# Drug target interaction prediction via multi-task co-attention

Yuyou Weng, Xinyi Liu, Hui Li and Chen Lin*

Department of Computer Science,
Xiamen University,
Xiamen, Fujian, China
Email: yuyouweng@stu.xmu.edu.cn
Email: xinyiliu@stu.xmu.edu.cn
Email: hui@xmu.edu.cn
Email: chenlin@xmu.edu.cn
*Corresponding author

Yun Liang

Department of Information,
South China Agricultural University,
Guangzhou, Guangdong Province, China
Email: sdliangyun@163.com

**Abstract:** Drug-Target Interaction (DTI) prediction is a key step in drug discovery and drug repurposing. A variety of machine learning models are considered to be effective means of predicting DTI. Most current studies regard DTI prediction as a classification task (that is, negative or positive labels are applied to indicate the intensity of interaction) or regression tasks (numerical value is used to measure detailed DTI). In this article, we explore how to balance bias and variance through a multi-task learning framework. Because the classifier is more likely to produce higher bias, and the regression models are more prone to create a significant variance and overfit the training data. We propose a novel model, named Multi-DTI, that can predict the precise value and determine the correct labels of positive or negative interactions. Besides, these two tasks are performed with similar feature representations of CNN, which is adopted with a co-attention mechanism. Detailed experiments show that Multi-DTI is superior to state-of-the-art methods.

**Keywords:** multi-task learning; scientific data management; data integration; drug target interaction.

**Reference** to this paper should be made as follows: Weng, Y., Liu, X., Li, H., Lin, C. and Liang, Y. (2020) 'Drug target interaction prediction via multi-task co-attention', *Int. J. Data Mining and Bioinformatics*, Vol. 24, No. 2, pp.160–176.

**Biographical notes:** Yuyou Weng received her BS degree from Department of Computer Science of Xiamen University, in 2016. Currently, she is pursuing the MS degree from the School of Information Science and Technology, Xiamen University, China. Her research interests include machine learning and data mining.

Xinyi Liu is a first-year master student in the School of Informatics, Xiamen University. She received her BEng degree from Xiamen University, in 2018. Her professional interests include data mining and information retrieval.

Hui Li is currently an Assistant Professor in the School of Informatics, Xiamen University. He received his BEng degree in Software Engineering from Central South University, in 2012, and MPhil and PhD degrees in Computer Science from University of Hong Kong, in 2015 and 2018, respectively. His research interests include data mining and data management with applications in recommender systems and knowledge graph.

Chen Lin received her BEng and PhD degrees from Fudan University, China in 2004 and 2010, respectively. Currently, she is an Associate Professor with School of Informatics, Xiamen University, China. Her research interests include web mining and recommender systems.

Yun Liang is an Associated Professor in the College of Mathematics and Informatics, South China Agriculture University, Guangzhou, China. She received her MSc and PhD degrees in the School of Information Science and Technology at Sun Yat-sen University, in 2005 and 2011, respectively. From 2016 to 2017, she worked in Simon Fraser University. Her research interests include computer vision, image computation and machine learning.

*This article is a revised and expanded version of a paper entitled 'Drug target interaction prediction using multi-task learning and co-attention', presented at the 'Proceedings – 2019 IEEE International Conference on Bioinformatics and Biomedicine BIBM 2019', San Diego, CA, USA, 18–21 November 2019.*

## 1 Introduction

Drug-Target Interaction (DTI) prediction has played a vital role in drug repurposing and drug discovery. Identifying the biological principle of potential intervention goals can enable efficient drug development. It is environment-friendly and superb for drug development to identify the organic starting origin of a disease, and the manageable ambitions for intervention. Naturally, there has been numerous research for DTI prediction in the bioinformatics community (Palma et al., 2014). Mainly, significant lookup interest has these days dedicated to computational DTI systems (Pahikkala et al., 2015; Luo et al., 2016; He et al., 2017; Öztürk et al., 2018; Tsubaki et al., 2018) to exchange ordinary biochemical experimental methods.

