Computational Biology and Chemistry 110 (2024) 108058

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

Advancing Drug-Target Interaction prediction with BERT and subsequence embedding

Zhihui Yang, Juan Liu *, Feng Yang, Xiaolei Zhang, Qiang Zhang, Xuekai Zhu, Peng Jiang

Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, 430072, Hubei province, China

ARTICLE INFO

Keywords:
Drug-Target interaction
BERT
Transfer learning
Subsequence embedding
Deep learning

ABSTRACT

Exploring the relationship between proteins and drugs plays a significant role in discovering new synthetic drugs. The Drug-Target Interaction (DTI) prediction is a fundamental task in the relationship between proteins and drugs. Unlike encoding proteins by amino acids, we use amino acid subsequence to encode proteins, which simulates the biological process of DTI better. For this research purpose, we proposed a novel deep learning framework based on Bidirectional Encoder Representation from Transformers (BERT), which integrates high-frequency subsequence embedding and transfer learning methods to complete the DTI prediction task. As the first key module, subsequence embedding allows to explore the functional interaction units from drug and protein sequences and then contribute to finding DTI modules. As the second key module, transfer learning promotes the model learn the common DTI features from protein and drug sequences in a large dataset. Overall, the BERT-based model can learn two kinds features through the multi-head self-attention mechanism: internal features of sequence and interaction features of both proteins and drugs, respectively. Compared with other methods, BERT-based methods enable more DTI-related features to be discovered by means of attention scores which associated with tokenized protein/drug subsequences.

We conducted extensive experiments for the DTI prediction task on three different benchmark datasets. The experimental results show that the model achieves an average prediction metrics higher than most baseline methods. In order to verify the importance of transfer learning, we conducted an ablation study on datasets, and the results show the superiority of transfer learning. In addition, we test the scalability of the model on the dataset in unseen drugs and proteins, and the results of the experiments show that it is acceptable in scalability.

1. Introduction

The pursuit of novel therapeutics for disease treatment is a driving force behind drug discovery. However, traditional methods for drug discovery often incur extensive costs and demand significant labor (Santos et al., 2017). The identification of potential compounds is particularly hard, necessitating rigorous assay experiments and the evaluation of over 97M potential compounds within candidate databases (Broach and Thorner, 1996; Huang et al., 2021). In light of these challenges, virtual drug screening has emerged as a potent tool for drug discovery.

At the key of virtual drug screening lies the determination of whether a target-drug candidate interaction exists. The incentive for identifying interactions with already approved drugs is two-fold: it can significantly decrease new drug development costs, and mitigate safety concerns associated with novel drug candidates. Thus, predicting drug-target interactions (DTIs) via computer-aided drug discovery (CADD) is of paramount importance.

Initial approaches to DTI identification within the scope of CADD focused on leveraging drug and target structure information. For instance, the Quantitative Structure-Activity Relationship (QSAR) method, which relies on structural similarity to predict bioactivity, was introduced. However, QSAR's performance is hampered by limitations such as sample size and the non-specific nature of the active molecule (Zanni et al., 2014; Ewing et al., 2001; Österberg et al., 2002). Other methodologies incorporated molecular docking techniques to anticipate DTIs, including one group that employed AutoDock in tandem with the relaxed-complex method to reveal a novel mode of HIV integrase inhibition (Schames et al., 2004). Molecular docking endeavors to design drugs by understanding the structural characteristics of the receptor, along with the nature of receptor-drug molecule interactions (Sethi et al., 2019). However, due to insufficient accuracy and efficiency in scoring binding energy functions, these methods have fallen short of precise DTI prediction (Yuriev et al., 2011).

* Corresponding author.
E-mail address: liujuan@whu.edu.cn (J. Liu).

https://doi.org/10.1016/j.compbiolchem.2024.108058
Received 8 August 2023; Received in revised form 1 February 2024; Accepted 12 March 2024
Available online 5 April 2024

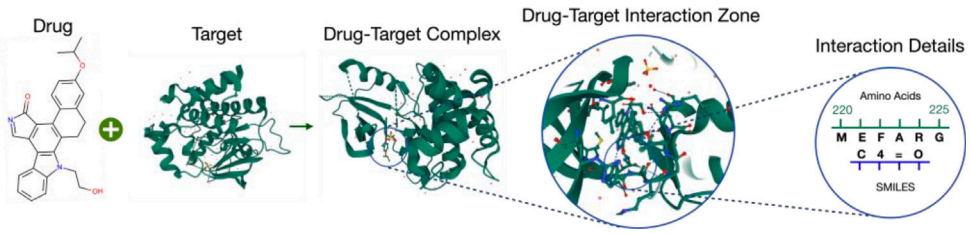

Fig. 1. The binding process of the crystal structure of mixed-lineage kinase MLK1 complexed with VIN 6331.

The recent surge in biological data and advancements in machine learning have spurred the proposal of novel machine learning algorithms. These algorithms, informed by drug and protein representation sequences and interaction networks, address the constraints posed by limited structural data and the high costs of molecular docking. For example, Cheng et al. established a Network-Based Inference (NBI) model to deduce new DTIs using drug-target bipartite network topology similarity (Cheng et al., 2012). Similarly, He et al. put forward the SimBoost approach that employs gradient boosting machines to address the issue of missing endpoints for non-interacting drug-target pairs (He et al., 2017).

To elevate the representation of the DTI network, Chen et al. designed a heterogeneous network-based model, NRWRH, by integrating drug-drug similarity networks, protein-protein similarity networks, and drug-target interaction networks (Chen et al., 2012). In an interesting turn, Wen et al. developed a DeepDTI model based on the Deep Belief Network (DBN), a conglomeration of Restricted Boltzmann Machines (Wen et al., 2017).

In the wake of rapid advancements in deep learning, numerous notable methodologies for predicting Drug-Target Interactions (DTIs) have emerged (Pan et al., 2022). Given the predominant representations of molecules as sequences and graphs, deep learning has seen extensive application in these areas. For instance, Hakime et al. formulated a Convolutional Neural Network (CNN)-based deep learning model that solely employs sequence information of targets and drugs for DTI prediction (Öztürk et al., 2018). Likewise, DeepAffinity, another sequence-based model, utilizes an auto-encoder module in conjunction with a Recurrent Neural Network-CNN (RNN-CNN) for DTI prediction (Karimi et al., 2019).

