SMT-DTA: Improving Drug-Target Affinity Prediction with Semi-supervised Multi-task Training

Qizhi Pei[1][*], Lijun Wu[2][†], Jinhua Zhu[3], Yingce Xia[2], Shufang Xie[2],
Tao Qin[2], Haiguang Liu[2], Tie-Yan Liu[2]
[1]University of Science and Technology of China; [2]Microsoft Research Asia;
[3]CAS Key Laboratory of GIPAS, University of Science and Technology of China;
`{peiqz, teslazhu}@mail.ustc.edu.cn`
`{lijuwu, yinxia, shufxi, taoqin,tyliu}@microsoft.com`

Abstract

Drug-Target Affinity (DTA) prediction is an essential task for drug discovery and pharmaceutical research. Accurate predictions of DTA can greatly benefit the design of new drug. As wet experiments are costly and time consuming, the supervised data for DTA prediction is extremely limited. This seriously hinders the application of deep learning based methods, which require a large scale of supervised data. To address this challenge and improve the DTA prediction accuracy, we propose a framework with several simple yet effective strategies in this work: (1) a multi-task training strategy, which takes the DTA prediction and the masked language modeling (MLM) task on the paired drug-target dataset; (2) a semi-supervised training method to empower the drug and target representation learning by leveraging large-scale unpaired molecules and proteins in training, which differs from previous pre-training and fine-tuning methods that only utilize molecules or proteins in pre-training; and (3) a cross-attention module to enhance the interaction between drug and target representation. Extensive experiments are conducted on three real-world benchmark datasets: BindingDB, DAVIS and KIBA. The results show that our framework significantly outperforms existing methods and achieves state-of-the-art performances, e.g., 0.712 RMSE on BindingDB IC$_{50}$ measurement with more than 5% improvement than previous best work. In addition, case studies on specific drug-target binding activities, drug feature visualizations, and real-world applications demonstrate the great potential of our work. The code and data are released at `https://github.com/QizhiPei/SMT-DTA`.

1 Introduction

With the rapid development of healthcare management, drug related research is becoming more and more important, and there are various applications with practical value, such as Drug-Drug Interaction (DDI) [61], drug repurposing [57], drug synergy prediction [27], Drug-Target Interaction (DTI) [8] and so on. Among these tasks, Drug-Target Affinity (DTA) prediction is one with crucial importance in drug research [34, 56], which aims to accurately predict the binding effect of a drug to a protein target. Computational methods, such as molecular docking [67, 71] and molecular dynamics simulations [63, 42], are representative approaches for DTA prediction with satisfied accuracy, but these methods are not efficient with high computational cost.

In recent years, deep learning based approaches have been applied to DTA prediction due to their efficiency and high accuracy, and good progress has been made [68, 19, 28, 26, 7]. Though deep

[*] This work is conducted at Microsoft Research Asia.

[†] Corresponding author.

learning models are effective, their success highly depends on the availability of a large-scale of supervised training data. Unfortunately, the labeled binding affinities for the recognized targets and drugs are in limited size since it is costly and time-consuming to collect labeled data through wet experiments. For example, the widely used DTA dataset, BindingDB [41], contains about $2M$ binding data with drug-target pairs, which seem to be a huge amount. However, as pointed out in [28], the dataset is of low quality and the affinity labels for same compound-protein pairs in the dataset are noisy due to different reasons (e.g., variations in experimental conditions, different data sources, etc). Following the same data filtration protocols as in [28], the remained clean data contains about $200k$, which is relatively small compared with datasets in NLP/CV fields (e.g., millions to billions) where deep learning has achieved great success. This limited scale of labeled data becomes a blocking issue for deep learning based DTA prediction.

To address this issue, in this work, we propose a Semi-supervised Multi-task Training (SMT) framework with three strategies to improve DTA prediction. Specifically, (1) we conduct a multi-task training on labeled drug-target pairs to better exploit the labeled data: in addition to the DTA prediction task, we incorporate masked language modeling (MLM) training task [14] on both drug and target, which can effectively improve the accuracy of DTA prediction. (2) In addition to the labeled pair data, inspired by the success of self-supervised learning [43, 14], we leverage large-scale unlabeled molecule and protein data to help with drug and target representation learning. We find that common pre-training (on unlabeled data) and fine-tuning (on supervised data) paradigm can not work well in DTA prediction, since the separate pre-training on unlabeled molecule and protein ignores the importance of interaction between paired drug and target when fine-tuning. Therefore, we propose a different method to mix unlabeled data together with paired data and perform semi-supervised training for DTA prediction. (3) To further improve the DTA prediction accuracy, we introduce an efficient cross-attention module that differs from existing methods, which explicitly captures the interaction information between drugs and targets, therefore benefiting the interaction prediction.

To evaluate SMT-DTA framework, we conduct experiments on several benchmarks, including the BindindDB [41], DAVIS [13], and KIBA [66]. Results show that SMT-DTA can significantly reduce the error in DTA prediction. For example, the root-mean-square-error (RMSE) on BindingDB IC_{50} dataset is reduced to near 0.712, which is about 5% improvement over existing state-of-the-art methods. Besides, we provide extensive case studies of the drug-target binding activities and we show our model can correctly identify important atomic groups and amino acids among binding sites. Furthermore, we visualize the drug features learned by our model and find it has good ability to group the drugs by their corresponding targets. Moreover, we also conduct real-world applications to detect targets for some drugs and the results show our SMT-DTA is with good generalization capability. These results can demonstrate the great potential of the SMT framework in DTA research.

The main contributions of our work are as follows:

- We propose a SMT-DTA framework with three strategies for DTA prediction to alleviate the data limitation issue: multi-task training with both DTA prediction and masked language modeling, semi-supervised learning with both labeled drug-target pairs and unlabeled molecules and proteins, and an efficient cross-attention module.
- We demonstrate the state-of-the-art performance of the SMT framework in DTA research by conducting experiments on DTA benchmark datasets.
- We compare different strategies to leverage unlabeled data and observe interesting findings, which can benefit the future direction of DTA prediction. For example, separate pre-training on unlabeled molecules and proteins can not benefit the interactions between drugs and targets in the fine-tuning stage.
- We provide extensive studies and analysis experiments to show the potential of our approach. For example, our model can well capture the structural knowledge with accurate binding information between drugs and targets.

2 Related Work

2.1 Drug-Target Interaction/Affinity Prediction

DTI/DTA is important in drug research and drug discovery. The prediction methods can be classified into structure-based and the structure-free approaches. For structure-based methods, the most widely adopted way is the molecular docking [67, 71, 33, 45] and molecular dynamics (MD) simulations [63, 42]. Docking-based methods predict the binding affinity given the 3D structure inputs of a drug compound and a protein, which utilize predefined force fields to estimate the binding affinity at the atomic level. However, the strong dependency of high-quality 3D structures leads to a major limitation to the cases without 3D structures of either drugs or targets. Structure-free predictions are developed to overcome this severe shortcoming. Similarity-based methods take the calculation of similarity metrics as descriptors for both drugs and targets to predict DTA [11, 2], such as KronRLS [55] and SimBoost [23]. Recently, with the rapid development of deep learning, deep learning based methods are promising to well exploit the local features of both molecule structures and protein sequences to predict DTA. There are many works proposed based on the deep learning models, such as DeepAffinity [28], MONN [38], and others [68, 19, 54, 64, 79, 36]. DeepAfftinity [28] is a semi-supervised method that utilizes the recurrent and convolutional neural networks to exploit the labeled and unlabeled data for predicting binding affinity. MONN [38] utilized a multi-objective training framework for both binding affinity and drug-target interaction matrix predictions.

2.2 Molecule and Protein Pre-training

Pre-training methods are widely exploited to learn effective and compact representations for molecule and protein recently, which take the spirit of pre-training and fine-tuning methodology from natural language processing (NLP) [43, 14] and image processing [35, 15, 21]. For molecule pre-training methods, there are two rough types based on molecule representations: the sequential model pre-training [18, 24, 46] based on the simplified molecular-input line-entry system (SMILES) [76] strings and the graph neural network (GNN) pre-training [75, 25, 40, 62] based on the molecule graph structures. The SMILES based pre-training typically takes the Transformer network, motivated by its excellent performances demonstrated on related fields, such as SMILES-BERT [74] and Chemberta [10] that adopt the masked language modeling (MLM) objective for pre-training. Different GNN models are adopted for graph structure pre-training, and the pre-training tasks include masked training that performed on different graph parts, such as N-gram [40], AttrMasking [25], ContextPredict [25], MotifPredict [62], subgraph prediction [37], and also the contrastive learning based methods [65, 58]. Protein pre-training depends on the amino acid protein sequences and it is implemented similar to text pre-training in NLP, which take the Transformer encoder for pre-training [5, 22, 16, 48]. Representative works include TAPE [59] and ESM [60]. TAPE introduced the first Transformer based protein pre-training model and downstream benchmark for evaluating the pre-trained models. ESM conducted the pre-training on different protein database to evaluate the effects of the protein data quality, and the training was performed on a much larger Transformer with 34 layers.