Most structures of DTI prediction models primarily based on machine learning methods (Öztürk et al., 2018; Tsubaki et al., 2018). The advantages of machine learning methods are scalable, time-saving, and labour-efficient. Machine learning-based DTI methods even have been extra promising with the growing quantity of publicly on hand data.

The chemical compound sequence of a drug and an amino acid sequence of a protein are usually treated as the input of DTI learners. The prediction results are processed and generated from extracted feature representations. Existing DTI prediction methods typically cope with one project only, i.e., precise numerical output values to measure the interaction (He et al., 2017; Öztürk et al., 2018), or output binary labels for positive or

negative signals (Tsubaki et al., 2018; Ni et al., 2018). The former models cope with numerical data to implement a regression task. And the latter performs a classification task with categorical data. From the view of multi-task learning, both the two kinds of DTI prediction methods only achieve a single task.

For most present multi-task DTI prediction methods, the problem needs to solve is the trade-off between bias and variance. Besides, numerical values tend to optimise with regression models. The numerical values often lead to an over-fitting result, so we will perhaps stumble upon significant variance. On the other hand, as a result of lacking numerical analysis with fine grain, classifiers can find classification segmentations. Thus, there may be high bias predicted on test data, which don't have observed labels. We consider that the DTI prediction models' assessment metrics are now and then conflicting; the trade-off between bias and variance is more and more severe. The generally comparison metrics is adopted to measure the performance of DTI prediction models, such as Area Under the Receiver Operating Characteristic (AUROC) curve and Mean Square Error (MSE). We can easily find that a DTI prediction classifier is probably to do better on metrics of classification, like AUROC (He et al., 2017), but on regression metrics, such as MSE (Öztürk et al., 2018), its performance may be poorly.

This paper is an extension of our previous work (Weng et al., 2019). In this paper, a DTI prediction model is proposed, which predicts the precise values of strength for DTI and determines the most probable boundary of positive or negative interactions. The balance between bias and variance are explored through a multi-task studying framework. With the rigorous experiment, we figure out that the overall performance of the DTI prediction with the aid of combining the two tasks can increase in phrases of a variety of well-known comparison metrics, like MSE, BCE, and AUROC.

The multi-task in the proposed DTI model, including regression and classification tasks, are both performed on a shared feature representations network. Previous systems use hand-crafted feature representations, e.g., many sorts of handmade features are mixed in He et al. (2017), consisting of PageRank ratings on homogeneous networks, occurrence information of drugs and targets, and so on. Different from them, Multi-DTI is expertise-driven. There has been a rapid development in feature representation in various fields nowadays, resulting from the latest methods of deep neural networks (Öztürk et al., 2018). In deep neural networks, the task-specific feature representation is studied and optimised all through the training process. To extract feature representation, CNN (Öztürk et al., 2018), GNN (Tsubaki et al., 2018) and GCN (Nguyen et al., 2019) are adopted in the researches of DTI prediction methods But these deep learning methods also have a shortage. For example, in terms of finding long-distance dependencies, it is hard for CNN and RNN to see the relationships between drug sequences and target sequences.

In this paper, a co-attention mechanism applies to CNN blocks to enhance the relationship between drug sequence and target sequence. We encode the long-distance dependencies of drug sequences and target sequences by way of inserting the influence of attention into elements in the corresponding sequence. The attention mechanism requires considerably much less time to train than CNN or RNN. Furthermore, the drug sequence is attended to the generation process of the target sequence with a co-attention mechanism; at the same time, the drug sequence is also influenced by the target sequence concurrently. As an extension of our previous work (Weng et al., 2019), we evaluate our model in another new metric, CI and $r_m^2$.

Generally speaking, our contributions to this work are summarised in the following two terms. (1) A new DTI model is proposed to perform multi-task. It can predict the numerical strength and binary interaction classification of DTI. (2) We make use of a novel co-attention into deep learning neural networks for representing drug/target sequences. To validate our assumption, we do enormous experiments on the proposed DTI model and prove that it performs better than state-of-the-art methods.

The structure of the paper is listed as follows. In Section 2, we shortly overview the related work. The framework of our DTI prediction model is introduced in Section 3. In Section 4, we present the experimental results and analyse the innovative effects. Finally, we make a conclusion to this work in Section 5.

