2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MTS-DTA: A drug target affinity prediction framework based on multi-task optimization and co-training

Bingchen ZHAO, Lei YU, Hongzhe TANG(✉)

Beihang University, Xueyuan Road 37, Beijing, China
{bingchenzhao, yulei, tanghongzhe}@buaa.edu.cn

Abstract. Drug-target affinity (DTA) prediction remains a critical challenge in AI-driven drug discovery yet suffers from severe scarcity of experimentally validated data due to the prohibitively high costs and time-intensive nature of biochemical assays. This data limitation not only amplifies overfitting risks but also compromises model generalizability under real-world distributional shifts. While existing approaches predominantly rely on molecular docking simulations and generative models—capable of producing synthetic data—they inadequately exploit available information due to inherent prior biases. To address these challenges, we propose MTS-DTA, a semi-supervised multi-task framework integrating co-training strategies with cross-task representation alignment. The framework introduces two core innovations: (1) multi-task synchronization, which enhances feature generalizability through joint optimization of representation and prediction tasks; (2) correlation-guided pseudo-labeling, dynamically generating pseudo-labels via inter-task dependencies to leverage unlabeled data while mitigating noise propagation. Benchmark evaluations confirm the framework's improved robustness against distributional biases, establishing a viable strategy to address data scarcity in drug discovery.

Keywords: Drug–Target Affinity Prediction, Multi-task Learning, Semi-supervised Learning, Masked Language Modeling

1 Introduction

With advancements in pharmacological sciences and biotechnology, drug-related properties, including drug-drug interactions, adverse drug reactions, therapeutic synergies, and drug-target interactions (DTIs)—have emerged as critical determinants of therapeutic outcomes, garnering substantial research attention. Among these, drug-target affinity (DTA) prediction holds pivotal importance for drug development success, as it directly governs pharmacodynamic mechanisms and serves as the fundamental basis for elucidating drug action principles.

Computational DTA prediction methodologies have evolved through three distinct phases: 1) early-stage molecular docking constrained by manual feature engineering [1], 2) machine learning-driven virtual screening limited by shallow feature representations [2,3], and 3) post-2018 deep learning paradigms leveraging graph neural networks and Transformers, enabled by advancements in GPU acceleration and large-scale

biochemical databases [4,5]. Although contemporary architecture automates molecular interaction modeling, persistent challenges arising from data scarcity and distributional biases continue to motivate methodological advancements in the field.

However, the efficacy of deep learning models in drug-target affinity (DTA) predictions remains critically dependent on the availability of substantial labeled training data requirement challenged by the prohibitively high costs and protracted timelines associated with experimental data acquisition. For instance, while the widely adopted BindingDB dataset comprises approximately 2 million drug-target binding records [6], only ~10% meet high-confidence standards due to inherent data heterogeneity caused by variable experimental protocols and inconsistent validation criteria. This scale starkly contrasts with the billion-sample datasets driving breakthroughs in natural language processing and computer vision, fundamentally constraining model predictive capacity.

Under such data scarcity, the chemical space coverage of existing labeled datasets (e.g., PDBbind and Davis [7]) likely represent <1% of pharmacologically relevant drug-target combinations. This disparity forces models to prioritize target-specific feature extraction over universal interaction pattern discovery, severely degrading performance on low-resource targets. Consequently, pharmaceutical R&D exhibits structural derivative bias—preferential synthesis of compounds with established scaffolds—which amplifies data distribution skewness and exacerbates generalizability limitations in novel chemical spaces.

Such data paucity not only amplifies overfitting risks but also severely constrains model generalizability to novel targets or unexplored chemical scaffolds, establishing a critical bottleneck for both DTA prediction accuracy and de novo drug development.

Current methodologies addressing severe data scarcity predominantly rely on molecular docking simulations and generative architectures (e.g., GANs [8], diffusion models), which artificially augment training data through two inherently limited approaches: (1) reliance on empirical force fields or predefined generative priors intrinsically restricts chemical diversity exploration; (2) generated molecular candidates necessitate validation via reliable affinity prediction models, paradoxically intensifying—rather than alleviating—the demand for experimentally verified labeled data.

These observations motivate two principled strategies to address data scarcity in DTA prediction:

First, we posit that heterogeneous affinity measurements can be unified across experimental protocols. While individual label-specific datasets (e.g. $k_d, IC50$) suffer from critical scarcity, numerous experimentally measured variants exist across different validation methodologies. Joint utilization through multi-task learning enables synergistic optimization across labels, thereby exposing latent interaction patterns inaccessible to single-task frameworks.

Second, we recognize that unlabeled chemical data—constituting >99% of available chemical space—exhibits inherently lower distributional bias compared to sparse labeled counterparts. Current DTA models predominantly exploit this resource through unsupervised pretraining yet fail to harness its full discriminative potential. Inspired by virtual screening workflows, we propose controlled pseudo-label assignment via semi-supervised co-training, enabling noise-resilient extraction of latent pharmacological patterns while maintaining training stability.