3 Methods

3.1 Preliminary

Before introducing the details, we first give the definition of the DTA problem and the necessary notations used in our work. Given a DTA dataset $\mathcal{DT} = \{(D, T, y)_i\}_{i=1}^{N}$, where (D, T, y) is the triplet of DTA data sample, N is the sample size, D is one drug from a drug (molecule) dataset, T is one target from a target (protein) dataset, and label y (a floating number) is the binding affinity number for the drug-target pair (D, T). The DTA prediction task is then a regression task to predict the binding affinity score between the drug and target pair. The goal is to learn a mapping function $\mathcal{F} : D \times T \rightarrow y$. We take the SMILES string as the representation of drugs, which is a sequence resulted by traversing the molecule graph using depth-first search and some specific rules. Specifically, for a drug D, the formulation is $D = \{d_i\}_{i=1}^{|D|}$, where $|D|$ is the length of SMILES and d_i is the token in the string. For a target T, we use its FASTA sequence representation that contains of amino acid tokens, where $T = \{t_i\}_{i=1}^{|T|}$, $|T|$ is the length and t_i is the amino acid token. Due to the sequence based representations for drug and target, we use Transformer [69] models $\mathcal{M}_{\mathcal{D}}$ and $\mathcal{M}_{\mathcal{T}}$ to encode drug and target representations. Other notations will be introduced in the following sections.

3.2 Multi-task Training with Masked Language Modeling

We first introduce our multi-task training strategy. Since we are targeting at alleviating the data limitation issue of DTA prediction, multi-task training [39, 49, 17] is a preferred method that widely adopted. Different from the conventional multi-task training method that each task takes its own dataset, we utilize the original paired drug-target data for multi-task training. To be specific, inspired from the general self-supervised methods that take the masked language modeling (MLM) objective for pre-training [14, 43], we take the same spirit in our DTA prediction. Besides the DTA prediction, we introduce the MLM training upon the paired drug-target data to form multi-task training.

Concretely, for a drug $D = \{d_1, d_2, .., d_{|D|}\}$ and a target $T = \{t_1, t_2, ..., t_{|T|}\}$, we randomly replace some of the tokens d_i, t_i in the sequence by a special `[MASK]` token with some mask ratios [14]. The masked sequences are then denoted as D' and T', and our goal is to reconstruct these masked tokens. Mathematically, the MLM training objectives for drug and target sequences are:

$$\mathcal{L}^{D}_{MLM} = -\sum_{m=1}^{M_D} \log P(d_m \| D'), \quad \mathcal{L}^{T}_{MLM} = -\sum_{m=1}^{M_T} \log P(t_m \| T'), \tag{1}$$

where d_m, t_m are the masked tokens, and M_D, M_T are the corresponding masked token numbers in drug and target sequences. During training, the MLM training objectives are then jointly optimized with the DTA prediction to have better representations for drugs and targets. In this way, though we do not incorporate more labeled DTA data, we maximize the potential of the original DTA data.

3.3 Semi-supervised Training with Large-scale Unlabeled Molecule and Protein Data

Recently, pre-training on large-scale unlabeled data in a self-supervised way has been promising and its success has been witnessed in natural language processing [14, 43, 12], image processing [6, 4], video processing [73, 44] and so on. Specific to chemistry and biology domain, there are also some great works which achieve promising results with pre-training, such as Chemberta [10] pre-trained on molecule SMILES, AttrMasking [25] pre-trained on molecule graph, and ESM [60] model that pre-trained on protein sequences. In this work, we also leverage the large-scale unlabeled molecules/proteins for representation enhancement to overcome the limited DTA data issue.

Different from the widely adopted pre-training (on unlabeled data) and fine-tuning (on supervised data) strategy [14] in previous works [10, 43], we propose to train the tasks on unlabeled data and supervised data together and perform semi-supervised learning. One major problem of existing methods is that the separately pre-training on molecule and protein only can not well capture the interaction between drug and target, while the interaction design is the key to DTA prediction. This leads to the poor performance of pre-training and fine-tuning method for DTA prediction (see Section 5.5). Therefore, we propose a different semi-supervised training method. Specifically, same as the above multi-task training on paired drug-target data samples, we also adopt the MLM training on the unlabeled molecules and proteins in a semi-supervised multi-task training. That is, both the drugs and targets in the paired data and the unlabeled molecules and proteins are incorporated with the MLM training objective. Denote the unlabeled molecule and protein dataset as $\mathcal{M} = \{M_i\}_{i=1}^{|\mathcal{M}|}$ and $\mathcal{P} = \{P_i\}_{i=1}^{|\mathcal{P}|}$ respectively, where M_i is the molecule sequence and P_i is the protein sequence. For each molecule sequence $M = \{m_j\}_{j=1}^{|M|}$ and protein sequence $P = \{p_j\}_{j=1}^{|P|}$, we also take MLM training. The objectives are the same as Eqn. (1),

$$\mathcal{L}^{M}_{MLM} = -\sum_{s=1}^{S_M} \log P(m_s \| M'), \quad \mathcal{L}^{P}_{MLM} = -\sum_{s=1}^{S_P} \log P(p_s \| P'), \tag{2}$$

where m_s, p_s are the masked tokens and M', P' are the sequences with masked tokens, S_M and S_P are the number of masked tokens. We put discussions about our joint training strategy to leverage the unlabeled data and the common pre-training and fine-tuning strategy in Section 3.6.

3.4 Interaction with Cross-attention Module

The goal of DTA prediction is to predict the binding affinity score caused by the *interaction* between drug and target. Therefore, it is crucial to well model the interaction in a good way. However, previous works [26, 64] usually directly take the drug and target as input and concatenate their encoded representations to feed for the regression prediction, which ignores the importance of special interaction module design. Recently, [38] and [28] pay attention to the interaction design and introduce the attention mechanism [52] into the interaction module. They both present the joint

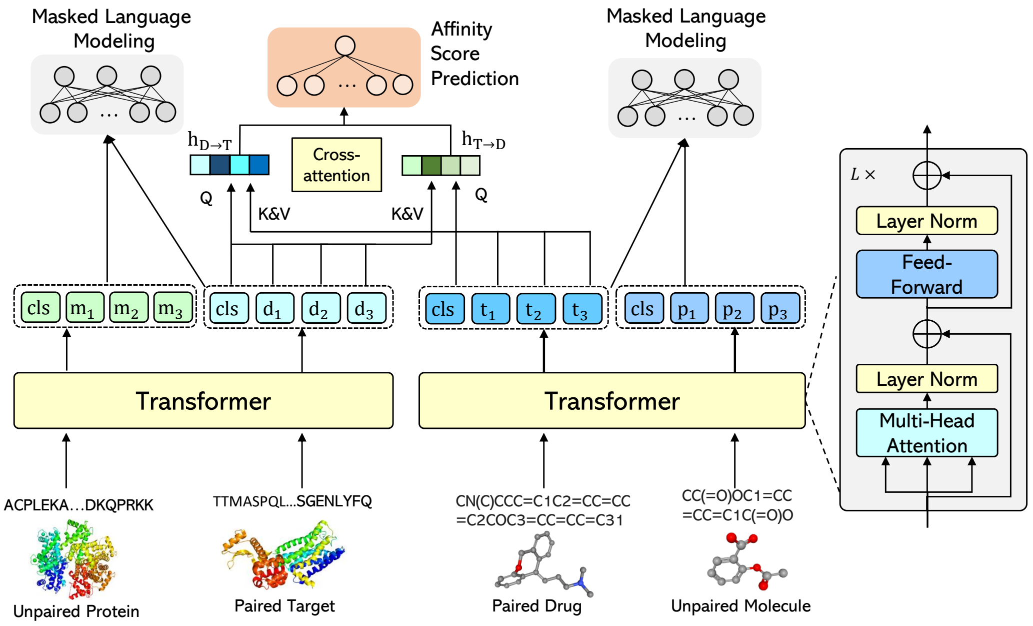


Figure 1: The overall framework of our SMT-DTA. The two Transformer models are encoding the molecule/drug and protein/target sequences. The MLM (Masked Language Modeling) training is conducted on both unlabeled (unpaired) data and paired data. The cross-attention module is used between drug-target paired data, where Q is query, K&V are key and value.

attention mechanism that performs on each token of the drug and target sequences, which leads to a pairwise interaction attention matrix over all the tokens.

We introduce a different cross-attention module in this paper, which is much simpler than the pairwise interaction mechanism. Before feeding the paired drug and target sequences into the Transformer models, we first add a special `[CLS]` token at the beginning of the sequences. Specifically, the drug D and target T are formed as $D = \{\texttt{[CLS]}_D, d_1, d_2, ..., d_{|D|}\}$ and $T = \{\texttt{[CLS]}_T, t_1, t_2, ..., t_{|T|}\}$. After that, D and T are encoded by the Transformer models to get the hidden representations $H_D = [h_{\texttt{[CLS]}_D}, h_{d_1}, h_{d_2}, ..., h_{d_{|D|}}]$ and $H_T = [h_{\texttt{[CLS]}_T}, h_{t_1}, h_{t_2}, ..., h_{t_{|T|}}]$. To calculate our cross-attention, we only take the $h_{\texttt{[CLS]}_D}$ and $h_{\texttt{[CLS]}_T}$ out of H_D and H_T to represent the drug D and target T. This is similar to the pre-training methods [14, 43] that use the `[CLS]` token to represent the whole sentence. Then cross-attention is performed over the `[CLS]` tokens and the corresponding sequences. Specifically, during the attention calculation, the query is the $h_{\texttt{[CLS]}_D}$ or $h_{\texttt{[CLS]}_T}$, and the key and value are H_T or H_D. Mathematically, the cross-attention mechanism is as follow:

$$\begin{aligned} h_{D\to T} &= \texttt{Attn}_{D\to T} = \texttt{softmax}(\frac{(h_{\texttt{[CLS]}_D}W_1)(H_T W_2)^{\mathrm{T}}}{\sqrt{d}})(H_T W_3), \\ h_{T\to D} &= \texttt{Attn}_{T\to D} = \texttt{softmax}(\frac{(h_{\texttt{[CLS]}_T}W_4)(H_D W_5)^{\mathrm{T}}}{\sqrt{d}})(H_D W_6), \end{aligned} \tag{3}$$

where `Attn` and `softmax` are attention function and softmax operation, d is the hidden dimension, and Ws are parameter matrices. After the cross attention, the attended representations are then concatenated for DTA prediction.