## 2 Related work

Related works are briefly reviewed with two lines as follows.

### *2.1 DTI prediction*

For drug discovery and design, DTI is crucial. Because it is exceptionally highly-priced and time-consuming to use experimental biochemical strategies for DTI identification, computational DTI prediction techniques have acquired developing recognition in literature. To predict DTIs, there are two main traditional computational strategies called ligand-based methods (Keiser et al., 2007) and molecule docking methods (Cheng et al., 2007). When goal proteins have little binding ligands, the first strategies will be useless. Contrastly, when 3D buildings of goal proteins are no longer available, the second technique will be computationally highly-priced and unable to provide correct predictions (Chen et al., 2016). Thus, to infer DTI, researchers have come up with a lot of machine learning-based techniques. There are two most important kinds of DTI learners.

One kind equals DTI prediction with binary classification tasks in which we mark the known DTIs as positive and unknown DTIs as negative (Ding et al., 2013) or without a label (i.e., PU Learning) (Peng et al., 2017). The latest work (Ni et al., 2018) defines unknown DTIs as lacking labels. Like Random Forest (RF) (Pahikkala et al., 2015; Li et al., 2015) and Support Vector Machine (SVM) (Shar et al., 2016), researchers also adopt traditional regression models. The other kind tries to get a numerical value named drug-binding affinity. Gradient boosting method (He et al., 2017) and deep neural networks are included in the regression models. Deep neural networks consider regression loss and are used most recently (Guney et al., 2016; Zhang et al., 2017).

### *2.2 Representation learning*

Machine learning-based techniques include regression methods and classification methods. They function on drugs and targets' feature representations. Previously, the area expertise largely influences the feature representations, e.g., molecule docking and descriptors (Cheng et al., 2007; Zhang et al., 2017). Benefit from the deep learning's super success, for drug and target representations, some network descriptors have been used. Most of them concentrate on drug-target pairs to extract their topological similarity. For example, DBN (Wen et al., 2017) builds a stack of Restricted Boltzmann Machine (RBM) (Wang and Zeng, 2013), DeepWalk (Zong et al., 2017) computes similarities in a

linked tripartite network. Convolutional Neural Network (CNN) (Krizhevsky et al, 2012, 2017) is a network building that can perform adequately with grid data, and in lots of computer vision tasks, it has been utilised efficiently. Because DTI prediction is also relevant to grid-like data, CNN is applied in a range of deep CTI forecast indicators, like CNN grading function (Ragoza et al., 2017), DeepDTA (Öztürk et al., 2018), OnionNet (Zheng et al., 2019) and so on. Furthermore, DeepCPI (Tsubaki et al., 2018) makes use of Graph Neural Network (GNN) (Ying et al., 2018).

Because of the poor scaling characteristics of CNN and GNN, it is not easy for them to grasp long-distance dependencies in the sequence. The self-attention mechanism solves this problem by associating a single sequence's different positions to calculate the representation of the same process. The self-attention mechanism produced gratifying results when used in lots of natural languages processing models, for example, in transformer (Vaswani et al., 2017). About the promotion of self-attention, one way is focusing attention jointly on two sequences so that it can turn into the co-attention mechanism (Ma et al., 2017). Some co-attention models are coarse-grained, and some are fine-grained (Fan et al., 2018). Coarse-grained models utilise the embedding of other data as a query to calculate the attention of each input. In this work, we retain drug and protein sequences' topology information by undertaking the co-attention mechanism.

## 3 Method

In this section, a novel model, Multi-DTI, is proposed as a DTI predictor with a multi-task studying framework. We have introduced some previous work in our recent paper (Weng et al., 2019). The multi-task means Multi-DTI aim to perform regression and classification tasks at the same time. Multi-DTI is a kind of supervised model, i.e., the model is fed with given labels. In DTI prediction, the input of target is amino acid sequence, the input for drugs is Simplified Molecular-Input Line-Entry System (SMILES) representations of the chemical compound sequences, as well as the supervision signals as a value between 0 and 1. In the multi-task setting of Multi-DTI, supervision indicators consist of the precise strength value of DTI and the positive or negative DTI labels, which is also called numerical values and binary variables. Firstly, the process of generating the supervision alert signals is described for the input of Multi-DTI. Next, the network structure of Multi-DTI is introduced in detail and introduce every part of its components, whose structure is similar with our previous work (Weng et al., 2019). Finally, we set a loss function to mix the regression task and classification task.