In parallel, there has been significant work pertaining to drug molecular graph representations. A case in point is GNN-CPI, a method that integrates a Graph Neural Network (GNN) with compounds and a CNN with proteins to predict compound-protein interactions (Tsubaki et al., 2019). Other examples include GraphDTA and DeepConv-DTI, which rely on graph and sequence data and incorporate a global max pooling operation into their models (Nguyen et al., 2021; Lee et al., 2019).

However, these approaches primarily focus on protein sequences and drug molecule graphs to model DTIs, while overlooking the underlying biological processes at the heart of DTI. In biological contexts, DTIs are predominantly driven by specific fragments within protein and drug molecule sequences (Bai et al., 2016), giving rise to a growing interest in fragment-based drug design (Hajduk and Greer, 2007). Fig. 1 illustrates this concept with the example of mixed-lineage kinase MLK1 (UniProtKB ID: P80192) complexed with a compound of VIN (PDBBind ID: 3DTC), demonstrating that specific amino acid fragments bind to ligand structure counterparts (Hudkins et al., 2008).

Deep learning methods such as MolTrans and FragDPI employ the high frequency of protein and drug subsequences to model the DTI process, thereby enhancing the interpretability of the interaction process (Huang et al., 2021; Yang et al., 2022). While MolTrans pioneered the use of subsequences for protein and drug encoding, it struggles with learning protein subsequence features without pretraining from large DTI datasets. In addition, the subsequences in the paper refers to a contiguous or sequence of amino acids and SMILES characters within a biological molecule, like drug and protein.

While the subsequence representation approach is promising, MolTrans exhibits limitations, as it only learns DTI features specific to the training dataset, lacking comprehensive prior knowledge about DTI information. To address this issue, we introduce an approach for DTI prediction based on the Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018). Our key contributions are as follows:

- **Transfer Learning.** This is an attempt to consider transfer learning of large DTI datasets to apply it on the specific DTI prediction task. The transfer learning conduces the model to learn the tokenized subsequences interaction feature of the protein and drug sequences. Meanwhile, the pretraining assists the model learn drug subsequence and protein subsequence interior features.
- **Subsequence Embedding.** Different from previous works of encoding drugs and proteins that use individual amino acids or atoms (bonds), we adopt subsequences vocabulary to embed drugs and proteins sequence. This approach allows sequences to be encoded according to functional modules, preserving the functional units of the sequence well.
- **Interaction Mechanisms Mining.** Instead of previous works that encode protein sequences and drug sequences separately, our BERT-based model attempts to combine them together at the encoding stage and then calculate the attention score between protein and drug subsequences, making it easier to explore the interaction module between them.

The rest of the content is organized as follows. Section 2 describes the proposed materials and methods for the model. Experiments set up and results are provided in Section 3. Finally, we make a conclusion and further research topics about some relevant issues in Section 4.

The code and datasets are available at the http://lanproxy.biodwhu.cn:9099/YangZhihui/subsequence_for_dpi.git.

2. Materials and methods

2.1. Benchmark dataset

2.1.1. BindingDB dataset

BindingDB is a public database that hosts a wealth of experimentally determined binding affinities for protein-ligand complexes (Liu et al., 2007). For model pretraining, we utilize 263,584 unlabeled drug-target pairs, and for model fine-tuning, we employ 10,665 drugs and 1413 proteins. To ensure balanced training, an equal number of negative DTI pairs are drawn in the training set to match the positive samples.

The pretraining dataset from BindingDB, sourced from deepaffinity (Karimi et al., 2019), comprises items with positive drug-target affinity scores and lacks negative samples. Therefore, the pretraining stage focuses on learning the binding feature for drug and target protein sequences.

Table 1
Fine-tune datasets statistics.

Dataset	Drug numbers	Proteins numbers	Positive item numbers	Negative item numbers
BIOSNAP	4510	2181	9619/1374/2748	9619/1374/2748
DAVIS	68	379	1043/160/303	1043/2846/5708
BindingDB	10 665	1413	6334/927/1905	6334/5717/113 384

Note: The data number in the positive and negative item numbers is divided as train/validate/test.

2.1.2. DAVIS dataset

The DAVIS dataset is a comprehensive resource featuring the interactions of 72 kinase inhibitors with 442 kinases, covering more than 80% of the human catalytic protein kinome. The dataset comprises 68 drugs and 379 proteins and serves as one of the fine-tuning datasets in our study (Davis et al., 2011).

2.1.3. BIOSNAP dataset

The BIOSNAP dataset, derived from Huang et al. (2021), encompasses 4510 drug nodes and 2181 protein targets, as well as 13,741 DTI pairs obtained from DrugBank (Wishart et al., 2018; Zitnik et al., 2018). This dataset considers interactions between small chemicals (drugs) and target proteins, all of which have been experimentally validated via biological experiments or formal pharmacological studies.

The datasets above provide the fine-tune training to learn the specific DTI feature-related field based on the pretrained model. The statistics information about the datasets is shown in Table 1.

2.2. Conservative subsequences vocabulary

The initial step of our model involves encoding drug and protein sequences using a conservative subsequence vocabulary. As the primary coding dictionary, this vocabulary supplies all subsequence contents along with their corresponding frequencies. This vocabulary was sourced from Huang et al. (2021), who used the Frequent Consecutive Subsequence (FCS) algorithm to mine 23,532 drug subsequences and 16,693 protein subsequences from the Uniprot dataset, encompassing 560,823 unique protein sequences (Consortium, 2019), and the ChEMBL database, which includes 1,870,461 drug SMILES strings (Gaulton et al., 2012). The FCS algorithm is designed to decompose each sequence of proteins and drugs hierarchically into subsequences, smaller subsequences, and individual atoms and amino acid symbols. Details on the FCS algorithm are available in Sennrich et al. (2015) and Huang et al. (2021).

By employing the FCS algorithm, we obtain a dictionary of hierarchical frequent subsequences for sequences, which act as a key part to encode the DTI pairs. The dictionary is make up of the rank, frequency and subsequence from drug and protein.