Compared with pairwise interaction mechanism, one advantage of our cross-attention mechanism is that our method takes less computational cost. Concretely, suppose the drug sequence length is $|D|$, the target sequence length is $|T|$, then our cross-attention costs computation about $O(|D| \times d)$ and $O(|T| \times d)$, where the former is for attention $\texttt{Attn}_{T\to D}$ and the latter is for attention $\texttt{Attn}_{D\to T}$, and the total is then $O((|D| + |T|) \times d)$. Here d is the dimension of hidden state. For the pairwise attention, the computations are $O(|D| \times |T| \times d)$ since the tokens in one sequence are attending to all the tokens in the other one, which costs much than our $O((|D| + |T|) \times d)$. In the experiments, we have shown that our cross-attention module is not only efficient but also very effective.

3.5 Overall Framework

The overall SMT-DTA framework is shown in Fig 1 and the proposed components are all presented in the figure. Our training data contains three parts, the paired drug-target data, the unlabeled molecule data and the unlabeled protein sequences. The molecule and drug sequences are processed by one Transformer [69] model, and the protein and target sequences are processed by another Transformer model. For the paired data (D, T), they perform regression task of DTA prediction and also MLM training. After encoding the drug and target sequences and obtaining the attended representations $h_{D\to T}$ and $h_{T\to D}$ through our cross-attention module, they are concatenated together and the resulted representation is used for final affinity score prediction, which are as follows:

$$H = \texttt{Concat}(h_{D\to T}, h_{T\to D}), \quad y' = \texttt{MLP}(H), \tag{4}$$

where `Concat` and `MLP` are the concatenation and linear transformation layer, y' is the predicted binding affinity score. The loss function is the mean squared error (MSE):

$$\mathcal{L}_{MSE} = (y - y')^2. \tag{5}$$

For the unlabeled molecule data M and protein data P, the loss functions are MLM training objective $\mathcal{L}^M_{MLM}, \mathcal{L}^P_{MLM}$ as shown in Eqn. (2). Then together with the MLM loss $\mathcal{L}^D_{MLM}, \mathcal{L}^T_{MLM}$ on the paired drug-target as in Eqn. (1), the final training objective is to minimize the following term,

$$\mathcal{L} = \mathcal{L}_{MSE} + \alpha * (\mathcal{L}^D_{MLM} + \mathcal{L}^T_{MLM}) + \beta * (\mathcal{L}^M_{MLM} + \mathcal{L}^P_{MLM}), \tag{6}$$

where α and β are the coefficients to control the weights of different losses. For simplicity, we set α to be the same as β.

3.6 Discussion

One can find that in our work, the way to leverage the large-scale unlabeled molecule and protein data is different from the conventional pre-training methods. The commonly adopted method is to first pre-train the Transformer models with the unlabeled data only. After pre-training, the Transformer models are used as either (1) good initialization for downstream fine-tuning (fine-tuning based utilization) or (2) good feature extractors for downstream model training (feature based utilization). These two methods all separate the pre-training and the downstream task training into two stages, while our strategy proposed in this paper is to jointly train the pre-training task and the downstream task in a multi-task framework. The difference of our strategy is clear, but the advantages and disadvantages are also obvious. Our training method can learn both good representations from unlabeled data by MLM training and also the task-specific representations from downstream task training, which hence can avoid the catastrophic forgetting problem [32, 9] existed in pre-training/fine-tuning based method. Besides, the general strategy ignores the interaction modeling during pre-training, which leads to a huge gap between the pre-training task and fine-tuning prediction [72], while the representations learned by our method can well capture the interaction information between drug and target. However, due to the semi-supervised training with large-scale unlabeled data, the training cost is increased comparing to directly fine-tune on an already pre-trained Transformer model (if already existed). From the experimental studies (details in Section 5.5), we find that our SMT method achieves superior performances compared with the common two-stage training methods. In the future, we will focus on more efficient methods to better leverage the unlabeled data.

4 Evaluation Results of Binding Affinity Prediction

4.1 Experimental Settings

We first roughly introduce the experimental settings of binding affinity prediction, including the datasets, model, evaluation, and compared baselines.

The unlabeled molecules and proteins are from **PubChem** [29] and **Pfam** [50] datasets. We randomly sampled $10M$ molecules and proteins[3] for semi-supervised training. For supervised DTA data, we take from three widely acknowledged benchmark datasets, BindingDB, DAVIS, and KIBA.

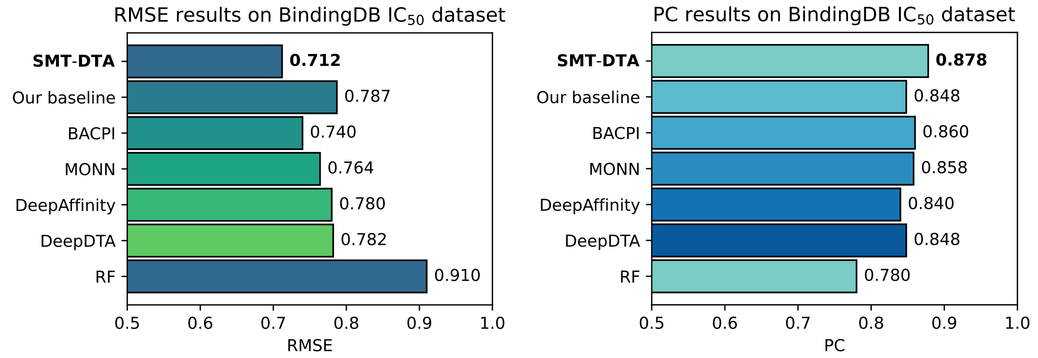
[3] We also study other sizes in Supplementary Materials.

BindingDB [41] is a database[4] of measured binding affinities, focusing on the interactions of targets with small drug-like molecules. Previous works like DeepAffinity [28], MONN [38] and BACPI [36] have been evaluated on half-maximal inhibitory concentration (IC$_{50}$) values derived from the BindingDB database. For consistency purposes, we use the same BindingDB dataset, where the dataset is randomly split to 6:1:3 as training/valid/test sets. We study on the IC$_{50}$ and K_i measurement. **DAVIS** [13] contains selectivity assays of the kinase protein family and the relevant inhibitors with their respective dissociation constant (K_d) values. **KIBA** [66] dataset includes kinase inhibitor bioactivities measured in K_i, K_d, and IC$_{50}$ metrics. DAVIS and KIBA are randomly split by 7:1:2 as train/valid/test dataset as in DeepPurpose [26].

We use two Transformer encoders for molecule encoder $\mathcal{M}_{\mathcal{D}}$ and protein encoder $\mathcal{M}_{\mathcal{T}}$, and each follows RoBERTa$_{\texttt{base}}$ architecture that consists of 12 layers. The regression prediction head is a 2-MLP layers with `tanh` activation function. As for comparison, our SMT-DTA is compared with DeepDTA [54], DeepAffinity [28], MONN [38], BACPI [36], KronRLS [55], GraphDTA [51], and DeepPurpose [26], which are representative and competitive methods in previous works. As for evaluation metrics, we use (i) mean square error (MSE), (ii) root mean square error (RMSE), (iii) pearson correlation coefficient (PC) [1], (iv) corcondance index (CI) [20] to evaluate the performance of our model on DTA regression task. Results on BindingDB dataset are evaluated on RMSE and PC, and results on DAVIS and KIBA datasets are evaluated by MSE and CI scores.

4.2 Performances on BindingDB Benchmark

We first present the results on BindingDB dataset. Since IC$_{50}$ measurement is adopted in previous works [28, 54], we mainly compare the performance on IC$_{50}$ with other strong models. The results are shown in Fig. 2, and the evaluation metrics are RMSE and PC. In the figure, 'Our baseline' refers to our implemented baseline model with two separate encoders, cross-attention module, and a feed-forward prediction layer, without the semi-supervised multi-task training. From the figure, we can see that our SMT-DTA method achieves the best performance in terms of both two metrics. For example, the RMSE is reduced from 0.787 to 0.712 with more than 7% improvement by comparing our baseline and SMT-DTA. When comparing with previous state-of-the-art models, such as MONN [38] (0.764) and BACPI [36] (0.740), our model surpasses their performances by about $3\% - 4\%$. These numbers clearly show the effectiveness of our SMT-DTA framework for binding affinity prediction.