In the paper, lower-case letters are used for indices, and upper-case letters are used for functions and scalars; lower-case bold-face letters are used for vectors, while upper-case bold-face letters are used for matrices.

Supposed that there are $M$ drugs and $N$ targets, which are denoted as $D=[D_1,...,D_M]$ and $T=[T_1,...,T_N]$ respectively. We represent their supervision signals as $Y\in\mathcal{R}^{M\times N}$. There are two ways to generate supervision signal $Y$, which have been mentioned in the previous Section 1. To perform a regression task, the supervision signal $Y_{i,j}\in(0,+\infty)$, that is, $Y_{i,j}=R_{i,j}$, where $R_{i,j}$ represents the normalised DTI value. To perform a classification task, the supervision signal $Y_{i,j}\in\{0,1\}$, where a positive label of

DTI is represent as $Y_{i,j}=1$, otherwise, $Y_{i,j}=0$. It is worth mentioning that pairs between drugs and targets with unknown interactions are often ignored and didn't take into account.

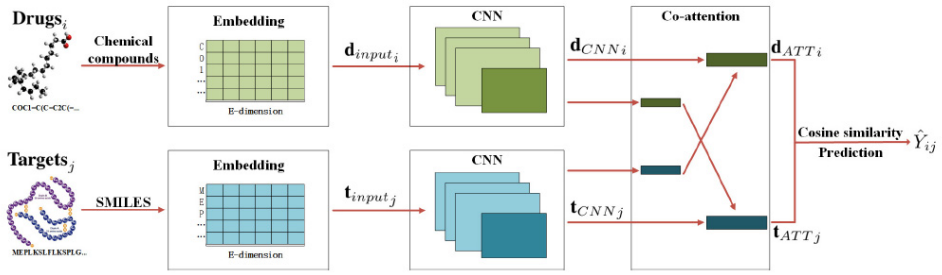
We combine the classification and regression tasks to complete the multi-task. Therefore, the supervision signals with a given range are constructed in the form of a matrix as follows.

$$Y_{ij}=\begin{cases}0, & if\ R_{ij}=unknown\\ \dfrac{R_{ij}}{\max(R)}, & else\end{cases} \tag{1}$$

It is evident that the real observed DTI values $R$ are equal to $Y$ due to the fact they point out the strength of DTI. We reset the normalised strength of DTI for multi-task. In detail, We reduce each element with the same proportion of $R$ according to the maximal value in $R$. In the meantime, to mine more information, we mark it as zero if the DTI has unknown labels.

The overview of Multi-DTI is shown as Figure 1, the embeddings of drug and target sequence $d_i, t_j$ go with the flow through an embedding layer, a feature representation component, and a prediction layer. Multi-DTI is aiming to generate the prediction value of $\hat{Y}_{ij}$. The model parameters of Multi-DTI is given as $\Omega$. $\hat{Y}_{ij}=F\left(d_i,t_j\mid\Omega\right)$, to approximate $Y_{ij}$ with these parameters.

**Figure 1** Network architecture of model multi-DTI




## *3.1 Embedding layer*

Firstly, we embed integer/label encoding to characterise each data as inputs. For drug sequences, we scan about 2000 SMILES sequences, which is collected the Pubchem Database. We allocate 64 labels for each element. For example, letters "C", "N", "=" is allocated with different labels. Each label is represented with a special integer, e.g. "C":1, "=":22, "N":3 and so on. The process is similar to our previous work (Weng et al., 2019). For example, we transform a label embedding for a drug sequence "CN=C=O", the result is given like below: $[C\ N=C=O]=[1\ 3\ 22\ 1\ 22\ 5]$

Secondly, an embedding function $G(\cdot)$ is utilised to change the specific sequence into a $E$-dimensional float vector as feature representation, where $V$ is the count

number of drug label tags. That is, $G(V) \in \mathcal{R}^E$, Throughout the whole training process, the learning of the embedding function is continued. In order to get separate integer label embeddings, we concatenate all integer label embeddings of the same drug sequence in rows. For intending operations on CNN, we assemble a matrix for every drug $i$. Next, we will leave out subscript index $i$ and $j$ without ambiguity, and $D_{input} \in \mathcal{R}^{D \times E}$ is used to represent the input of CNN, where $D$ is the maximal size of a drug sequence.