Table 2 exhibits the top 10 subsequences for proteins and drug SMILES in the dictionary, excluding single amino acid symbols. This consecutive subsequence vocabulary supplies foundational and semantically meaningful biomedical content, which correlates with the frequently recurring fundamental units of drugs and proteins. This vocabulary offers a guideline for encoding drug SMILES and protein sequences. Additionally, the length of the protein subsequence varies from a single amino acid (1aa) to a chain of eleven (11aa).

2.3. Drug and protein representation

Drug and protein representation in our model is rooted in the domain knowledge of Drug-Target Interactions (DTIs) occurring at the subsequence level, shown as "Conservative sub-sequence vocabulary" in Fig. 2. These dictionaries help us encode the DTI pairs that need to be processed.

Initially, we have the DTI pair which should be encoded represented on the left side of Fig. 2. Then we use the conservative subsequences vocabulary to encode the DTI pair as depicted "Protein and Drug conservative subsequences vocabulary" in Fig. 2, which is used to decompose the amino acids and drug SMILES subsequence. According to the encode strategy, we get two encoding list for the DTI pair, shown as the "Protein and Drug sequence fragments" in Fig. 2. Subsequently, we collate the rank list as the initial representation vector to the corresponding drug and protein sequence. Each of these subsequences is later embedded into a unique representation vector, forming the foundation for our representation model.

Table 2
Top 10 subsequences of drug and protein ().

Rank	Drug subsequence	Frequency	Protein subsequence	Frequency
1	cc	10 073 898	LL	1 908 409
2	O)	4 088 952	GL	1 181 538
3	(=	3 440 472	EL	1 173 174
4	(=O)	3 141 450	SL	1 159 393
5	ccc	2 546 300	GG	1 070 493
6	(C	2 445 313	SS	1 057 957
7	C(=O)	2 102 814	EE	1 030 207
8	[C	1 968 643	DL	984 725
9	[C@	1 962 590	PL	770 883
10	H]	1 924 333	TG	710 749

Note: This is a prt of the conservative subsequences vocabulary.

Therefore, by leveraging the FCS algorithm and the conserved subsequences dictionaries for proteins and drugs, we are able to effectively encode DTI pairs and capture the essential information necessary for our representation model. This approach ensures a rigorous and logical representation process for DTIs.

2.4. Pretraining for the model

Our pretraining phase begins with the self-supervised training of the language model using BindingDB dataset, after which the model is restructured into a task-specific model, such as a DTI prediction model. In the field of pretraining for protein and drug sequence community embedding learning, several recent studies (Elnaggar et al., 2020; Brandes et al., 2022; Zhang et al., 2021; Lee et al., 2020; Chithrananda et al., 2020) have explored the efficacy of self-supervised learning in either protein sequences or SMILES sequences to enhance generalizability for downstream tasks.

However, the exploration of interaction embedding learning has remained relatively uncharted. Prior research tends to focus on the representation of individual amino acids and chemical characters within sequence contexts, often overlooking motifs and subsequences as collective entities. To rectify this, we apply self-supervised learning to large unlabeled DTI datasets in order to unearth the interaction components within sequences.

Our pretraining method is specifically designed to capture the overarching feature of the subsequence in the course of DTI. To learn critical interaction-related features, the pretraining task is aligned with masked token learning. By employing this task, the model can learn the attention score of each token contributing to the interaction, thus enhancing its predictive power.

2.5. The model framework

This section provides an overview of our proposed model, as depicted in Fig. 3. The model consists of two main components: a novel

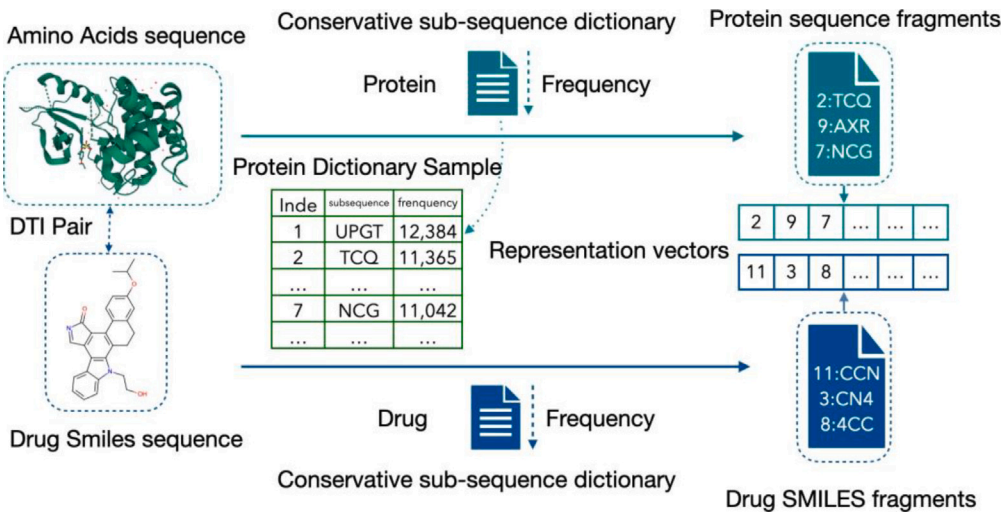

Fig. 2. The drug and protein representation process.