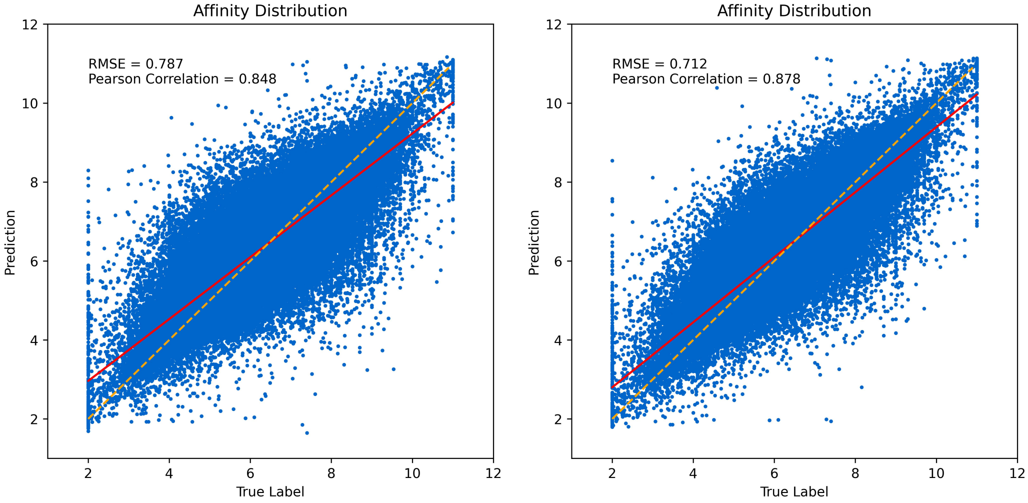
(a) RMSE results.

(b) PC results.

Figure 2: Performance evaluation of different approaches on the BindingDB IC$_{50}$ test dataset. (a) shows the RMSE evaluated results and (b) shows the PC results. The compared methods are DeepDTA [54], DeepAffinity [28], MONN [38], BACPI [36] and Random Forest [28]. 'Our baseline' refers to our implemented baseline model with cross-attention module.

Affinity Score Distribution. To get more detailed information on the performances of the prediction models, we compare the predicted values with the ground-truth affinity scores (i.e., values from the dataset) for all drug-target pairs in the BindingDB IC$_{50}$ test set. The results are shown in Fig. 3 in the form of scattering plots. The plots are from our baseline model and SMT-DTA model. The yellow dashed diagonal lines refer to the exact match between the prediction and the ground-truth. We can

[4] `https://www.bindingdb.org/`



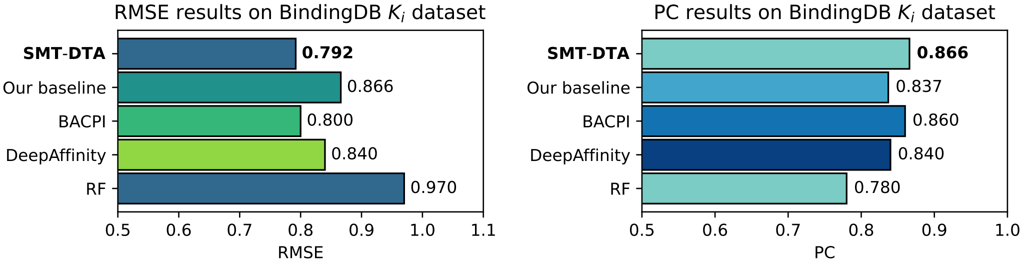
(a) Our Baseline Model.

(b) SMT-DTA Model.

Figure 3: Distribution comparison of predicted and ground-truth affinity scores. (a) is from our baseline model and (b) is from our SMT-DTA model. The yellow dashed diagonal lines refer to the exact match between the prediction and the ground-truth, the rd line is the actual match between prediction and ground-truth.

see that the red line from regression analysis of SMT-DTA model (Fig. 3b) is closer to the diagonal line, compared to the baseline model (Fig. 3a), indicating that the SMT-DTA predicted affinity scores are closer to the experimental scores.The distribution of the scattered points around the fitted lines are also narrower, justifying the smaller RMSE values of the SMT-DTA model.

The SMT-DTA model was also trained and tested on the inhibitory constant values K_i on BindingDB dataset, and the performance is summarized in Fig. 4. The performance of K_i was not widely analyzed or reported in previous works, so the comparison is limited to the DeepAffinity [28], BACPI [36], and our approaches. The test set results show that the RMSE is improved from 0.840 to 0.792, demonstrating the application of SMT-DTA model in K_i predictions.



(a) RMSE results.

(b) PC results.

Figure 4: Performance evaluation of different approaches on the BindingDB K_i dataset. (a) shows the RMSE results and (b) shows the PC results. The compared methods are DeepAffinity [28] and Random Forest [28]. ‘Our baseline’ refers to our implemented baseline with cross-attention module.

4.3 Performances on DAVIS and KIBA Benchmarks

The performance comparison on DAVIS and KIBA datasets are reported in Table 1. Besides the works from KronRLS [55], GraphDTA [51], DeepDTA [54], we also show the performance of DeepPurpose implemented with different model architectures for encoding molecule and protein [26], such as MPNN+CNN, CNN+Transformer and so on. The evaluation metrics are MSE and Concordance Index (CI) scores. The data sizes of these two datasets are smaller than BindingDB dataset, but from

Table 1: Performance evaluation of different prediction approaches on the DAVIS and KIBA datasets. The ↓ and ↑ indicate the direction of better performances. The compared methods are DeepDTA [54], KronRLS [55], GraphDTA [51], and DeepPurpose [26]. 'Our baseline' refers to our implemented baseline model with cross-attention module.

Dataset		**DAVIS**		**KIBA**	
Method	**Model**	**MSE↓**	**CI↑**	**MSE↓**	**CI↑**
KronRLS [55]	-	0.329	0.847	0.852	0.688
GraphDTA [51]	-	0.263	0.864	0.183	0.862
DeepDTA [54]	-	0.262	0.870	0.196	0.864
DeepPurpose [26]	CNN+CNN	0.254	0.879	0.196	0.856
	MPNN+CNN	0.271	0.858	0.222	0.825
	MPNN+AAC	0.242	0.881	0.178	0.872
	CNN+Trans	0.282	0.852	0.240	0.818
	Morgan+CNN	0.271	0.858	0.229	0.825
	Morgan+AAC	0.258	0.861	0.233	0.823
	Daylight+AAC	0.277	0.861	0.252	0.808
Our baseline	-	0.237	0.875	0.162	0.893
SMT-DTA	-	**0.219**	**0.890**	**0.154**	**0.894**

the tables, we can also find that our method outperforms previous works with clear improvements. Specifically, on DAVIS dataset, the MSE and CI scores of our implemented baseline are 0.237 and 0.875, which are already better than most existing works, such as DeepDTA [54]. Our proposed SMT-DTA model further improves the performances to be 0.219 MSE and 0.890 CI scores. On KIBA dataset, similarly, our baseline model surpasses the previous best work DeepPurpose [26] (MPNN+AAC). With the SMT-DTA framework, it further obtains 0.154 MSE and 0.894 CI scores.

5 Study

In this section, we present case visualizations of binding activities between drugs and targets, the application of our model for target detection, visualization of drug feature learning, and also other studies to show the impact of our SMT-DTA framework. More studies can be found in appendix.

5.1 Case Study of Drug-Target Binding

Deep learning models are often lack of interpretability. Based on Transformer architecture and cross-attention mechanism, we gain a better understanding of the DTA prediction through the interactive attention. In this subsection, we provide three case studies to visualize the atomic level attention on compound molecules and the amino acid level attention on proteins. For atomic level attention, we carry out the experiment to see the attention values of target→drug attention calculation, where the query is target protein and the key/value is compound. Therefore, the attention weights on each atom reflect the importance of the compound atoms for one specific target. Similarly, for amino acid level attention, the experiment is to see attention values of drug→target attention calculation, where the query is compound and the key/value is target protein. The attention weights on each amino acid of a protein can reflect the importance of these amino acids to the corresponding drug compound.

Following [7], we choose drug prochlorperazine (PCP) and its target S100A4 (UniProt ID), whose atomic structure is experimentally determined (PDB ID: 3M0W)[5], for one example analysis. Prochlorperazine is a phenothiazine antipsychotic medicine used to treat anxiety or schizophrenia, and its structure-activity relationship (SAR) has been thoroughly explored. From Fig. 5a, we can see that the attention highlighted atoms of PCP are consistent with the SAR features, demonstrating that our model is capable of capturing key atomic groups interacting with proteins. The ground-truth of structural binding site information can be clearly visualized from the protein-drug complex structure (PDB ID: 3M0W) in the same figure. Among 10 residues (out of 100) with the highest attention scores, 4 are

[5] `https://www.rcsb.org/structure/3M0W`

located in the vicinity of the binding site. Interestingly, three residues (Leu42, Leu79 and Met85) are from one chain, and one residue (Cys3) is from the other chain. The second case is GABA$_A$ receptor protein, a ligand gated ion channel from Erwinia chrysanthemi (ELIC) (UniProt ID: P0C7B7). We analyze the interaction between GABA$_A$ and the drug molecule flurazepam, whose key atoms are highlighted in Fig. 5b, corresponding to four atomic groups. Similar to the first case, the prediction results are assessed using the complex structure determined with crystallography method (PDB ID: 2YOE)[6]. As shown in Fig. 5b, four of the amino acids around the binding sites are identified based on the attention score using the flurazepam as query input. These identified amino acids are labelled to emphasize their close contacts with the drug molecule. In the last case study, we investigate the application of our model in predicting SARS-CoV-2 main protease (M^{pro}) and an inhibitor molecule YTJ (2-3-[3-chloro-5-(cyclopropylmethoxy)phenyl]-2-oxo[2H-[1,3'-bipyridine]]-5-ylbenzonitrile)[78]. The key atoms of YTJ and the important residues near the binding site (e.g., Cys44, Cln189, Cys145 and Pro168) are highlighted in Fig. 5c.