To address these challenges, we propose MTS-DTA—a semi-supervised framework integrating co-training paradigms with multi-task optimization. The innovation of our approach manifests in two principal dimensions:

(1) Joint Representation Learning: Simultaneous optimization of molecular and protein characterization tasks enhances feature specificity for improved predictive accuracy.

(2) Dynamic Pseudo-label Optimization: A boosting-enhanced pseudo-label generation mechanism enforces multi-task consistency by leveraging biochemical correlations among affinity metrics (Ki, Kd, IC50).

We evaluated MTS-DTA on three benchmark datasets: BindingDB[6], DAVIS[9], and KIBA[10]. For comparative analysis, baseline models spanning machine learning[11] and deep learning[12] paradigms were selected, including KronRLS[13], SimBoost[14], DeepDTA[7], and WideDTA[15]. Detailed implementation protocols for these baseline comparators are elaborated in Section 2.

2 Related Work

Unlike the DTI prediction task based on binary classification, the prediction goal of the DTA prediction task is the drug-target affinity value of the combination to measure the strength of the binding interaction between the drug and the target, which is usually regarded as a regression problem. In recent years, many methods have emerged in the field of drug-target affinity prediction, including traditional methods based on machine learning [11] and methods based on deep learning [12].

In the past few decades, traditional machine learning methods have become common in DTA prediction tasks. In early studies, machine learning methods were mainly used for DTA prediction, and regression modeling was achieved by artificially designing the characteristics of molecules and targets. Typical methods include: KronRLS [13] is based on the nuclear regularization least squares method, which uses the similarity matrix of compounds and proteins for prediction. SimBoost [14] predicts affinity by integrating molecular fingerprints with protein sequence similarity and gradient boosting tree regression. Such methods rely on expert experience to design features (such as molecular fingerprints and amino acid composition), which have problems such as limited feature expression ability, poor generalization across targets, and difficulty in capturing complex molecular-target interactions.

At the same time, deep learning methods can extract task-specific features from raw data without human intervention, and many deep learning methods can significantly improve model performance by automatically learning the characterization of molecules and targets. According to the type of input data, it can be divided into two categories:

- Sequence information-driven: Extract features through sequence modeling using molecular smiles and protein amino acid sequences as inputs. Representative Methods: DeepDTA [7] uses two CNNs encoding compound and protein sequences, respectively, for regression prediction by the fully linked layer (FC). WideDTA [15] introduces protein domain annotation and molecular physicochemical properties as additional

features. SimCNN-DTA [16] combines sequence similarity matrix with deep feature fusion.

• Structural information-driven: Directly leverage 3D molecular maps or protein structure information to model spatial interactions. Representative Method: GraphDTA [17] represents molecules as graph structures, encoded by graph neural networks (GNNs). FusionDTA [18] enhances binding site perception by fusing molecular and protein contact maps. AttentionDTA [19] greatly optimizes the model effect by introducing a cross-modal attention mechanism to align the molecular substructure and protein binding domain. AttentionMGT-DTA [20] uses a graph transformer and attention mechanism for multimodal drug target affinity prediction. TransVAE-DTA [21] combines Transformer and variational autoencoder (VAE) to fuse drug and target coding features.

Although the structural information-driven method has better results, the acquisition of training data is also more constrained due to the difficulty of obtaining structural information, and the consumption of computing resources is relatively larger. Although the above methods have made breakthroughs in feature extraction, their performance is still limited by the scarcity of labeled data, and their performance is significantly reduced in cold-start scenarios (new targets/new compounds), and the generalization of the model is still insufficient.

In order to alleviate the problem of data scarcity, some studies have tried to optimize representation learning using unlabeled data: SSM-DTA [22] improves prediction accuracy by enhancing the characterization of drugs and targets with unlabeled data through MLM tasks.However, these unlabeled data are only used for training with pretrained models, and have limited effect on the generalization of the model.

In summary, the existing studies have not made full use of the unlabeled data and dissimilar labeled data in the vast chemical space, in view of this situation, our model proposes the idea of joint representation learning and dynamic pseudo-label optimization, which can make full use of the data with different unlabeled data and labeling systems, which has not been mentioned in the current research.

3 Methods

To efficiently optimize protein and molecule representations, and to take advantage of the huge amount of unlabeled data, we propose our semi-supervised multi-task DTA prediction architecture, MTS-DTA, which has a rough structure as shown in Figure 1.