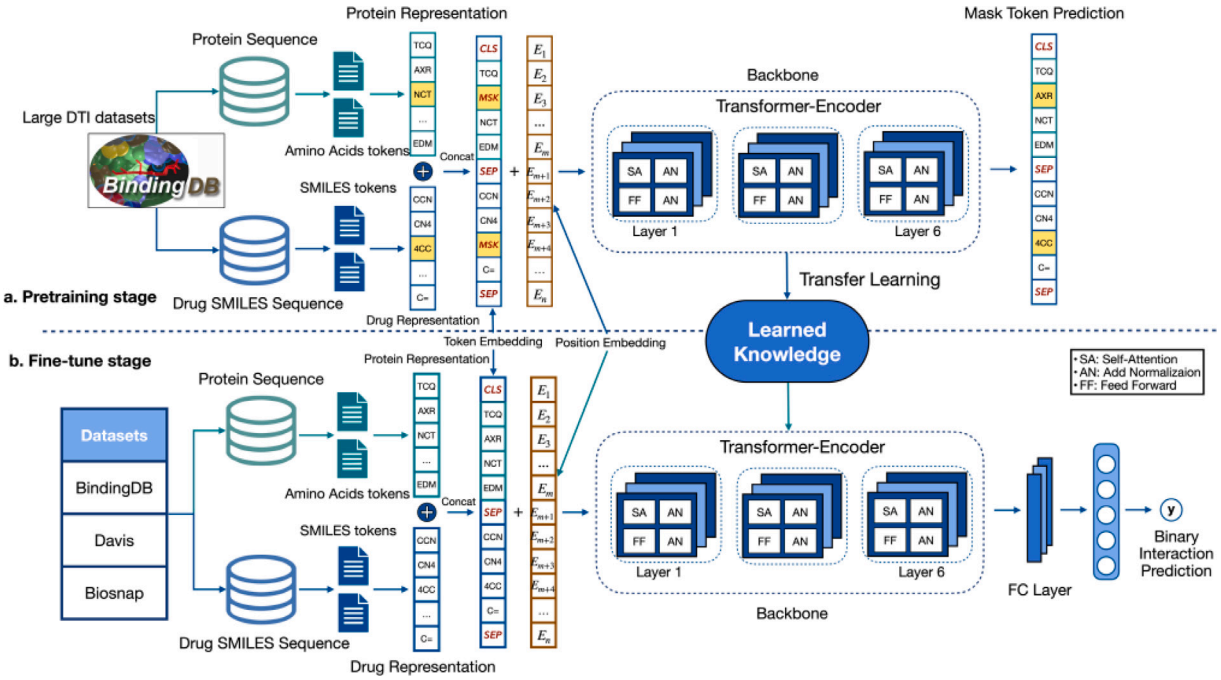

Fig. 3. The pipeline of our BERT-based model. The model mainly based on the two stages, pretraining and fine-tune, and they share the same network structure except another SoftMax layer in the fine-tune stage. a. In the pretraining stage, the model learns attention features associated with tokenized protein/drug subsequences by predicting masked tokens on a large number of unlabeled data. b. In the fine-tune phase, the model adds task-specific parameters to transfer learning to specific supervised learning tasks for DTI prediction based on the pretrained weights.

BERT-based network and a two-stage model training process. The two-stage training process encompasses a pretraining stage that utilizes masked token prediction, followed by a fine-tuning supervised learning task aimed at DTI prediction.

2.5.1. BERT-based model

BERT, a transformative model in the pretraining field, was introduced as a benchmark (Devlin et al., 2018). It attained state-of-the-art performance in multiple NLP tasks, thereby instigating a new era in pretraining models. We used BERT to encode drug-target interactions by modeling the relationships between drug subsequences (tokens) and protein subsequences (tokens). This approach is distinct from the encoding methods that take a single sequence input and only consider the relationships between amino acids or atoms. The BERT-based model initiates with the subsequence representation from proteins and drugs.

2.5.2. Subsequence representation

Through the subsequence mining process for drugs and proteins, we obtain a dictionary of subsequence tokens, as shown the Amino Acids tokens and SMILES token in Fig. 3, which build two token dictionaries with index of the rank, just like the sequence fragments in Fig. 2. For instance, we deploy the FCS algorithm on the drug-protein pair p_i and d_i, leading to the acquisition of the protein representation R_p and the drug representation R_d, as indicated by the equations below. The length of subsequences from drugs and proteins are not limited to three letters.

$$R_p = [TCO, AXRF, \ldots, NCT, EDM] \tag{1}$$

$$R_d = [CCN, CN4, \ldots, 4CC, C=] \tag{2}$$

2.5.3. Subsequence mask and tokenization

By masking the subsequence, we derive the masked representation of the protein and drug. The BERT model's input sequence comprises the drug-protein pair p_i and d_i, separated by a $[SEP]$ token, and an initial "classification token" $[CLS]$ utilized for prediction. Therefore, the input sequence becomes R_i. Furthermore, we mask 15% of the tokens, as illustrated in the yellow box in Fig. 3. The masking process's details are as follows: an 80% probability that the token is replaced by $[MSK]$, a 10% probability that the token is replaced by another random token, and a 10% probability that the token remains unchanged.

$$R_i = [[\mathbf{CLS}], TCO, [\mathbf{MSK}], \ldots, EDM, [\mathbf{SEP}], CCN, CN4, \ldots, [\mathbf{MSK}], C =, [\mathbf{SEP}]] \quad (3)$$

Before performing the embedding operation, the input needs to be tokenized to convert the input sequence tokens into their corresponding index in the subsequence vocabulary. Consequently, we receive the sequence tokenization, which is a one-dimensional vector containing the token number of each recurring subsequence fragment.

To retain the subsequence and position information in the sequence, we employ two kinds of embeddings for the tokens: token embedding and position embedding. Token embedding aims to transform the subsequence into representation space, where each token is represented by a unique vector. Position embedding, on the other hand, is used to preserve the order of the subsequence, thereby encoding the position information of a subsequence into a feature vector.

2.5.4. Transformer encoder module

After the subsequence embedding operation, we add the corresponding elements of the two embedding vectors to derive the final sequence representation vector. This source vector is then fed into the Transformer encoder module (Vaswani et al., 2017) to uncover the relationship between the subsequences of the drug and protein. The transformer encoder employs multi-head self-attention to identify the various patterns in drug and protein subsequences.

It is important to note that the Transformer encoder utilizes token embeddings and positional embeddings to facilitate this process, which provide the representation and relationship between subsequences from protein and drug. Token embeddings encode the semantic meaning of individual tokens within the input drug and protein sequences. They capture the contextual information in the sequence and contribute to the understanding of the relationships between different tokens. On the other hand, positional embeddings encode the positional information of the tokens, enabling the model to discern the relative positions of the tokens within the sequence. By incorporating token embeddings and positional embeddings into the Transformer encoder, our model can effectively leverage the rich information contained within the drug and protein subsequences.

