virus. Cell, 2020, 183(3): 730–738.e13
10. Shu X, Royant A, Lin M Z, Aguilera T A, Lev-Ram V, Steinbach P A, Tsien R Y. Mammalian expression of infrared fluorescent proteins engineered from a bacterial phytochrome. Science, 2009, 324(5928): 804–807
11. Pahikkala T, Airola A, Pietila S, Shakyawar S, Szwajda A, Tang J, Aittokallio T. Toward more realistic drug-target interaction predictions. Briefings in Bioinformatics, 2015, 16(2): 325–337
12. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013, 1025−1033
13. Özturk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics, 2018, 34(17): i821–i829
14. Nguyen T, Le H, Venkatesh S. GraphDTA: prediction of drug−target binding affinity using graph convolutional networks. BioRxiv, 2019: 684662
15. Devlin J, Chang M W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019, 4171−4186
16. Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, Gao J, Zhou M, Hon H W. Unified language model pre-training for natural language understanding and generation. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 1170
17. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9
18. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21: 1–67
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6000−6010
20. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound−protein affinity through unified recurrent and convolutional neural networks. Bioinformatics, 2019, 35(18): 3329–3338
21. Liu T, Lin Y, Wen X, Jorissen R N, Gilson M K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. Nucleic Acids Research, 2007, 35(S1): D198–D201
22. Kuhn M, Von Mering C, Campillos M, Jensen L J, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Research, 2008, 36(S1): D684–D688
23. Suzek B E, Wang Y, Huang H, McGarvey P B, Wu C H, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 2015, 31(6): 926–932
24. Li M, Lu Z, Wu Y, Li Y. BACPI: a bi-directional attention neural network for compound−protein interaction and binding affinity prediction. Bioinformatics, 2022, 38(7): 1995–2002
25. Leonard T A, Różycki B, Saidi L F, Hummer G, Hurley J H. Crystal structure and allosteric activation of protein kinase C βII. Cell, 2011, 144(1): 55–66
26. Sutton R B, Sprang S R. Structure of the protein kinase cβ phospholipid-binding C2 domain complexed with Ca^{2+}. Structure, 1998, 6(11): 1395–1405
27. Thao T T N, Labroussaa F, Ebert N, V'kovski P, Stalder H, Portmann J, Kelly J, Steiner S, Holwerda M, Kratzel A, Gultom M, Schmied K, Laloli L, Hüsser L, Wider M, Pfaender S, Hirt D, Cippà V, Crespo-Pomar S, Schröder S, Muth D, Niemeyer D, Corman V M, Müller M A, Drosten C, Dijkman R, Jores J, Thiel V. Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. Nature, 2020, 582(7813): 561–565
28. Tzenaki N, Papakonstanti E A. p110δ PI3 kinase pathway: emerging roles in cancer. Frontiers in Oncology, 2013, 3: 40
29. Takahashi Y, Hayakawa A, Sano R, Fukuda H, Harada M, Kubo R, Okawa T, Kominato Y. Histone deacetylase inhibitors suppress *ACE2* and *ABO* simultaneously, suggesting a preventive potential against COVID-19. Scientific Reports, 2021, 11(1): 3379
30. Volz H P, Gleiter C H. Monoamine oxidase inhibitors. Drugs & Aging, 1998, 13(5): 341–355
31. Kumar A, Redondo-Muñoz J, Perez-García V, Cortes I, Chagoyen M, Carrera A C. Nuclear but not cytosolic phosphoinositide 3-kinase beta has an essential function in cell survival. Molecular and Cellular Biology, 2011, 31(10): 2122–2133

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

Feng Yang is a PhD candidate in the School of Computer Science, Wuhan University, China. His current research interests include machine learning, retrosynthesis prediction and metabolic pathway design.

Juan Liu is a professor in the School of Computer Science, Wuhan University, China. Her research interests include machine learning, data mining, bioinformatics, pattern recognition, and artificial intelligence methods for medicine.

Qiang Zhang is a PhD candidate in the School of Computer Science, Wuhan University, China. Her current research interests include retrosynthesis prediction, metabolic pathway design, bioinformatics, and machine learning.

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

Hayat Ali Shah received his MS degree in Computer Science from Virtual University of Pakistan, Pakistan in 2018. He is currently a PhD candidate in the School of Computer Science, Wuhan University, China. His research interests are simulated alignments, multiple sequence alignments, machine learning, prediction and reconstruction of metabolic pathways.