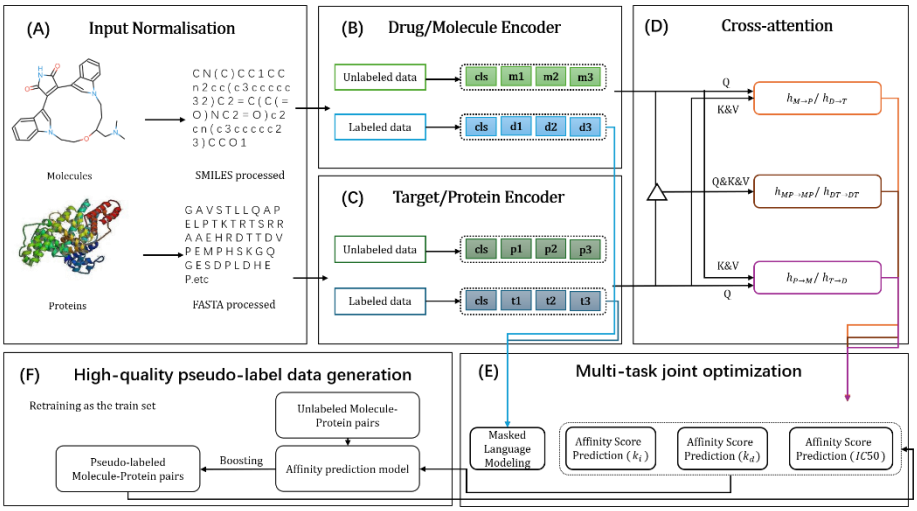

Fig. 1. The overall architecture of the MTS-DTA. (A) Enter the standardization module, where we convert the data for drugs and proteins into a format that we can use. (B) and (C) Molecular and protein encoder modules, we use two Transformer to encode molecules and proteins separately to extract features. (D) Cross-attention module, we utilize three different cross-attention mechanisms to generate a joint characterization of drug-target pairs. (E) Joint optimization module, where we will jointly optimize the characterization task and the prediction task. (F) Pseudo-label data generation module, where we generate high-confidence pseudo-label data according to certain rules and re-iterate the model.

In terms of overall structure, our model architecture can be divided into three major parts, namely the Generation of drug-target pair joint characterization, the multi-task joint optimization, and the High-quality pseudo-label data generation. This division is described in more detail below.

3.1 Generation of drug-target pair combination characterization

The data in this study consists of three parts: labeled drug-target pair data, unpaired drug molecule data, and unpaired protein sequences. After the unpaired data are generated by embedding, the potential drug-target pairs are expanded by using a combination strategy of molecule and protein, and since most of the combinations are low-activity pairings, the design can reduce the interference of pseudo-label noise by enhancing the data differentiation.

Input Normalization

To achieve efficient characterization of protein sequences and drug molecules, a differentiated data cleaning and embedding framework was designed in this study.

For the protein FASTA sequence, the description line (starting with >) and non-standard amino acid characters were removed first, and only 20 standard amino acids were retained. Then, the sequence was truncated to a maximum length of 1,024

characters, and the ultra-long sequence (accounting for <5%) was encoded in a sliding window segment to balance the computing resource consumption. Finally, a single-letter token sequence (such as "M A G V") is generated through character-level segmentation, which preserves the natural order of amino acids and adds vacancies to assist in parsing.

For drug molecule SMILES, RDKit-based chemical compliance verification screened legal molecules, SMILES was broken down into atomic, bond, and ring markers (e.g., "Cl C = O") by regularization tokenization strategy, and the maximum length was limited to 512 tokens, and short sequences were aligned by filler characters.

Molecule Encoder and Protein Encoder

In the embedding generation stage, two independent RoBERTa-base [23] encoders were used in this study. The input layer maps tokens to 768-dimensional space and constructs learnable position coding matrices for adapted proteins (up to 1,024) and molecules (up to 512), respectively. Through dynamic masking language modeling training (masking 15% of the input token), the model learns biological and chemical semantic features. During feature extraction, a 768-dimensional global sequence embedding vector is generated by time-series average pooling based on the token-level latent state output by the last layer of Transformer. The framework optimizes computational efficiency through differentiated sequence truncation strategies (sliding window vs. fixed length), while enhancing data robustness with a large amount of unlabeled data..

For the sake of later description, we define D as a sequence of a tagged drug molecule, then $|D|$ is the length of its sequence, in the same way, M represents an untagged drug molecule, T denotes a tagged protein target sequence, and P denotes an untagged protein sequence.Then we named the two encoders used for molecular code generation and protein code generation, respectively E_{mol} and E_{prot}。 We mark the beginning of the division sequence with [cls], so we get H_D, H_M, H_T, H_P four groups of hidden states which are shown in equations 1,2,3,4.

$$H_D = \left\{h_{[cls]_D}, \{h_{d_i}\}_{i=1}^{|D|}\right\} \tag{1}$$

$$H_M = \left\{h_{[cls]_M}, \{h_{m_i}\}_{i=1}^{|M|}\right\} \tag{2}$$

$$H_T = \left\{h_{[cls]_T}, \{h_{p_i}\}_{i=1}^{|T|}\right\} \tag{3}$$

$$H_P = \left\{h_{[cls]_P}, \{h_{p_i}\}_{i=1}^{|P|}\right\} \tag{4}$$

Where $h_{[cls]}$ represents the beginning mask of each embedding and h_i represents the bit of the embedding.