2.5.5. Fine-tune of DTI prediction task

Upon completion of the pretraining stage, the model has acquired the common features of DTI. To further enhance the model's specific capabilities in downstream tasks, we fine-tune the model using three datasets: BIOSNAP, DAVIS, and BindingDB.

2.6. Implementation

We have implemented our Bert-based model using Pytorch (Paszke et al., 2019) and trained it on eight *Nvidia RTX 3090 GPUs*. The operating system is 64-bit Ubuntu 18.04.6 with Python 3.7.13. Our GPU server platform provides outstanding hardware and software support for the algorithm.

In token construction, the subsequence vocabulary comprises 23 532 drug substructures and 16 693 protein substructures, so we apply 40 230 tokens (including five special tokens) to the sequence. Regarding the source length of the drug and protein sequences, we set the maximum length of the tokenized sequence at 512.

In the BERT model, we employ six hidden transformer encoder layers and allocate 12 attention heads to each encoder with an intermediate dimension of 1536. In the fine-tuning stage, the size of the final network of the FC layer is 384. During the training process, we set the batch size and the number of epochs to 32 and 100, respectively. The learning rate and dropout rate are 1e−5 and 0.1.

3. Experiments and results

3.1. Experimental setup

3.1.1. Compared methods

To validate the advantages of our BERT-based model, we compared it against seven existing methods, encompassing two traditional methods and five state-of-the-art deep learning methods, all of which have been applied in the field of DTI. The traditional methods are Logistic Regression (Rogers and Hahn, 2010; Cao et al., 2013) and a DNN with three layers and a hidden size of 1024.

The five advanced deep learning baseline methods are following:

- **GNN-CPI** (Tsubaki et al., 2019): Utilizes a graph neural network to encode drugs and a CNN to encode proteins.
- **DeepDTI** (Wen et al., 2017): Based on Restricted Boltzmann Machines.
- **DeepDTA** (Öztürk et al., 2018): Applies CNN on drug and protein sequences.
- **MolTrans** (Huang et al., 2021): Mines substructural patterns through a transformer encoder. This model served as our inspiration.
- **HyperAttentionDTI** (Zhao et al., 2022): Uses stacked 1D-CNN layers to learn features and infers an attention vector for each amino acid–atom pair.
- **BACPI** (Li et al., 2022): Employs graph attention network and convolutional neural network (CNN) to learn the representations of compounds and proteins. Develops a bi-directional attention neural network model to integrate the representations.
- **DrugBAN** (Bai et al., 2023): Works on drug molecular graphs and target protein sequences. Utilizes conditional domain adversarial learning to align learned interaction representations across different distributions for better generalization on drug–target pairs.

3.1.2. Evaluation metrics

To obtain a more realistic assessment of our model's prediction performance as a binary classification task, we established ROC-AUC (area under the receiver operating characteristic curve) and PR-AUC (area under the precision–recall curve) as the two main evaluation metrics to measure the models' performance.

ROC-AUC measures the model's ability to discriminate between positive and negative samples across different classification thresholds. It considers both true positive rate (sensitivity) and false positive rate (1 - specificity). A higher ROC-AUC value indicates better discriminative power and overall performance.

PR-AUC, on the other hand, focuses on the trade-off between precision and recall. It measures the model's ability to correctly identify positive samples while minimizing false positives.

Furthermore, to gauge the capability of the methods in handling both positive and negative samples, we used sensitivity and specificity as two additional metrics.

3.2. Experimental results

To validate the significant performance of our BERT-based model, we conducted a series of experiments comparing it with all seven baseline models. Given the pivotal role of the pretraining stage in our model, we further investigated its impact by performing an ablation study. Finally, we conducted an additional experiment to test the scalability of the model by predicting Drug-Target Interactions (DTIs) of unseen drugs and proteins.

Table 3
Comparison results of baselines on datasets.

Method	ROC-AUC	PR-AUC	Sensitivity	Specificity
Dataset 1: BIOSNAP				
LR	0.638 ± 0.004	0.654 ± 0.011	0.583 ± 0.039	0.610 ± 0.018
DNN	0.729 ± 0.003	0.745 ± 0.010	0.716 ± 0.040	0.788 ± 0.024
GNN-CPI	0.849 ± 0.007	0.820 ± 0.004	0.780 ± 0.014	0.819 ± 0.012
DeepDTI	0.826 ± 0.005	0.846 ± 0.006	0.794 ± 0.027	0.815 ± 0.017
DeepDTA	0.834 ± 0.005	0.849 ± 0.006	0.726 ± 0.015	0.813 ± 0.012
MolTrans	0.851 ± 0.002	0.832 ± 0.004	0.754 ± 0.032	**0.836 ± 0.014**
HyperAttentionDTI	0.825 ± 0.006	0.830 ± 0.011	0.800 ± 0.020	0.812 ± 0.017
BACPI	0.882 ± 0.004	0.840 ± 0.009	0.810 ± 0.015	0.815 ± 0.016
DrugBAN	**0.893 ± 0.003**	0.839 ± 0.007	0.825 ± 0.012	0.820 ± 0.014
Our Model	0.885 ± 0.002	**0.856 ± 0.004**	**0.828 ± 0.020**	0.787 ± 0.029
Dataset 2: DAVIS				
LR	0.645 ± 0.010	0.192 ± 0.023	0.529 ± 0.051	0.692 ± 0.033
DNN	0.778 ± 0.008	0.232 ± 0.022	0.688 ± 0.041	0.774 ± 0.034
GNN-CPI	0.756 ± 0.011	0.242 ± 0.018	0.626 ± 0.042	0.758 ± 0.035
DeepDTI	0.775 ± 0.002	0.208 ± 0.005	0.676 ± 0.014	0.767 ± 0.011
DeepDTA	0.792 ± 0.007	0.272 ± 0.040	0.688 ± 0.041	0.778 ± 0.018
MolTrans	0.816 ± 0.002	0.363 ± 0.014	0.720 ± 0.020	0.789 ± 0.012
HyperAttentionDTI	0.841 ± 0.080	0.483 ± 0.010	0.708 ± 0.056	0.722 ± 0.014
BACPI	**0.843 ± 0.079**	0.478 ± 0.008	0.824 ± 0.053	0.810 ± 0.013
DrugBAN	0.836 ± 0.078	0.484 ± 0.008	0.871 ± 0.050	0.769 ± 0.012
Our Model	0.799 ± 0.092	**0.515 ± 0.008**	**0.896 ± 0.074**	**0.850 ± 0.010**
Dataset 3: BindingDB				
LR	0.657 ± 0.002	0.451 ± 0.015	0.635 ± 0.013	0.736 ± 0.011
DNN	0.889 ± 0.002	0.581 ± 0.013	0.743 ± 0.024	0.898 ± 0.017
GNN-CPI	0.877 ± 0.003	0.547 ± 0.013	0.720 ± 0.013	0.888 ± 0.010
DeepDTI	0.827 ± 0.002	0.407 ± 0.004	0.619 ± 0.022	0.872 ± 0.021
DeepDTA	0.902 ± 0.002	0.604 ± 0.011	0.757 ± 0.032	0.904 ± 0.014
MolTrans	0.905 ± 0.001	0.601 ± 0.006	0.781 ± 0.004	0.883 ± 0.006
HyperAttentionDTI	0.900 ± 0.004	**0.876 ± 0.009**	0.801 ± 0.032	0.791 ± 0.016
BACPI	0.907 ± 0.002	0.872 ± 0.007	0.811 ± 0.019	0.911 ± 0.015
DrugBAN	**0.912 ± 0.003**	0.868 ± 0.006	0.822 ± 0.025	**0.912 ± 0.016**
Our Model	0.910 ± 0.006	0.659 ± 0.008	**0.857 ± 0.049**	0.818 ± 0.046