Finally, we perform similar operations for protein sequences. In terms of a target sequence, we scan 550,000 protein sequences from UniProt Database. Next, we extract 25 different labels for each element to get a separate embedding. We encode the target sequences with integer label encodings. Then in order to assemble the input of CNN $T_{input} \in \mathcal{R}^{T \times E}$, the encodings are concatenated in rows.

It is evident that both drug and target sequences have various lengths. We determined to select a size limit to create a fixed representation form, i.e., we set $D$ is the maximum length of drugs and $T$ for targets. The drug sequences which are shorter than $D$ would be filled with zeros, while the part out of range would be cut-off.

### *3.2 Feature representation component*

$D_{input}$ and $T_{input}$ are utilised as input in this component. The CNN networks made up the most kernel part of the feature representation component to encode sequential information. There are two CNN blocks. One of them is set for drug sequence and another for the target sequence. The output of these CNN blocks is $D_{CNN}$ and $T_{CNN}$. Every CNN block is consists of three consecutive 2D-convolutional layers and a max-pooling layer, which are used to filter data.

Next, a co-attention mechanism is put behind CNN blocks to cope with $D_{CNN}$ and $T_{CNN}$, respectively, and output $D_{ATT}$ and $T_{ATT}$. In intuition, some drugs and unique targets affect each other. The underlying assumption of the co-attention mechanism is that the attention weights could capture the influence of drugs and targets.

Specifically, with a given drug vector $d_{CNN}$ and a target vector $t_{CNN}$, the drug output vector $d_{ATT}$ is generated in the co-attention layer. Each element in the input of drug feature representation $d_{CNN}$ is multiplied with its corresponding attention weight of target feature representation $d_{ATT}$:

$$d_{ATT} = d_{CNN} \odot \Gamma\left(t_{CNN}\right) \tag{2}$$

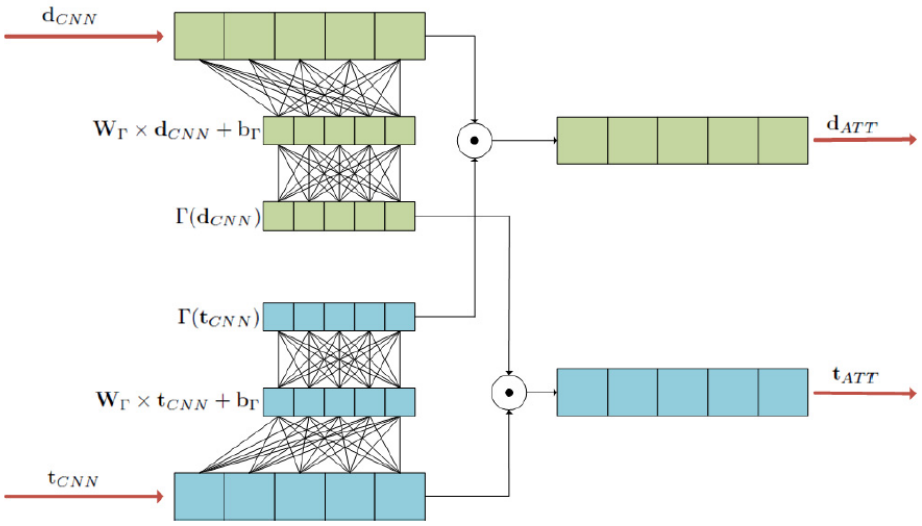
where $\Gamma(\cdot)$ function is used to calculate the importance of $t_{CNN}$, which is part of attention mechanism. We adopt a single layer feed forward network to get the attention weights, which is defined as below.

$$\Gamma\left(t_{CNN}\right) = A\left(W_\Gamma \times t_{CNN} + b_\Gamma\right) \tag{3}$$

where $W_\Gamma$ and $b_\Gamma$ represent the matrix of weight parameters and the vector of bias parameters of the connected network, respectively. To normalise the attention weights parameters, we defined softmax function $A(\cdot)$.