These three case studies clearly demonstrate the power of the present model in identifying important atomic groups or amino acids. We would like to stress that the structure information was only used when assessing the predicted key amino acids in these case studies instead of our training method. Although other amino acids not within immediate vicinity of the binding site may also have higher attention scores, we found that the model prediction results are significantly meaningful. For example, in the case of S100A4-PCP interactions, the model predicted 10 amino acids as key residues out of the whole sequence (100 amino acids), and 4 of these predicted residues are near the binding site. Similar results were observed for other cases, suggesting that the SMT-DTA model learns important rules in drug-target interactions via the proposed training framework.

5.2 Target Detection on DrugBank Dataset

To evaluate the application values and generalization capability of our model, we perform an experiment using the DrugBank [77] dataset, where most drugs and targets are excluded from the BindingDB training data. DrugBank contains the real-world drug-target pairs with different interaction types, and we take the drug-target pairs with 'inhibitor' like[9] interactions, whose activity is quantified as IC$_{50}$ scores. The DrugBank dataset includes 4351 targets, which interact with at least one drug. We carry out the test by limiting the drug dataset to those interact with less than 20 target proteins (most of the drugs only interact with one target, making it more difficult to identify true targets). We randomly selected 100 drugs among those satisfy the selection criteria, and predict the affinity scores between 100 drugs and 4351 targets. For each drug, the targets are ranked based on the predicted affinity scores, and we analyze the ranking position of the true targets out of 4351 candidates. The performance is evaluated by counting the number of correctly identified targets for these 100 drugs in the top ranked candidates (labelled as Top-N). For these 100 drugs in this test, $\mathbf{13}$ drugs have correctly identified their best candidate target (i.e., Top-1), a significant improvement comparing to the baseline model that only predicts $\mathbf{5}$ pairs in the Top-1 category. If the candidate pool is relaxed to Top-5, our SMT-DTA model correctly identifies the targets for $\mathbf{18}$ drugs, while the baseline model finds $\mathbf{10}$. This test result shows that our SMT-DTA model can be potentially useful in drug repurposing research, by predicting the targets for a given drug.

5.3 Drug Feature Learning and Drug Grouping

In reality, multiple drugs can interact with the same target and very often these drugs bind to similar regions on the target. Due to this intrinsic correspondence, drugs for the same target possess some common properties that may not be directly visualizable with conventional statistical parameters. Deep learning models can capture hidden features, which are more efficient in describing drug properties. We evaluate the drug grouping performance on 285 drugs that interact with 5 targets, which are all from the DrugBank dataset[10]. The 285 drugs are grouped by t-SNE algorithm based on the learned features of three models: our baseline model, SMT-DTA trained model, and a strong

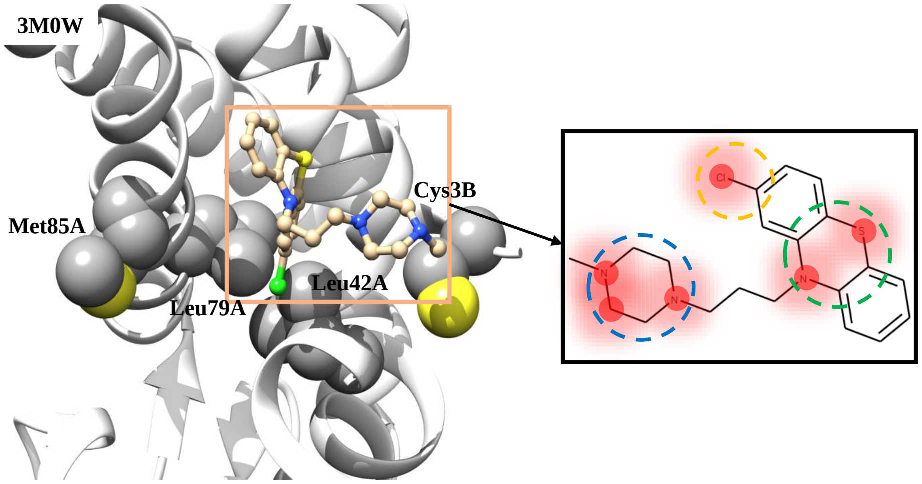
[6] `https://www.rcsb.org/structure/2YOE`

[7] `https://www.rcsb.org/structure/7M8X`

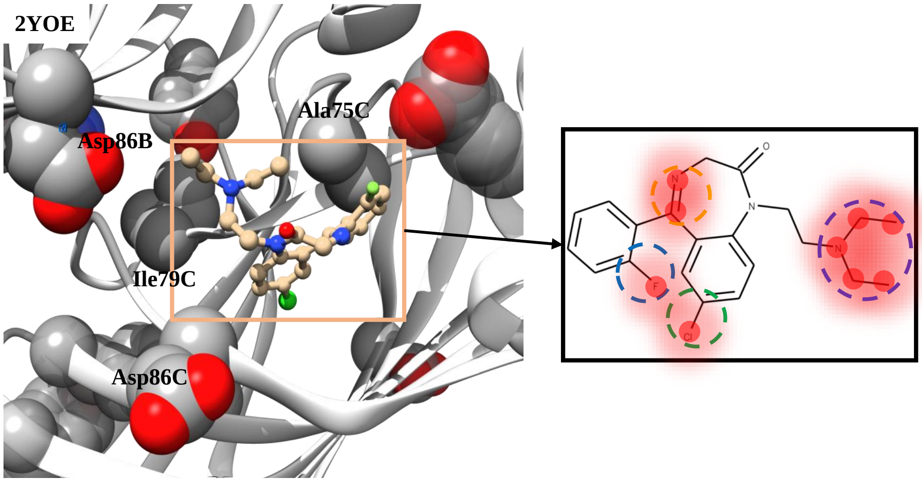
[8] YTJ is its PDBe ligand code.

[9] Including 'aggregation inhibitor', 'weak inhibitor', 'inhibitory allosteric modulator', 'inhibitor' and 'translocation inhibitor'.

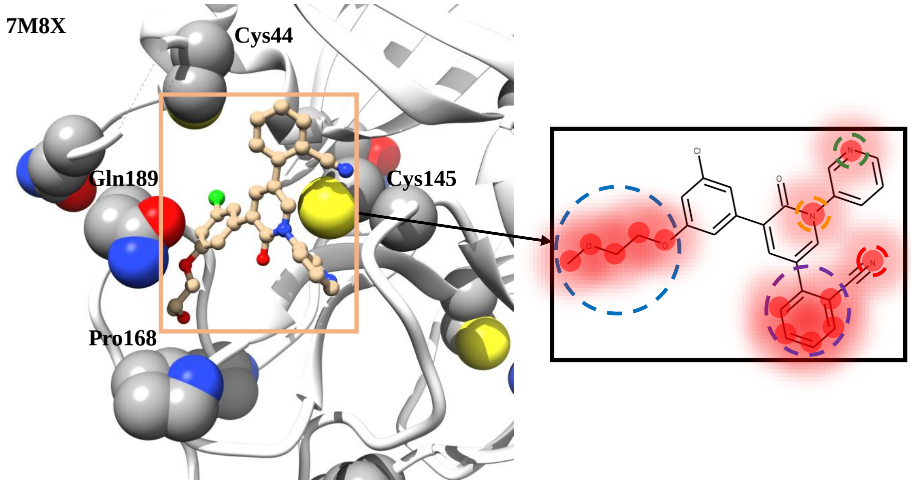
[10] The detailed information of these drugs and targets is in Supplementary Materials.



(a) Case: S100A4 target and prochlorperazine (PCP). Left panel shows the 3D view of S100A4 with PCP. Right panel shows the 2D representation of PCP, whose atoms with high attention scores are highlighted in red color. The pharmacophore groups according to SAR are enclosed by dashed circles: (1) The blue circle is a nitrogen-containing basic group. The side chain substituted with piperazine has the strongest effect; (2) The green circle contains sulfur at 5-position and nitrogen at 10-position, which are associated with antipsychotic activity; and (3) The yellow circle is electron-withdrawing group at 2-position enhancing drug activity.



(b) Case: GABA$_A$ receptor and flurazepam. The 3D structure of the complex (PDB ID: 2YOE) and 2D representation of the flurazepam are shown on the left and right panels. Atoms of flurazepam with high attention scores are highlighted in red color, and the important groups according to SAR are circled: (1) The blue and green circle are electron-withdrawing groups that enhance activity; (2)The yellow circle is a saturated double bond at 4,5-position, which increases sedative and antidepressant effects; and (3) The group enclosed by the purple circle prolongs the efficacy.