Note: The bolded numbers indicate the best performance, and underlined numbers represent the second-best performance.

3.2.1. Experimental results with baselines

Table 3 illustrates the performance results of our BERT-based model and other baseline methods on the BIOSNAP, DAVIS, and BindingDB datasets. Our BERT-based model outperforms all others in terms of sensitivity, achieving a score of 89.6% on the DAVIS dataset, 2.5% higher than the similar-task DrugBAN model, and surpassing 82% on the BIOSNAP and BindingDB datasets. These results attest to the superior capability of our BERT-based model in predicting positive associations between drugs and proteins.

Additionally, our model demonstrated superior ability in terms of PR-AUC for the DAVIS and BIOSNAP datasets, outperforming MolTrans by 15.2% on the DAVIS dataset. Even though MolTrans also utilizes subsequence representation, it lacks the common DTI features in the pretraining stage, rendering it less effective than our BERT-based model on certain datasets.

The overall results highlight that the two primary methods that incorporate the attention mechanism and protein subsequence representation (DrugBAN and BERT-based models) consistently achieve higher overall metric scores across all datasets. This indicates that subsequence embedding can provide more detailed information about interactions compared to pure amino acid sequences, thus enhancing the DTI prediction task. The graph representation for protein is another further direction for optimize the DTI prediction.

However, our model did not outperform comparison methods in some metrics (ROC-AUC and Specificity). Upon analyzing the structure of the comparison methods, it appears that the CNN structure significantly aids in feature extraction, and thus, its integration might be considered in next studies.

3.2.2. Ablation study

Considering the significant role of transfer learning in our model, we conducted an ablation study to examine the effect of pretraining.

Table 4
Experimental results of importance of pretraining on datasets.

Training strategy	ROC-AUC	PR-AUC	Sensitivity	Specificity	Accuracy
Dataset 1: BIOSNAP					
w/o pre-training	0.8607	0.8407	0.8090	0.7631	0.8062
with pre-training	0.8849	0.8562	0.8282	0.7866	0.8249
Dataset 2: DAVIS					
w/o pre-training	0.7096	0.4618	0.8740	0.8488	0.8613
with pre-training	0.7986	0.5147	0.8959	0.8502	0.8749
Dataset 3: BindingDB					
w/o pre-training	0.8895	0.6176	0.7559	0.8418	0.7988
with pre-training	0.9102	0.6589	0.8566	0.8179	0.8365

Specifically, we depicted the ROC-AUC and PR-AUC evaluation metrics during the training process in Fig. 4.

The results reveal the training process of the two models on different datasets, indicating that the model with pretraining exhibits a faster and more stable convergence rate throughout the process. This is attributable to pretraining aiding the model in learning the features of drug and protein sequences prior in the pretraining phase, thereby enabling quicker convergence. Moreover, to validate the performance of the final trained model, we compared the prediction results of both models on the test set across three datasets, as outlined in Table 4. The results demonstrate that our model significantly outperforms the model without pretraining, suggesting that transfer learning has effectively learned features beneficial for DTI prediction.

It is important to highlight that, despite the pretraining, the specificity of dataset 3 is observed to be weaker compared to the model without pretraining. This discrepancy primarily come from utilizing the BindingDB dataset for both pretraining and fine-tuning. Addressing this aspect is a priority for future optimization efforts. Upon thorough analysis, we identified a potential source of the issue: an imbalance