We also carry out a comparable construction of co-attention on every goal vectors $t_{CNN}$ extracted utilising CNN. In Figure 2, we display the architecture and application of the co-attention mechanism.

**Figure 2** Illustration of parallel coarse-grained co-attention




### *3.3 Prediction layer*

$D_{ATT}$ and $T_{ATT}$ are utilised as input in this layer. In the last component, the feature representation sequences of drugs and targets have been transformed into new vectors in latent space. We utilised the similarity between feature vectors if drugs and targets, which is measured to predict the final result, i.e., computing the cosine similarity between $d_{ATTi}$ and $t_{ATTj}$. In the prediction layer, In intuition, The DTI value would be higher if its corresponding drug feature representation is more similar to the target feature representation. Therefore, we describe the output of $\hat{Y}_{ij}$ as:

$$\hat{Y}_{ij} = \cos ine\left(d_{ATTi}, t_{CNNj}\right) = \frac{d_{ATTi}^{T} t_{ATTj}}{\left\|d_{ATTi}\right\| \cdot \left\|t_{ATTj}\right\|} \tag{4}$$

To perform a classification task, Normalised Cross Entropy (NCE) loss (Xue et al., 2017) is adopted to be part of final loss function . Given $\hat{Y}_{ij}$ as the predicted label and $Y_{ij}$ as the real label, we describe NCE as:

$$NCE = \sum_{\forall(i,j)} \left[ Y_{ij} \log \hat{Y}_{ij} + \left(1 - Y_{ij}\right) \log\left(1 - \hat{Y}_{ij}\right) \right] \tag{5}$$

It is worth to mention that in case of $Y_{ij} \in \{0,1\}$, NCE as equation (5) could be regarded as equal to the traditional Binary Cross Entropy. Therefore, the NCE loss function will lead to a more definite boundary between positive and negative prediction results.

To perform a regression task, we need to use the regression loss function as Mean Square Error (MSE).

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left( Y_{ij} - \hat{Y}_{ij} \right)^2 \tag{6}$$

where $M, N$ means the maximum count number of drugs and targets, respectively.

Finally, we make a comprehensive consideration to MSE and NCE as the loss function of Multi-DTI, which is described as below:

$$Loss = \delta \times MSE + (1 - \delta) \times NCE \tag{7}$$

where $\delta$ is a defined parameter that adjusts the mixture proportion of MSE and NCE.

*Discussion*: The way to output a high-quality prediction with a given feature space of DTI has attracted many research interests. It is crucial to figure out that a multi-layer perception has been used in numerous existing researches (Öztürk et al., 2018; Tsubaki et al., 2018; Nguyen et al., 2019). In this paper, we use a new idea different from these works. Cosine similarity between drug and target feature representation has been used in Multi-DTI directly. The advantage of the new idea is easy for learning. Besides, fewer parameters of a model could speed up the training process and save more time.

## 4 Experiment

The next lookup questions are studied in this part.

*RQ1*: Will the latest techniques work worse than the Multi-DTI model?

*RQ2*: Will the result of Multi-DTI is affected by the parameters?[1]

### *4.1 Experimental setup*

To validate our model, the Kinase Inhibitor BioActivity (KIBA) data set[2] is exploited. DTI strength is an integration of Kd, Ki, and IC50 scores, composing our data set. About 25% of the labels are positive labels, while the rest are negative labels. The fundamental information of the data sets is illustrated in Table 1. To improve the stability and robustness of the model, we apply 5-fold cross-validation.