Cross-attention mechanisms

The core of the DTA prediction task is to predict binding affinity by modeling the interaction between drug and target, and the key is to efficiently characterize the interaction process of drug molecule intercalation and target protein intercalation. The traditional method uses a pairwise interaction mechanism to generate an interaction matrix by traversing all the positions of the two sequences. While this fully ligated approach covers potential sites, it is computationally expensive when dealing with long sequences.

Since the actual drug-target effect usually occurs only in the observation of a few key groups (the rest of the sequences are mostly responsible for structural functions), we introduce a cross-attention module to replace the traditional mechanism. By focusing on key interaction areas, the attention mechanism significantly reduces computational complexity while maintaining the prediction effect.

In order to obtain a more comprehensive interactive information, we use three different cross-attention modes [24] to focus on different drug-target forms of action, which correspond to the drug-to-target mode of action, the target-to-drug mode of action, and the overall binding mode of the target-drug conjugate. First we perform a cross-attention mechanism between $H_{D.}\ or, H_{M.}$ And $H_{T}.or, H_{P.}$ respectively, where Query is $h_{[cls]_D}(h_{[cls]_M})$ or $h_{[cls]_T}(h_{[cls]_P})$, and the Key & Value are $H_T(H_P)$ or $H_D(H_M)$, Taking the combination of D and T as an example, we can get the results in equations 5 and 6.

$$h_{D\to T} = softmax(\frac{\left(h_{[cls]_D}W_1\right)(H_TW_2)^T}{\sqrt{d}})(H_TW_3) \tag{5}$$

$$h_{T\to D} = softmax(\frac{\left(h_{[cls]_T}W_4\right)(H_DW_5)^T}{\sqrt{d}})(H_DW_6) \tag{6}$$

where d is the dimension of the hidden layer, and W_s. is the parameter matrix.

In addition to these two, the third type of cross-attention is actually achieved through self-attention, and we hope that this characterization focuses on the local interaction between the protein and the drug. Therefore, we stitch together the representations of drugs and proteins to achieve a third cross-attention mechanism through sparse self-attention, and each token only focuses on k positions before and after, so as to reduce computational complexity and supplement global cross-attention.

Finally, the weight matrix of the cross-attention mechanism is transformed into a diagonal matrix to realize the transformation of the three cross-attention mechanisms into the same output dimension, and the final encoder output is obtained through the weighted summation mechanism, as shown in Equation 7.

$$H_{final} = \alpha \cdot h_{mol\to prot} + \beta \cdot h_{prot\to mol} + \gamma \cdot SelfAttn \tag{7}$$

where the weights α, β, γ are learnable parameters.

For different prediction heads, we will initialize with different gating parameters, and iterate the gating parameters with different components of the loss function to achieve the differentiation of the encoder output.

3.2 Multi-task joint optimization

The data with the largest stock of DTA label data are three kinds of labels k_i, k_d and $IC50$, and considering the problem of data volume, we choose the regression tasks of these three labels as our three main prediction tasks. At the same time, considering that the quality of representation in deep learning is often directly related to the effect of downstream tasks, we also regard the loss of representation tasks as part of our multi-task joint optimization.

After obtaining the encoder output suitable for subsequent regression tasks, we need to use three convolutional neural networks (CNNs) as the regression prediction heads, and at the same time, the three prediction tasks of k_i, k_d and $IC50$ are jointly optimized with the representation tasks of MLM to obtain representations that are more suitable for downstream tasks, so as to improve the prediction effect of the model.

The joint optimization of the four tasks is realized by the weighted sum of the loss function, and the prediction task is the mean square error (MSE) due to the characteristics of the regression task, and its calculation method is as follows in Equation 8.

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \tag{8}$$

In order to make the downstream prediction task more compatible with the upstream representation task, we also used the MLM task [25] to optimize the representation model on the basis of using the existing pre-trained model. For the sequence of the drug molecule and the sequence of the protein, we randomly replace some Tokens according to Bert's idea, by replacing them with some proportions of [MASK]Token, and then trying to restore the original sequence by the mask sequence, we can get a loss of predicted original sequence from the mask sequence, as shown in Equation 9.

$$L_{MLM}^{D} = -\frac{1}{\mathrm{M_D}}\sum_{k=1}^{M_D}\log P(d_k||D'), L_{MLM}^{T} = -\frac{1}{\mathrm{M_T}}\sum_{k=1}^{M_T}\log P(t_k||T') \tag{9}$$

where d_k and t_k are the mask tokens,M_D and M_T are the number of mask markers. This loss measures the ability to predict the replaced part from the context of the sequence, i.e., the representation of the encoder.