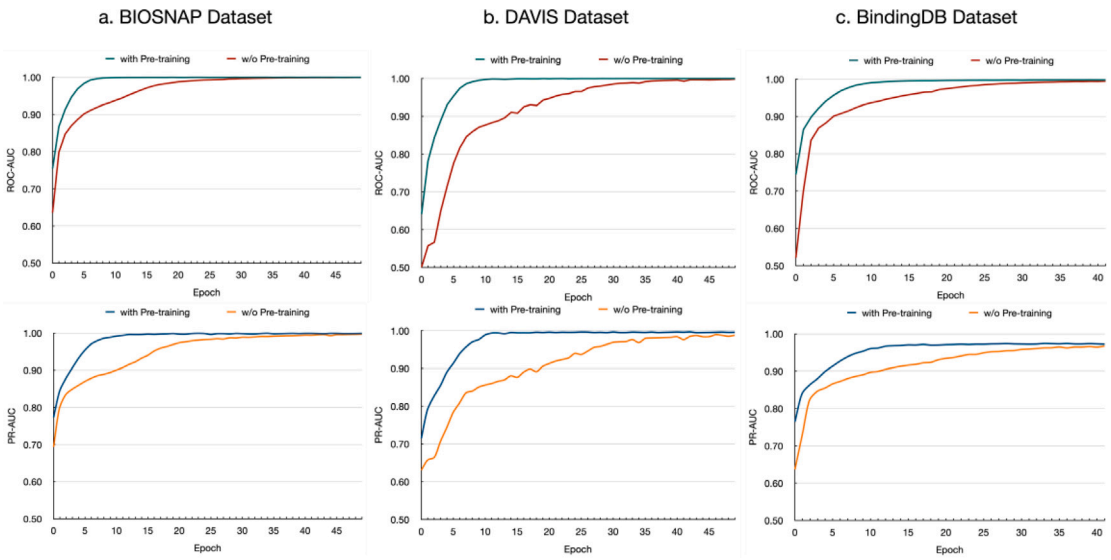

Fig. 4. Comparison of ROC-AUC and PR-AUC variation during training.

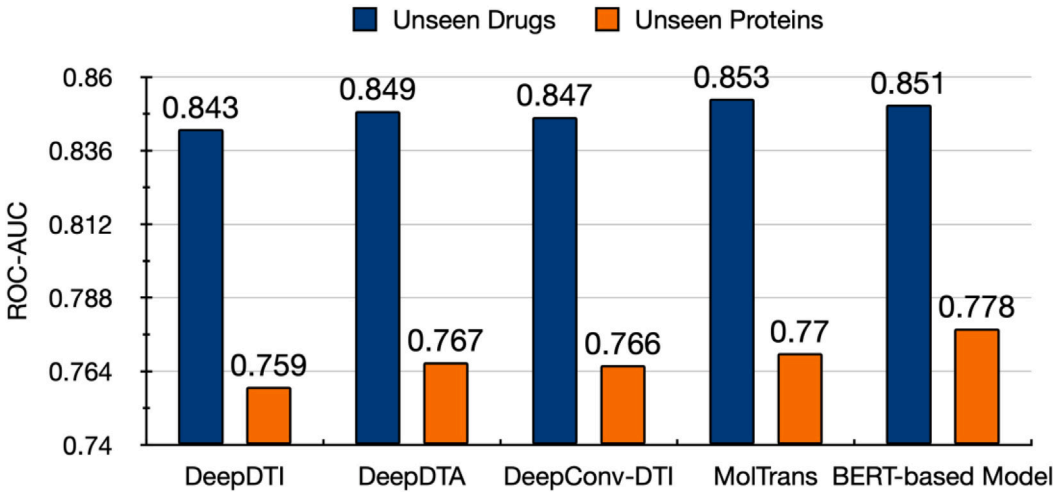

Fig. 5. The prediction results on unseen drugs and unseen proteins.

in negative samples during the fine-tuning stage. Specifically, there is an overrepresentation of protein or drug entries identical to the positive samples in the pretraining stage. Consequently, during fine-tuning, the model might erroneously classify numerous negative DTI pairs as positive, resulting in decreased specificity when compared to a model without pretraining. This insight informs our focus for refining the model's performance in subsequent work.

This study thus establishes that transfer learning can enhance the contextual relevance of proteins or compounds to their labels, even when learned exclusively from unlabeled datasets.

3.2.3. Scalability study

To evaluate the model's performance on unknown samples, we conducted additional studies on unseen drugs and unseen proteins (dataset sourced from Huang et al. (2021)). The results, as shown in Fig. 5, indicate that our BERT-based model exhibits a reliable scalability capability, as observed from the ROC-AUC scores.

Even though our BERT-based model did not secure the top performance for unseen drugs, it outscored DeepDTI, DeepDTA, and DeepConv-DTI. Remarkably, the BERT-based model achieved the highest ROC-AUC for unseen proteins.

Although our approach does not significantly outperform the baselines, we have examined the existing methods for unseen drugs and unseen proteins at this stage. We found that for both datasets, the performance improvement of various models is not substantial. For example, a multi-modal transformer network (Kroll et al., 2023) and NHGNN-DTA (He et al., 2023).

We need to conduct more in-depth studies on the protein and drug features in unseen drugs and unseen proteins. It would be beneficial to consider incorporating protein datasets with similar functionalities as part of our pretraining dataset. This approach could potentially enhance the model's generalization ability further.

In other hand, large-scale BERT models are built upon extensive and diverse datasets, and high-quality and complex pretraining of BERT has been shown to effectively improve model prediction performance. Upon analyzing our own pretraining data, we found that our dataset, sourced partially from BindingDB, lacks sufficient complexity and richness. Therefore, based on our research, we aim to obtain additional categories of DTI data from DrugBank and PDBSum. We will organize and combine them with BindingDB to create a new pretraining dataset, thereby enhancing the complexity of the entire pretraining data and

improving the model's generalization capabilities. Ultimately, this will lead to improved prediction performance for unseen drugs and proteins.

4. Conclusion

In this study, we proposed a BERT-based model for DTI prediction tasks, incorporating subsequence representation and pretraining on large datasets. Owing to the remote feature extraction ability of the BERT model, it can capture interaction features between molecular SMILES subsequences and protein subsequences after their concatenation. In essence, the BERT-based approach provides a transfer learning model capable of learning binding features from vast drug-target interaction datasets. Compared with other methods, the BERT-based model allows for the discovery of more DTI-related subsequences attention features of proteins and drugs through transfer learning.

Through extensive experimentation on three datasets, the model demonstrated state-of-the-art performance. Furthermore, the ablation study showed how the pretraining stage aids the model in understanding interaction features more comprehensively. Additionally, the BERT-based approach presented a promising method for identifying potential candidate drugs for viral proteins, such as those of COVID-19 and Monkeypox.

Looking ahead, we intend to focus on the following topics:

- The relationship between enzymes and compounds in metabolic reactions has long been a central theme of research. For certain potential biochemical reactions, an appropriate enzyme needs to be identified for catalysis. Our BERT-based model offers a possibility to mine for candidate enzymes based on the reactions.
- During the binding of a drug to a target, certain existing bonds break while new ones form. To better characterize this reaction process, we plan to incorporate features into our model that specifically account for bond changes in the reactions.