**Table 1** Statistics of the KIBA data set

| | *# Number of drugs* | *# Number of targets* | *# Number of drug-target pairs* |
|---|---|---|---|
| KIBA | 2008 | 185 | 92,706 |

For SMILES, we set a size limit of maximal 100 characters, and for protein sequences, we set the limit of maximum as 1000 in this experiment. As Öztürk et al. (2018) said, the maximum size covers at least 80% of the proteins and 95% of the compounds. We set the embedding dimension as $E = 128$ and use 256 as batch size. Adam is chosen as the optimiser, and with the learning rate as $1e-5$, convergence is declared for 200 Epochs.

**Table 2** Comparison result

| *Method* | *Drugs* | *Targets* | *Prediction* | *Loss* | *MSE* | *BCE* | AUROC | AUPR |
|---|---|---|---|---|---|---|---|---|
| PUDTI | Descriptor | Descriptor | Concate | Hinge loss | 0.201 | 0.553 | 0.804 | 0.278 |
| DTINet | RWR | RWR | Matrix completion | MSE | 0.292 | 0.794 | 0.695 | 0.368 |
| DeepCPI | GNN | CNN | Concate Attention | BCE | 0.197 | 0.693 | 0.476 | 0.250 |
| DeepDTA | CNN | CNN | Concate FFN | MSE | 0.002 | 0.046 | 0.814 | **0.436** |
| CNN-basic | CNN | CNN | Cosine | BCE | 0.002 | 0.043 | 0.853 | 0.402 |
| CNN-Multi | CNN | CNN | Cosine | Multi-task | 0.002 | 0.041 | 0.848 | 0.419 |
| Multi-DTI | CNN Attention | CNN Attention | Cosine | Multi-task | **0.002** | **0.034** | **0.888** | 0.424 |

### *4.2 Comparative study*

Multi-DTI model is compared with the baseline models listed as follows:

1) *PUDTI (Peng et al., 2017)*: an optimisation model based on SVM. It is learned on negative labels that are extracted primarily based on positive-unlabelled learning. On the basis of on PaDEL-Descriptors of drugs, PAACs and PSSM of targets, We can describe every pair of input explicitly. The loss function of PUDTI is the Hinge loss function.

2) *DTINet (Luo et al., 2017)*: a regression model. It uses a given heterogeneous network to predict DTI. In the heterogeneous network, numerous information related to drugs is integrated. We use Random Walk with Restart (RWR) to extract the feature representations. With given drug feature representations $P$ and target feature representations $Q$, The model can learn the feature space mapping $Z$ by minimising MSE loss function $\min_Z \left(\hat{Y} - PZQ\right)^2$.

3) *DeepCPI (Tsubaki et al., 2018)*: an end-to-end deep learning model. A Graph Neural Network (GNN) block can learn drug representation, and with a CNN block, it can learn the protein representation. On the basis of the concatenation of drug and protein feature representations, a neural attention mechanism is adopted by DeepCPI to predict DTI. The loss function of DeepCPI is BCE.

4) *DeepDTA (Öztürk et al., 2018)*: a deep neural architecture with two separate CNN modules, which can study the feature representations from drugs and targets, respectively. To predict the value of DTI, a fully-connected feed-forward layer is performed by DeepDTA on the concatenation of drug and protein representations. The loss function of it is MSE.

To validate the influence of co-attention and multi-task learning, we made a comparison between Multi-DTI and its variations.

5) *CNN-basic*: a Multi-DTI model's variant. It uses primary CNN modules on drug and target sequences and then calculates their cosine similarity. The loss function of it is BCE.

6) *CNN-Multi*: another Multi-DTI model's variant, which utilises multi-task loss (i.e., equation (7)) on CNN blocks.

We consider the methods in metrics as MSE, BCE, AUROC, and AUPR.