During the training process, the total loss function is the weighted sum of the loss function of the three prediction tasks and the loss function of the MLM task [26], as shown in Equation 10.

$$L_{total} = \sum_{i=1}^{3}\left(\frac{1}{2\sigma_i^2}L_{MSE_i} + \log\sigma_i\right) + \left(\frac{1}{2\sigma_{MLM}^2}L_{MLM} + \log\sigma_{MLM}\right) \tag{10}$$

where σ is the task-specific noise figure, which is the learnable parameter.

Assuming that the prediction error of each task obeys a Gaussian distribution, the uncertainty of the task can be measured by its noise figure (variance σ^2)[27], and the advantage of this dynamic weight adjustment method is that if the loss of a task increases, the noise σ^2 will increase significantly, thus reducing the weight of the loss function, which means that the model considers the task unreliable. With this approach, the weight distribution of the loss function can be dynamically adjusted to quickly adapt to the uncertainty of the task.

Considering that there is no artificial standard for MLM tasks, and the magnitude of loss is often greater than that of supervised regression tasks, we set its initial loss function to a larger value when the weights are initialized, so as to reduce its initial weights and prevent them from dominating the training. At the same time, considering that the joint optimization of MLM tasks is only to enhance the encoder representation ability, not to directly optimize downstream tasks, we also set a lower learning rate for MLM tasks to stabilize training.

To avoid the conflict between the optimization direction of the MLM task and the downstream task, we also set the loss truncation, that is, when the MLM loss is lower than the threshold, the gradient backhaul is prevented.

Combined with the above strategies, we can obtain three well-trained DTA prediction models at the end of this training phase for subsequent pseudo-label generation tasks.

3.3 Dynamic semi-supervised training

After obtaining three relatively well-trained prediction models for predicting k_i, k_d, $IC50$, we can remove the votes of low-confidence data according to the relationship between the three indicators according to the idea of boosting. At the same time, we will also use labeled data to control the generation progress of pseudo-labeled data, so as to avoid a large amount of untrustworthy noise data entering the training set and worsening the effect of the model.

Using three prediction models, we can make predictions on the prepared data, in order to facilitate progress control, we randomly incorporate labeled data into the unlabeled data at a ratio of 1:10, and predict batch by batch, when the deviation between the predicted result and the true value of the labeled data is greater than three percent of the set threshold, the batch of data is considered invalid.

At the same time, using the typical values of the k_i, k_d, $IC50$ data, we can also remove drug-target pairs containing outliers from the output of the three predictors, so as to avoid introducing excessive noise into subsequent training.

After obtaining enough pseudo-label data, we repeat the above process until the metrics of the prediction model reach a desired goal, and with this method, we obtain a DTA prediction model that is better than the current baseline.

The overall structure of this part is shown in Algorithm 1.

Algorithm 1. Multi-Strategy DTA Pseudo-Labeling (MTS-DTA)

Input : $\mathcal{M} = \{M_{k_i}, M_{k_d}, M_{IC_{50}}\}$ – Trained models
$\mathcal{D}_L = \{(x_i, y^{k_i}, y^{k_d}, y^{IC_{50}})\}$ – Labeled data
$\mathcal{D}_U = \{x_j\}$ – Unlabeled data
$\tau_{\text{dev}} = 0.03$ – Relative error threshold
$\tau_{\text{out}} = 3$ – Outlier cutoff

Output: $\mathcal{M}_{\text{final}}$ – Optimized ensemble model

Initialize:
$\mathcal{D}_{\text{mix}} \leftarrow \text{Merge}(\mathcal{D}_L, \mathcal{D}_U, 1:10)$
$B \leftarrow \text{Partition}(\mathcal{D}_{\text{mix}})$
while *ValidationLoss* $> \epsilon$ **do**
  foreach $B_k \in B$ **do**
    Prediction Phase:
    $\forall x \in B_k$:

$$\hat{y}^{k_i} \leftarrow M_{k_i}(x)$$
$$\hat{y}^{k_d} \leftarrow M_{k_d}(x)$$
$$\hat{y}^{IC_{50}} \leftarrow M_{IC_{50}}(x)$$

    Validation:
    if $B_k \cap \mathcal{D}_L \neq \emptyset$ **then**
      Compute $\delta = \frac{1}{3}\sum_m (y^m - \hat{y}^m)^2$
      if $\delta > \tau_{dev}$ **then**
        Discard B_k **continue**
    Filtering:
    $\forall x \in B_k$: **if** $\exists m \in \{k_i, k_d, IC_{50}\} \mid |\hat{y}^m - \mu_m| > \tau_{out}\sigma_m$ **then**
      Remove x from B_k
    Pseudo-labeling:
    $\forall x \in B_k$:

$$\tilde{y} = [\hat{y}^{k_i}, \hat{y}^{k_d}, \hat{y}^{IC_{50}}]$$

    $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \{(x, \tilde{y})\}$
  Model Update:
  $\mathcal{M} \leftarrow \text{Retrain}(\mathcal{D}_L)$
return $\mathcal{M}_{\text{final}} = \{M_{k_i}, M_{k_d}, M_{IC_{50}}\}$

4 Experiment

4.1 Datasets

The datasets we used for training and evaluation were BindingDB[6], Davis[9], and KIBA[10], and after cleaning the outlier samples and eliminating the duplicate samples, we obtained the following benchmark dataset, as shown in Table 1.