(c) Case: SARS-CoV-2 main protease (M^{pro}) and drug compound YTJ. The complex structure is shown on the left panel (PDB ID: 7M8X) and the 2D representation of the drug is shown on the right. Atoms with high attention scores match five pharmacophore groups enclosed by dashed circles.

Figure 5: Case studies of protein targets and the corresponding ligands. The amino acids with high attention scores are shown in van der Waals representations and the drug compound molecules are in ball-stick format in 3D models. The amino acids near the binding sites are labelled with their names and residue indexes.

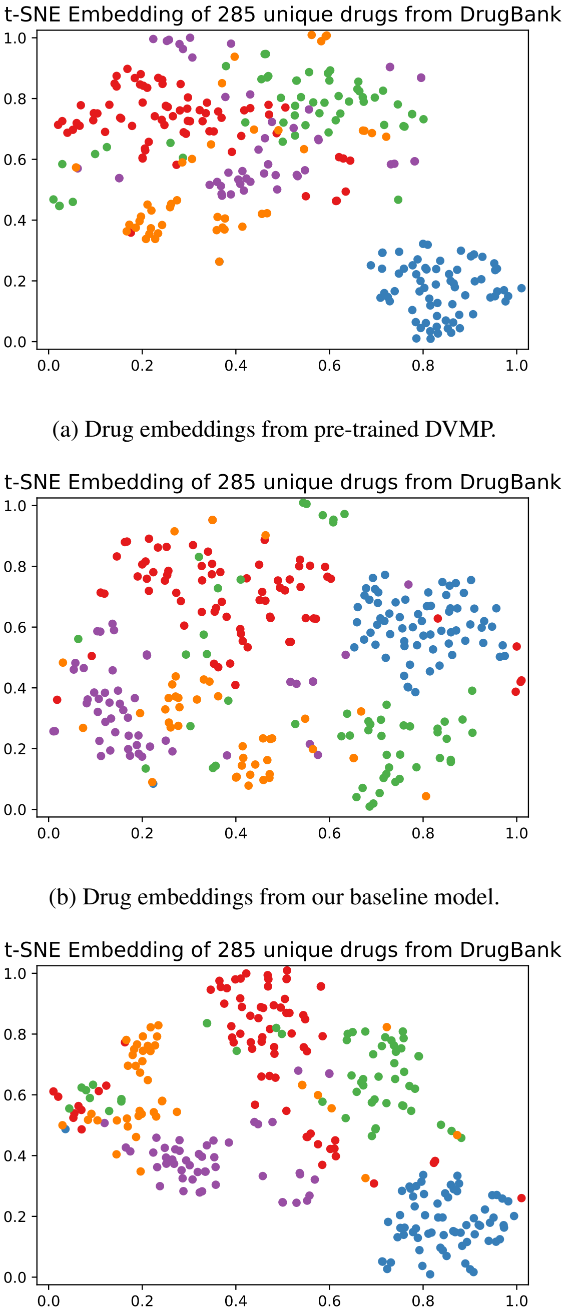
pre-trained molecule model (DVMP [80]). The embedding of these drugs in the manifold is shown in Fig. 6, where the colors are coded based on the corresponding targets. From the figure, we can see that both the baseline DTA and SMT-DTA models show distinguishable clusters, which can be mapped to the corresponding targets. In the case of pre-trained DVMP model, only the two apparent clusters can be identified (the blue dots and the rest), suggesting that the target-specific features are not learned by the model. The SMT-DTA model further improves drug embedding, manifested in better defined clusters than those of the baseline model. For example, the drugs in the green/orange groups are broadly spread in the baseline model embedding, but they are centralized into clusters in the SMT-DTA model embedding (Fig. 6c). This test result demonstrates that the drugs can be better grouped according to their targets. Such target-specific features learned by the SMT-DTA model can be a foundation for the improved performance in DTA prediction.

Table 2: Ablation study of our proposed components. We study the effect of the MLM multi-task training, the cross-attention module and the unlabeled data joint training. 'PC' stands for Pearson Correlation.

Module			**Validation MSE/Test RMSE↓**	**Valid/Test PC↑**
Cross-attention	Paired MLM	Unlabeled Data		
✓	×	×	0.625/0.787	0.846/0.848
✓	×	✓	0.601/0.772	0.853/0.855
✓	✓	×	0.545/0.735	0.866/0.868
✓	✓	✓	**0.513/0.712**	**0.875/0.878**

5.4 Ablation Study of Components in SMT-DTA

We then provide the ablation study to verify the effectiveness of each of our proposed components: the MLM multi-task training, the cross-attention module, and the joint training with unlabeled data. The results are presented in Table 2, and we report both the validation and test MSE/RMSE and Pearson Correlation scores on BindingDB IC_{50} dataset. Note that the 'Paired MLM' in the table refers to the MLM training only on paired drug-target data, and our baseline model is with the cross-attention module. From the table, we observed the following facts: (1) The MLM multi-task training on the



(a) Drug embeddings from pre-trained DVMP.

(b) Drug embeddings from our baseline model.

(c) Drug embeddings from our SMT-DTA model.

Figure 6: Visualized embeddings of 285 drugs that correspond to 5 targets from DrugBank dataset. Each color represents the drugs for one specific target. (a) shows the drug embeddings from DVMP pre-trained model, (b) is from our baseline model, and (c) is from our SMT-DTA model.

original paired data plays the most important role for improving DTA prediction. For instance, the validation MSE is decreased from 0.625 to 0.545 with a large margin, and the corresponding test RMSE 0.735 is already the best performance among previous works, e.g., BACPI [36] (0.740). (2) The joint training on unlabeled data (without paired MLM training) also improves the performance, e.g., validation MSE reduction from 0.625 to 0.601 and test RMSE from 0.787 to 0.772. (3) After combing the joint training on unlabeled data with the paired MLM training, the validation/test results are further improved to 0.513 and 0.712. The Pearson Correlation score is also increased with these components. Therefore, it is obvious that each of our component has enhancement effects for improving the DTA prediction performance.

Table 3: Different training strategies to leverage the unlabeled molecule and protein data. In this table, 'our baseline' refers to the setting of the third row in Table 2, which contains paired MLM training and cross-attention module. 'PC' stands for Pearson Correlation.

Training Strategies	**Validation MSE**↓	**Valid PC** ↑	**Test RMSE**↓	**Test PC** ↑
Our baseline	0.545	0.866	0.735	0.868
Feature based tuning	0.638	0.840	0.795	0.843
Pre-training/fine-tuning	0.536	0.868	0.738	0.867
SMT-DTA	**0.513**	**0.875**	**0.712**	**0.878**

5.5 Training Strategies Comparison

As we introduced before, we implemented a different strategy to leverage the large-scale unlabeled data. This is different from the common pre-training (on large-scale unlabeled data) and fine-tuning (on supervised labeled data) strategies (discussions of these training strategies are in Section 3.6). To compare the performance of these different strategies, we conduct DTA prediction experiments on BindingDB IC_{50} dataset. The following three different training strategies are compared: (1) *Feature based tuning*. Pre-training with large-scale unlabeled data, and then the pre-trained model is fixed and used as feature extractor to conduct subsequent tuning with the same cross-attention module; (2) *Pre-training and fine-tuning*. Pre-training with large-scale unlabeled data, and then the pre-trained model with newly added cross-attention module are all trained to fine-tune DTA prediction; and (3) *Semi-supervised multi-task training*. This is our proposed SMT-DTA strategy that the DTA prediction and MLM training on unlabeled data are jointly optimized in a multi-task framework. We report the results in Table 3 with performances on both validation and test datasets. The feature based training can not achieve a good prediction mainly due to the limited tuning parameters (only newly added cross-attention module is trained during the DTA paired data training). We also find that our semi-supervised multi-task training is the best strategy among these three, which surpasses the general pre-training and fine-tuning method. One reason is that separately pre-training on molecules or proteins ignores the importance of interaction for DTA prediction. The improved performance of the SMT framework shed light that on paired interaction related tasks, our training method can be a potential good candidate than pre-training and fine-tuning strategy, especially for the drug-target related applications.

6 Conclusions

Drug-Target Affinity (DTA) prediction is crucial in drug discovery. However, due to the limitation on the supervised data, DTA prediction using deep learning approaches has been difficult. To tackle this issue, we propose three strategies that improve the DTA prediction performance. We develop a SMT-DTA model based on paired MLM training with a multi-task framework leveraging large-scale unlabeled data, together with an efficient cross-attention module for drug-target interaction. Experiments on multiple DTA benchmark datasets show improved performance of the our method. The capability of identifying the key atom groups and amino acids is demonstrated by three case studies. The target-specific features embedded in the SMT-DTA are explored and demonstrated in drug grouping in accordance with their targets, providing an explanation to the improved performance of the proposed method.

References

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Massoud Amanlou, Jahan B Ghasemi, and Ali Masoudi-Nejad. Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. *Bioinformatics*, 36(17):4633–4642, 2020.

[2] Antti Airola and Tapio Pahikkala. Fast kronecker product kernel methods via generalized vec trick. *IEEE transactions on neural networks and learning systems*, 29(8):3374–3387, 2017.