CRediT authorship contribution statement

Zhihui Yang: Methodology, Software, Writing – original draft, Writing – review & editing. **Juan Liu:** Conceptualization, Funding acquisition, Supervision. **Feng Yang:** Data curation. **Xiaolei Zhang:** Data curation. **Qiang Zhang:** Data curation. **Xuekai Zhu:** Software. **Peng Jiang:** Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Key R and D Program of China (No. 2019YFA0904303)

References

Bai, P., Miljković, F., John, B., Lu, H., 2023. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nat. Mach. Intell. 5 (2), 126–136.

Bai, F., Morcos, F., Cheng, R.R., Jiang, H., Onuchic, J.N., 2016. Elucidating the druggable interface of protein- protein interactions using fragment docking and coevolutionary analysis. Proc. Natl. Acad. Sci. 113 (50), E8051–E8058.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M., 2022. ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics 38 (8), 2102–2110.

Broach, J.R., Thorner, J., 1996. High-throughput screening for drug discovery. Nature 384 (6604 Suppl), 14–16.

Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., 2013. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29 (7), 960–962.

Chen, X., Liu, M.-X., Yan, G.-Y., 2012. Drug–target interaction prediction by random walk on the heterogeneous network. Mol. Biosyst. 8 (7), 1970–1978.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y., 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 8 (5), e1002503.

Chithrananda, S., Grand, G., Ramsundar, B., 2020. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv: 2010.09885.

Consortium, U., 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47 (D1), D506–D515.

Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Zarrinkar, P.P., 2011. Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnol. 29 (11), 1046–1051.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al., 2020. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225.

Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D., 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. 15 (5), 411–428.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40 (D1), D1100–D1107.

Hajduk, P.J., Greer, J., 2007. A decade of fragment-based drug design: strategic advances and lessons learned. Nat. Rev. Drug Discov. 6 (3), 211–219.

He, H., Chen, G., Chen, C.Y.-C., 2023. NHGNN-DTA: A node-adaptive hybrid graph neural network for interpretable drug-target binding affinity prediction. Bioinformatics btad355.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., Ester, M., 2017. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J. Cheminform. 9 (1), 1–14.

Huang, K., Xiao, C., Glass, L.M., Sun, J., 2021. MolTrans: Molecular interaction transformer for drug–target interaction prediction. Bioinformatics 37 (6), 830–836.

Hudkins, R.L., Diebold, J.L., Tao, M., Josef, K.A., Park, C.H., Angeles, T.S., Aimone, L.D., Husten, J., Ator, M.A., Meyer, S.L., et al., 2008. Mixed-lineage kinase 1 and mixed-lineage kinase 3 subtype-selective dihydronaphthyl [3, 4-a] pyrrolo [3, 4-c] carbazole-5-ones: optimization, mixed-lineage kinase 1 crystallography, and oral in vivo activity in 1-methyl-4-phenyltetrahydropyridine models. J. Med. Chem. 51 (18), 5680–5689.

Karimi, M., Wu, D., Wang, Z., Shen, Y., 2019. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics 35 (18), 3329–3338.

Kroll, A., Ranjan, S., Lercher, M.J., 2023. Drug-target interaction prediction using a multi-modal transformer network demonstrates high generalizability to unseen proteins. bioRxiv. https://www.biorxiv.org/content/10.1101/2023.08.21.554147v2.

Lee, I., Keum, J., Nam, H., 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput. Biol. 15 (6), e1007129.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.

Li, M., Lu, Z., Wu, Y., Li, Y., 2022. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics 38 (7), 1995–2002.

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res. 35 (suppl_1), D198–D201.

Nguyen, T., Le, H., Quinn, T.P., Nguyen, T., Le, T.D., Venkatesh, S., 2021. GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics 37 (8), 1140–1147.

Österberg, F., Morris, G.M., Sanner, M.F., Olson, A.J., Goodsell, D.S., 2002. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins: Struct. Funct. Bioinform. 46 (1), 34–40.

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

Pan, X., Lin, X., Cao, D., Zeng, X., Yu, P.S., He, L., Nussinov, R., Cheng, F., 2022. Deep learning for drug repurposing: Methods, databases, and applications. Wiley Interdiscip. Rev.: Comput. Mol. Sci. e1597.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50 (5), 742–754.

Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., et al., 2017. A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. 16 (1), 19–34.

Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., McCammon, J.A., 2004. Discovery of a novel binding trench in HIV integrase. J. Med. Chem. 47 (8), 1879–1881.

Sennrich, R., Haddow, B., Birch, A., 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Sethi, A., Joshi, K., Sasikala, K., Alvala, M., 2019. Molecular docking in modern drug discovery: Principles and recent applications. Drug Discov. Dev.-New Adv. 2, 1–21.

Tsubaki, M., Tomii, K., Sese, J., 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35 (2), 309–318.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., Lu, H., 2017. Deep-learning-based drug–target interaction prediction. J. Proteome Res. 16 (4), 1401–1409.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al., 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46 (D1), D1074–D1082.

Yang, Z., Liu, J., Zhu, X., Yang, F., Zhang, Q., Shah, H.A., 2022. FragDPI: A novel drug–protein interaction prediction model based on fragment understanding and unified coding. Front. Comput. Sci..

Yuriev, E., Agostino, M., Ramsland, P.A., 2011. Challenges and advances in computational docking: 2009 in review. J. Mol. Recognit. 24 (2), 149–164.

Zanni, R., Galvez-Llompart, M., Galvez, J., Garcia-Domenech, R., 2014. QSAR multi-target in drug discovery: a review. Curr. Comput.-Aided Drug Des. 10 (2), 129–136.

Zhang, X.-C., Wu, C.-K., Yang, Z.-J., Wu, Z.-X., Yi, J.-C., Hsieh, C.-Y., Hou, T.-J., Cao, D.-S., 2021. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. Brief. Bioinform. 22 (6), bbab152.

Zhao, Q., Zhao, H., Zheng, K., Wang, J., 2022. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 38 (3), 655–662.

Zitnik, M., Sosic, R., Leskovec, J., 2018. BioSNAP datasets: Stanford biomedical network dataset collection.