Table 1. Overview of the preprocessed dataset

Datasets	Drugs	Proteins	Interactions
BindingDB	255328	2782	376751
Davis	68	361	24548
KIBA	2052	229	117178

Considering that only Binding has complete k_i, k_d and $IC50$ indicators in the three datasets, KIBA only has its calculated KIBA score, and DAVIS as a dataset for studying kinase mechanisms has only more k_d data. To verify the effectiveness of our MTS-DTA framework, we first trained it on BindingDB and compared it with other models.

4.2 Evaluate metrics and models for comparison

For the regression task, we used concordance index (CI), MSE, pearson correlation coefficient (PC) and regression toward the mean (r_m^2) to evaluate the performance of our model.

CI is an evaluation metric which reflects the correctness of the result, as we showed in Equation 11.

$$\text{CI}=\sum_{\delta_j>\delta_i} h(b_i - b_j) \tag{11}$$

Where $h(x) = \begin{cases} 0 & x < 0 \\ 0.5 & x = 0. \\ 1 & x > 0 \end{cases}$

MSE is a commonly used index to measure error. Given N samples with corresponding prediction value yi and ground truth value $\hat{y}_i$, as we showed in Equation 12.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

r_m^2 is a metric evaluating the external predictive performance. A model was regarded acceptable if and only if $r_m^2 \geq 0.5$. r_m^2 is defined as Equation 13.

$$r_m^2 = r^2 * (1 - \sqrt{r^2 - r_0^2}) \tag{13}$$

where r denotes the squared correlation coefficients between the observed and predicted values with intercepts and r_0 is the coefficient without intercepts.

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables, as shown in Equation 14.

$$P_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{14}$$

We used the following baseline model to compare with our model: KronRLS, SimBoost, DeepDTA, WideDTA, SimCNN-DTA, GraphDTA, AttentionDTA, TransVAE-DTA. Information about them is described in the section on related work.

4.3 Experimental results

Considering that some of the models used for comparison are missing the values of some metrics, we choose different metrics depending on the data set and the model.

Table 2. Performance of our model and baseline methods on the IC50 part of the BindingDB dataset.

Dataset	Model	MSE ↓	PC ↑
BindingDB	KronRLS[13]	0.774	0.760
	SimBoost[14]	0.662	0.783
	DeepDTA[7]	0.525	0.848
	WideDTA[15]	0.519	0.858
	SimCNN-DTA[16]	0.562	0.860
	GraphDTA[17]	0.574	0.817
	AttentionDTA[19]	0.492	0.878
	TransVAE-DTA[21]	0.453	0.882
	Our Model	**0.412**	**0.896**

Table 3. Performance of our model and baseline methods on the k_i part of the BindingDB dataset.

Dataset	Model	MSE ↓	PC ↑
BindingDB	KronRLS	0.814	0.760
	SimBoost	0.782	0.783
	DeepDTA	-	-
	WideDTA	-	-
	SimCNN-DTA	-	-
	GraphDTA,	0.574	0.817
	AttentionDTA	0.492	0.878
	TransVAE-DTA	-	-
	Our Model	**0.489**	**0.889**

As can be seen from Tables 2 and 3, our model is significantly better than the other models used for comparison on the BindingDB dataset.

On the Davis and KIBA datasets, since only one or two labels are used for fitting, we modified both the multi-task training part and the semi-supervised voting part of the model to fit the dataset and compare it with other models under the same conditions, and the experimental results are shown in Tables 4 and 5.

Table 4. Performance of our model and baseline methods on the Davis dataset.

Dataset	Model	MSE ↓	CI ↑	r_m^2 ↑
Davis	KronRLS	0.392	0.869	0.512
	DeepDTA	0.386	0.846	0.533
	WideDTA	0.394	0.853	0.497
	GraphDTA,	**0.301**	0.858	-
	AttentionDTA	0.399	0.865	-
	TransVAE-DTA	0.333	0.868	0.571
	AttentionMGT-DTA	0.308	0.872	0.584
	Our Model	0.317	**0.877**	**0.592**

Table 5. Performance of our model and baseline methods on the KIBA dataset.

Dataset	Model	MSE ↓	CI ↑	r_m^2 ↑
KIBA	KronRLS	0.262	0.806	0.567
	DeepDTA	0.472	0.782	0.521
	WideDTA	0.425	0.791	0.525
	GraphDTA,	0.482	0.780	-
	AttentionDTA	0.518	0.797	-
	TransVAE-DTA	0.278	0.822	**0.632**
	AttentionMGT-DTA	0.314	0.819	0.628
	Our Model	**0.257**	**0.828**	0.626

4.4 Analysis of experimental results

From the experimental results in the previous part, we can see that on the BindingDB dataset, our model has achieved a significant advantage over other models because it can use multi-task joint optimization and multi-task voting semi-supervised learning to obtain more information, while on the other two datasets, due to its relatively few fitting labels, our multi-task architecture is difficult to play an advantage, and our model is slightly behind those models that pay attention to representation in some evaluation indicators.