[3] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pp. 1–6. Ieee, 2017.

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[5] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: A universal deep-learning model of protein sequence and function. *bioRxiv*, 2021.

[6] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2959–2968, 2019.

[7] Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

[8] Ruolan Chen, Xiangrong Liu, Shuting Jin, Jiawei Lin, and Juan Liu. Machine learning for drug-target interaction prediction. *Molecules*, 23(9):2208, 2018.

[9] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7870–7881, 2020.

[10] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[11] Anna Cichonska, Balaguru Ravikumar, Elina Parri, Sanna Timonen, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu, and Tero Aittokallio. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS computational biology*, 13(8):e1005678, 2017.

[12] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[13] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[17] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, and Burkhard Rost. End-to-end multitask learning, from protein language to protein features without alignments. *bioRxiv*, pp. 864405, 2020.

[18] Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

[19] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In *IJCAI*, volume 2018, pp. 3371–3377, 2018.

[20] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[22] Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

[23] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9(1):1–14, 2017.

[24] Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

[25] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[26] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020.

[27] Aleksandr Ianevski, Anil K Giri, Prson Gautam, Alexander Kononov, Swapnil Potdar, Jani Saarela, Krister Wennerberg, and Tero Aittokallio. Prediction of drug combination effects with a minimal set of experiments. *Nature machine intelligence*, 1(12):568–577, 2019.

[28] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

[29] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[32] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[33] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

[34] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature reviews Drug discovery*, 3(8):711–716, 2004.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[36] Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li. Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics*, 38 (7):1995–2002, 2022.

[37] Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv preprint arXiv:2012.11175*, 2020.

[38] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10(4):308–322, 2020.

[39] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[40] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

[41] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[42] Xuewei Liu, Danfeng Shi, Shuangyan Zhou, Hongli Liu, Huanxiang Liu, and Xiaojun Yao. Molecular dynamics simulations and novel drug discovery. *Expert opinion on drug discovery*, 13(1):23–37, 2018.

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[44] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[45] Heng Luo, William Mattes, Donna L Mendrick, and Huixiao Hong. Molecular docking for identification of potential targets for drug repurposing. *Current topics in medicinal chemistry*, 16(30):3636–3645, 2016.

[46] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

[47] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

[48] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.

[49] Seonwoo Min, Seunghyun Park, Siwon Kim, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Pre-training of deep bidirectional protein sequence representations with structural information. *IEEE Access*, 9:123912–123926, 2021.

[50] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

[51] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[52] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[53] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[54] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[55] Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

[56] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

[57] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

[58] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, 2020.

[59] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[60] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[61] Tina Roblek, Tomaz Vaupotic, Ales Mrhar, and Mitja Lainscak. Drug-drug interaction software in clinical practice: a systematic review. *European journal of clinical pharmacology*, 71(2): 131–142, 2015.

[62] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

[63] Freddie R Salsbury Jr. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current opinion in pharmacology*, 10(6):738–744, 2010.

[64] Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, pp. 230–248. PMLR, 2019.

[65] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

[66] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3): 735–743, 2014.

[67] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[68] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2): 309–318, 2019.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[70] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017.

[71] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

[72] Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9161–9168, 2020.

[73] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pp. 504–521. Springer, 2020.

[74] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

[75] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molclr: molecular contrastive learning of representations via graph neural networks. *arXiv preprint arXiv:2102.10056*, 2021.

[76] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28 (1):31–36, 1988.

[77] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.

[78] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[79] Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2):134–140, 2020.

[80] Jinhua Zhu, Yingce Xia, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Dual-view molecule pre-training. *arXiv preprint arXiv:2106.10234*, 2021.

A Experimental Settings of Binding Affinity Prediction

The detailed experimental settings of our binding affinity prediction are introduced, including the datasets, model training, compared baseline models.

A.1 Datasets

Our method leverages both the limited DTA data and the large-scale unlabeled molecule and protein data. The introductions are as follows.

Drug-Target Affinity Prediction Datasets We use the widely adopted three benchmark datasets for DTA prediction, which are BindingDB, DAVIS and KIBA datasets.

BindingDB [41] is a public database[11] of measured binding affinities, focusing on the interactions of targets with small drug-like molecules. Previous works like DeepAffinity [28], MONN [38] and BACPI [36] have been evaluated on half-maximal inhibitory concentration (IC_{50}) values derived from the BindingDB database. For consistency purposes, we use the same BindingDB dataset in test. The dataset contains 376,751 IC_{50}-labeled samples, with 255,328 unique drugs and 2,782 unique targets. We randomly sample 70% as training (including 10% held out for validation) and 30% for test. In addition, there are three other measurements: concentration for 50% of maximal effect (EC_{50}), inhibition constant (K_i) and dissociation constant (K_d), where the data size is smaller than IC_{50}. In our experiments, we also study the K_i measurement. The same data splitting and labeling protocols are applied. In order to reduce the label range, the concentrations are transformed to logarithm scales as the following: $-\log_{10}(\frac{x}{10^9})$, where x is IC_{50} or K_i in the unit of nM.

DAVIS [13] dataset contains selectivity assays of the kinase protein family and the relevant inhibitors with their respective dissociation constant (K_d) values. It has 30,056 DTA pairs with 68 unique drugs and 379 unique targets. Following DeepPurpose [26], we randomly sample 70% as training set, 10% as validation set and 20% as test set. Same as [23, 54], we also transform the raw values to logarithm scales in the same way as BindingDB data.

KIBA [66] dataset includes kinase inhibitor bioactivities measured in three metrics, K_i, K_d, and IC_{50}. KIBA scores were constructed to optimize the consistency between K_i, K_d, and IC_{50} by using the statistical information they embedded in these quantities [54].

The final KIBA dataset we used has 118,254 DTA pairs with 2,068 unique drugs and 229 unique targets. We also randomly split 70%/10%/20% as train/valid/test dataset as in DeepPurpose [26].

Unlabeled Molecule and Protein Datasets The unlabeled molecule and protein datasets are from PubChem and Pfam respectively. Details are as follows.

Pfam Dataset [50] is a database[12] of protein families that includes their annotations and multiple sequence alignments generated using hidden markov models. We use amino acid sequence of protein extracted from Pfam database as our unlabeled protein data. The training set we randomly sampled consists of $10M$ protein sequences[13].

PubChem Dataset [29] is the largest collection of freely accessible chemical information. We use Isomeric SMILES of molecule extracted from PubChem database[14] as our unlabeled molecule data. The training set we randomly sampled also consists of $10M$ molecules, which is consistent with unlabeled protein sequences.

A.2 Model Configurations

We use two Transformer encoders for molecule encoder $\mathcal{M}_{\mathcal{D}}$ and protein encoder $\mathcal{M}_{\mathcal{T}}$, and each follows $\text{RoBERTa}_{\texttt{base}}$ architecture that consists of 12 layers. The embedding/hidden size and the dimension of feed-forward layer are 768 and 3,072 respectively. The max lengths for molecule

[11] https://www.bindingdb.org/

[12] http://pfam.xfam.org/

[13] We also study other sizes in Appendix.

[14] https://pubchem.ncbi.nlm.nih.gov/

and protein are 512 and 1,024 respectively. The regression prediction head is 2-MLP layers with `tanh` activation function and the hidden dimension is 768. Our model implementation is based on Fairseq [53] toolkit with version 0.10.2[15].

A.3 Training and Evaluation

Training. Our model is optimized by Adam [30] algorithm with learning rate $1e^{-4}$. The weight decay is 0.01. The dropout and attention dropout of two encoders are all 0.1. The learning rate is warmed up in the first 10k update steps and then linearly decayed. The batch size is 32 sentences and we accumulated the gradients 8 times during training. The maximal training step is 200k. We set the optimal coefficient α and β to be 2.0.

Evaluation. We use (i) mean square error(MSE), (ii) root mean square error(RMSE), (iii) pearson correlation coefficient (PC) [1], (iv) corcondance index (CI) [20] to evaluate the performance of our model on DTA regression task. To have a fair comparison with previous works, the results on BindingDB dataset are evaluated on (ii) and (iii), and the results on DAVIS and KIBA datasets are evaluated on (i) and (iv).

A.4 Compared Baselines

We compare our SMT-DTA with following baselines. We focus on state-of-the-art deep learning models as they have demonstrated superior performance over other methods.

- **DeepDTA** [54] uses CNN [3] on both SMILES and protein sequence to extract their features. Originally DeepDTA was evaluated on DAVIS and KIBA datasets, and MONN [38] evaluated DeepDTA on BindingDB dataset.
- **DeepAffinity** [28] uses RNN [47] on both SMILES and protein sequence for unsupervised pre-training to learn their representations. After that, CNN layers are appended after RNN for both molecules and proteins to make prediction. DeepAffinity was evaluated on BindingDB dataset only.
- **MONN** [38] uses a GCN module and a CNN module to extract the features for molecule and protein, respectively. Then, a pairwise interaction module is introduced to link the molecule and protein. MONN was evaluated on BindingDB dataset.
- **BACPI** [36] uses GAT [70] to encode molecule graph and CNN to encode protein sequence, respectively. A bi-directional attention is then introduced and the final integrated features are used to make affinity prediction. BACPI was evaluated on BindingDB dataset.
- **KronRLS** [55] employs the Kronecker Regularized Least Squares (KronRLS) algorithm that utilizes 2D compound similarity-based representation of drugs and Smith-Waterman similarity-based representation of targets [55]. KronRLS was evaluated on DAVIS and KIBA datasets.
- **GraphDTA** [51] uses GCN [31], GAT [70], GIN [78] and GAT-GCN to encode molecular graph, and CNN to encode protein sequence. Finally, the concatenated features are passed to feed-forward layers for prediction. GraphDTA was evalauted on DAVIS and KIBA datasets.
- **DeepPurpose** [26] supports training of customized DTA prediction models by implementing different molecule/protein encoders and various neural architectures. They were evaluated on DAVIS and KIBA datasets.