Considering that the Davis dataset contains only 68 compounds and 361 protein sequences, the evaluation results on the Davis dataset can effectively reflect the generalization performance of the model in the case of small data volume. Our model is only slightly lower than the best model on MSE metrics, leading both CI and r_m^2, which shows that our architecture can alleviate the problem of insufficient generalization caused by data scarcity. In our analysis, the main reason why our model on Davis did not achieve optimal in all the selected metrics is that Davis is a small dataset with only more than 20,000 drug-target affinity data and does not constitute the multitasking mode required for our core architecture.

From the experimental results, it can be concluded that the MTS-DTA framework we constructed can effectively improve the accuracy of model prediction with sufficient multi-task data, and we will conduct a series of ablation experiments to verify the effectiveness and value of each module.

4.5 Ablation experiments

In order to further study the impact of multi-task joint optimization and pseudo-label data participation training on the model performance, we conducted a series of ablation experiments on the BindingDB dataset.

First, we used two sets of controlled experiments to verify the effects of three independent cross-attention mechanisms and MLM task loss on the model, and the experimental results are shown in Table 6.

Table 6. Performance of each version of the model on BindingDB dataset

Dataset	Model	MSE ↓	PC ↑
IC50	Model with only Self-Attention	0.522	0.839
IC50	Model without Loss of MLM	0.584	0.807
IC50	Model normal	**0.412**	**0.896**

From the results of the ablation experiment, it can be seen that our three different cross-attention mechanisms can effectively improve the effect of the model, and in the absence of this mechanism, only self-attention is retained, and the MSE and PC of the model are not as good as the original model. At the same time, the loss of MLM task has a greater impact on the model, and the effect of the model is greatly reduced in the absence of MLM loss, which is even similar to the machine learning method, indicating that the reasonable optimization of the characterization task has an important impact on the downstream task.

At the same time, since the semi-supervised training process of our model will add pseudo-label data to the training set, we also designed experiments to study the effect of the number of pseudo-labeled data on the model indicators, and the specific results are shown in Table 7.

Table 7. Performance of each proportion of data on BindingDB dataset

Dataset	The proportion of labeled data in the training set	MSE ↓	PC ↑
IC50	20%	0.449	0.877
IC50	10%	0.412	**0.896**
IC50	5%	**0.407**	0.892

The trend is shown in the figure 2.

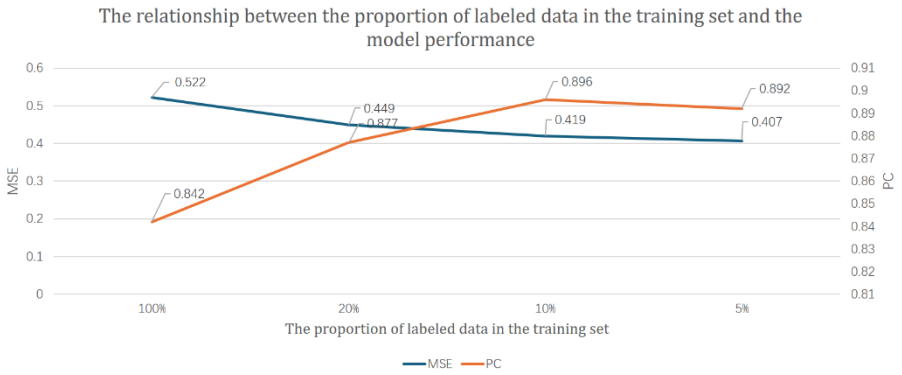

Fig. 2. The relationship between the proportion of labeled data in the training set and the model performance

With the increase of the proportion of pseudo-label data, the performance of the model also improves to a certain extent, but when too much pseudo-label data is added, some indicators of the model also decrease to a certain extent. This phenomenon shows that the data generated by our pseudo-label data production mechanism is effective, which can introduce more hidden information into the learning of the model, but at the same time, too much pseudo-label data may amplify some systematic errors and lead to the decline of the model's effectiveness. The introduction of too much pseudo-label data into the training will also have a greater demand for computing resources, considering the efficiency and cost, our final model is a version with a 10% incorporation ratio.

5 Conclusions

Compared with the existing model, our MTS-DTA can use the unlabeled data with higher efficiency, thereby improving the model performance, and showing a significant improvement in the DTA prediction task. Through multi-task joint optimization, we can unify scattered data labels and uncover hidden information. In addition, utilizing a dynamic semi-supervised learning training framework, we are able to generate high-confidence data and leverage it into training.