B Deeper Understanding of SMT-DTA

B.1 Training Analysis

We provide the training process analysis from the MSE value on both training and validation sets. We plot the corresponding values at each iteration in Fig 7 for both our baseline and SMT-DTA models. From the curve, we can observe that along the training process, the MSE values on both training and validation sets from our SMT-DTA method are lower than the baseline model. While the converged

[15] `https://github.com/pytorch/fairseq/tree/v0.10.2`

training MSE values are similar on training set, the validation MSE of SMT-DTA model is much better than the baseline model. Besides, the SMT-DTA model converges faster than the baseline model. These curves demonstrate that model training can benefit from the proposed SMT-DTA framework to get faster convergence and higher accuracy, and to achieve better performance in validation and test datasets.

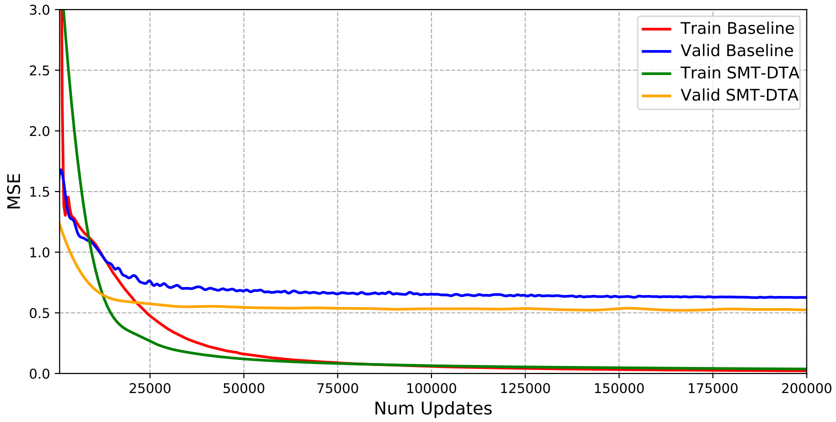
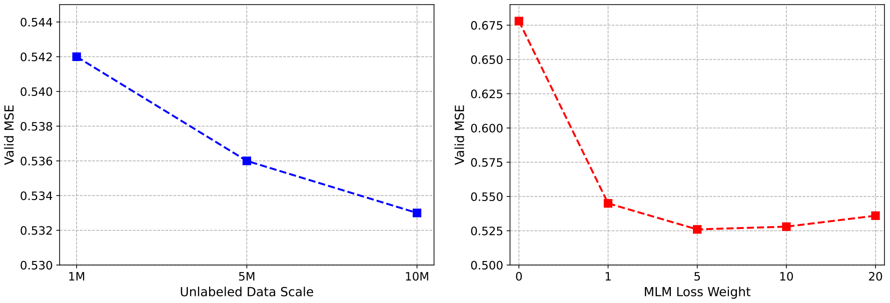


Figure 7: The MSE loss curves on BindingDB IC_{50} training and validation datasets along the training process. Results are from our implemented baseline model and our SMT-DTA method.



(a) Varied scales of unlabeled protein and molecule.

(b) Varied coefficient weight α and β, here $\alpha = \beta$.

Figure 8: Effects of (a) varied scales of unlabeled molecule and protein data, and (b) varied MLM loss coefficient α and β.

B.2 Effects of Unlabeled Data and MLM Loss Weight

We first study the effects of the different data scales of the unlabeled molecules and proteins. Specifically, we vary the unlabeled dataset to be $1M$, $5M$ and $10M$ scales during training and evaluate the effects of these unlabeled data scales on the performance of the trained model. The validation MSE scores are reported in Fig. 8a. From the figure, we see that larger unlabeled data scale will gradually improve the performance. Due to the computation resource limitation, we do not perform experiments on larger datasets. We suspect the reason of increased performance comes from data diversity. The larger unlabeled data can help the model to learn better generalized representations from more diverse data so as to enhance the DTA prediction task.

Our method introduces the coefficient weights α and β for MLM training, and we simply set α to be the same as β. Therefore, it is also necessary to investigate its effect. Here the weight value is varied among $\{0, 1, 5, 10, 20\}$. Note that we do not use cross-attention module here. Fig. 8b shows the results on validation MSE under different settings. We can see that the weight has a trade-off effect and the optimal weight is 5, around which the MSE value for the validation set is the lowest.

This is expected since the goal of MLM training is different from DTA prediction and the reduction in the DTA prediction MSE requires a proper incorporation with MLM training.

B.3 Drug and Target Information for ‘Drug Feature Learning and Drug Grouping’ Experiment

We list the detailed information of the 5 targets and their corresponding 285 drugs used in the ‘Drug Feature Learning and Drug Grouping’ experiment. The information is shown in Table 4, which contains the DrugBank ID for drugs and UniProt ID for targets.

Table 4: The 5 targets and corresponding 285 drugs that we used for embedding visualization on DrugBank dataset. Targets are provided by their UniProt ID and drugs are provided by their DrugBank ID.

Target (UniProt ID)	Drug (DrugBank ID)
P20309	DB06153, DB00462, DB00424, DB01238, DB00387, DB09089, DB00209, DB00725, DB00940, DB00622, DB01338, DB14185, DB01625, DB00477, DB01226, DB00434, DB00411, DB06702, DB01231, DB00202, DB01151, DB00458, DB00599, DB08897, DB01409, DB00934, DB09076, DB00363, DB00502, DB00193, DB01403, DB06787, DB11181, DB00505, DB06709, DB01062, DB01239, DB00280, DB04843, DB00835, DB03128, DB00809, DB00246, DB12278, DB00332, DB13720, DB01337, DB00572, DB00767, DB00342, DB13581, DB09262, DB00334, DB00986, DB00185, DB01085, DB01224, DB01019, DB01036, DB01142, DB00747, DB01591, DB00517, DB11235, DB00383, DB00496, DB09300, DB00408, DB04365, DB00376, DB09167, DB01069
P04150	DB00324, DB01185, DB09095, DB00860, DB00596, DB06781, DB14583, DB00367, DB00764, DB13003, DB14631, DB02210, DB13158, DB00838, DB09091, DB00396, DB01222, DB15566, DB14538, DB08906, DB14669, DB14539, DB00959, DB00421, DB13867, DB14649, DB01380, DB00620, DB00769, DB01395, DB14596, DB00223, DB01013, DB00351, DB08867, DB14544, DB01260, DB00288, DB01130, DB01234, DB00717, DB01047, DB00741, DB14512, DB00896, DB00180, DB00846, DB00834, DB00394, DB00687, DB00547, DB01410, DB01384, DB00240, DB00591, DB11921, DB04630, DB00663, DB14542, DB00253, DB05423, DB00635, DB00588, DB14540, DB00443, DB12637, DB14543, DB14541, DB11619
P27487	DB08530, DB07072, DB07271, DB07154, DB06993, DB04491, DB07081, DB06880, DB07851, DB07779, DB08024, DB07412, DB04578, DB04577, DB07092, DB11723, DB06011, DB07465, DB08429, DB08445, DB07021, DB07666, DB08588, DB07830, DB06939, DB07181, DB06203, DB06127, DB08672, DB01076, DB07356, DB06994, DB08164, DB08044, DB03660, DB03253, DB08743, DB02004, DB08382, DB07239, DB07482, DB01884, DB06335, DB08882, DB07135, DB08051, DB08504, DB07193, DB07067, DB01261, DB04876, DB07015, DB07328
P11217	DB03067, DB04013, DB07793, DB07807, DB03835, DB02471, DB04044, DB07949, DB04566, DB03286, DB00114, DB04643, DB02007, DB03479, DB04645, DB03496, DB04544, DB03354, DB06986, DB01843, DB02447, DB03133, DB08322, DB02519, DB02379, DB08151, DB04295, DB03218, DB02320, DB04195, DB04642, DB01823, DB02964, DB03392, DB02604, DB08503, DB04083, DB02719, DB07066, DB03383, DB04644, DB02843, DB04055, DB07792, DB02720, DB03657, DB04522, DB08500, DB03250
P42262	DB01354, DB07455, DB01355, DB01346, DB01483, DB03240, DB00463, DB00599, DB02818, DB04129, DB08304, DB00418, DB04000, DB03759, DB00142, DB05047, DB04152, DB01174, DB08303, DB00306, DB01352, DB04982, DB00898, DB02966, DB01353, DB01496, DB00794, DB08305, DB00241, DB00312, DB07598, DB02999, DB03319, DB02057, DB04798, DB02347, DB04599, DB01351, DB00237, DB01664, DB13146, DB00849