The results show that MTS-DTA is significantly better than the existing baseline model in low data scenarios and new target prediction, which greatly improves the accuracy of DTA prediction. For example, our model achieves a significant improvement of 4% on the IC50 dataset on BindingDB compared to the comparison baseline method. In general, the framework not only establishes the utilization paradigm of unlabeled biological data but also verifies the feasibility of multi-task collaborative optimization. The problem of overfitting and insufficient generalization performance caused by the scarcity of annotated data is greatly optimized.

However, the current dataset only has the corresponding conditions for BindingDB, and the task composition may need to be optimized in the future, to improve the effect of our model in the case of a single task.

References

1. Morris, G.M., Goodsell, D.S., Halliday, R.S., et al.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19(14), 1639–1662 (1998). https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14.
2. Cheng, F., Li, W., Zhou, Y., et al.: AdMetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J. Chem. Inf. Model. 52(11), 3099–3105 (2012). https://doi.org/10.1021/ci300367a.
3. Kellenberger, E., Rodrigo, J., Muller, P., Rognan, D.: Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins Struct. Funct. Bioinform. 57(2), 225–242 (2004). https://doi.org/10.1002/prot.20149.
4. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 40, D1100–D1107 (2011). https://doi.org/10.1093/nar/gkr777.
5. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H.: PubChem Substance and Compound databases. Nucleic Acids Research. 44, D1202–D1213 (2015). https://doi.org/10.1093/nar/gkv951.
6. Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K.: BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Research. 35, D198–D201 (2006). https://doi.org/10.1093/nar/gkl999.
7. Öztürk, H., Özgür, A., Ozkirimli, E.: DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 34(16), i821–i829 (2018). https://doi.org/10.1093/bioinformatics/bty593.
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. J. Jpn. Soc. Fuzzy Theory Intell. Inform. 29(5), 177–188 (2017). https://doi.org/10.3156/jsoft.29.5_177_2.
9. Davis, M.I., Hunt, J.P., Herrgard, S., et al.: Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29(11), 1046–1051 (2011). https://doi.org/10.1038/nbt.1990.
10. Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., Aittokallio, T.: Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative analysis. Journal of Chemical Information and Modeling. 54, 735–743 (2014). https://doi.org/10.1021/ci400709d.
11. Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H., Clevert, D.-A., Hochreiter, S.: Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chemical Science. 9, 5441–5451 (2018). https://doi.org/10.1039/c8sc00148k.
12. Karimi, M., Di, W., Wang, Z., Shen, Y.: DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics 35(16), 3329–3338 (2019). https://doi.org/10.1093/bioinformatics/btz118.
13. Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug–target interaction predictions. Brief Bioinform 2015;16:325–37.

14. He, T., Heidemeyer, M., Ban, F., et al.: SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J. Cheminform. 9(1), 24 (2017). https://doi.org/10.1186/s13321-017-0209-z.
15. Öztürk, H., Ozkirimli, E., Özgür, A.: WideDTA: prediction of drug-target binding affinity. arXiv preprint arXiv:1902.04166 (2019).
16. Shim, J.; Hong, Z.-Y.; Sohn, I.; Hwang, C. Prediction of drug–target binding affinity using similarity-based convolutional neural network. Sci. Rep. 2021, 11, 4416, DOI: 10.1038/s41598-021-83679-y.
17. Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; Venkatesh, S. GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics 2021, 37, 1140–1147, DOI: 10.1093/bioinformatics/btaa921
18. Yuan, W.; Chen, G.; Chen, C. Y.-C. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. Briefings in Bioinformatics 2022, 23, bbab506 DOI: 10.1093/bib/bbab506.
19. Zhao, Q.; Duan, G.; Yang, M.; Cheng, Z.; Li, Y.; Wang, J. AttentionDTA: Drug–Target Binding Affinity Prediction by Sequence-Based Deep Learning With Attention Mechanism. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2023, 20, 852–863, DOI: 10.1109/TCBB.2022.3170365.
20. Wu, H., Liu, J., Jiang, T., et al.: AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Neural Netw. 169, 623–636 (2024). https://doi.org/10.1016/j.neunet.2023.11.018.
21. Zhou, C.; Li, Z.; Song, J.; Xiang, W. TransVAE-DTA: Transformer and variational autoencoder network for drug-target binding affinity prediction. Computer Methods and Programs in Biomedicine 2024, 244, 108003, DOI: 10.1016/j.cmpb.2023.108003.
22. Pei, Q., Wu, L., Zhu, J., et al.: Breaking the barriers of data scarcity in drug–target affinity prediction. Brief. Bioinform. 24(6), bbad386 (2023). https://doi.org/10.1093/bib/bbad386.
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 6441–6451. Association for Computational Linguistics, Stroudsburg (2020)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998–6008. Curran Associates, Red Hook (2017)
25. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019), pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019)
26. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. arXiv preprint arXiv:1705.07115 (2018).
27. Gönen M., Heller G. Concordance probability and discriminatory power in proportional hazards regression Biometrika, 92 (4) (2005), pp. 965-970