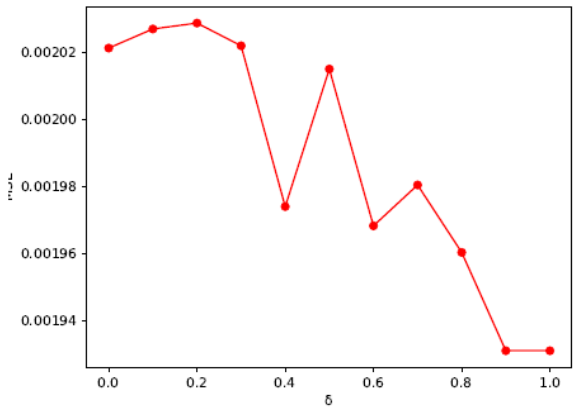
The following conclusions can be obtained from Table 2. (1) Multi-DTI gets the best result when we consider MSE, BCE and AUROC. It performs comparably on AUPR, too. (2) State-of-the-art methods can only cope with a single task. For example, DeepCPI defines its loss function as a classification, so when we consider BCE, AUROC, and AUPR metrics, it has the best performance. However, its performance in MSE is poor. Contrarily, Multi-DTI performs nicely in the comparison metrics because they undertake multi-task loss. (3) The prediction performance can be improved by the attention mechanism. As we can see, the BCE of CNN-Multi is increased by about 17% than Multi-DTI. (4) Multi-DTI achieves the second-best result, and DeepDTA obtains the best result on AUPR. The likely reason may be the trade-off of loss function between MSE and BCE. In the next subsection, we will find out about the relationship between $\delta$ and metrics.

### *4.3 Effect of parameters*

Here, we continue to explore *RQ2* and learn what effects will be caused by the parameters changing.

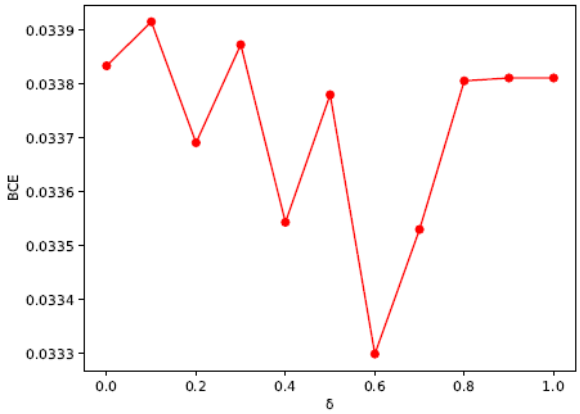
Firstly, we study the influence of the proportion parameter of $\delta$. In our loss function, $\delta$ controls the proportion of MSE. We set $\delta = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and the performance of MSE, BCE, AUROC and AUPR are presented in Figure 3.

**Figure 3** Performance with different values of $\delta$. (a) MSE (b) BCE (c) AUROC (d) AUPR

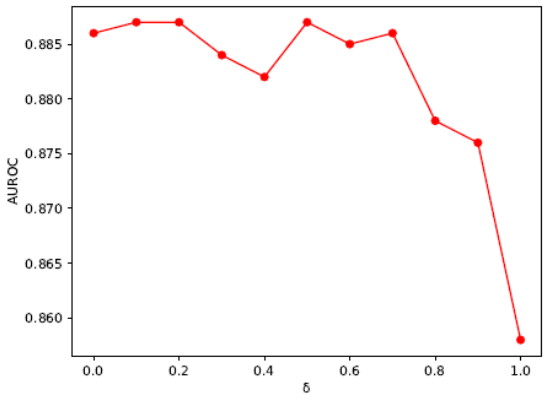



(a)

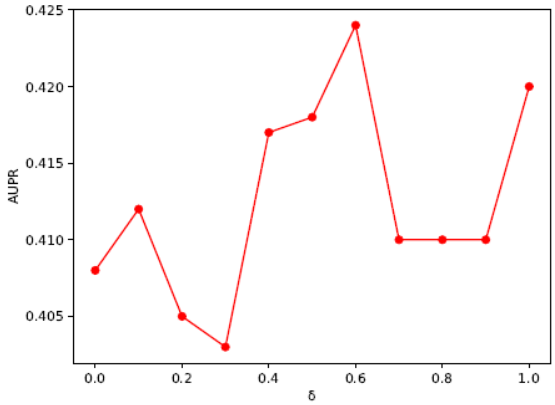
**Figure 3** Performance with different values of $\delta$. (a) MSE (b) BCE (c) AUROC (d) AUPR (continued)




(b)




(c)



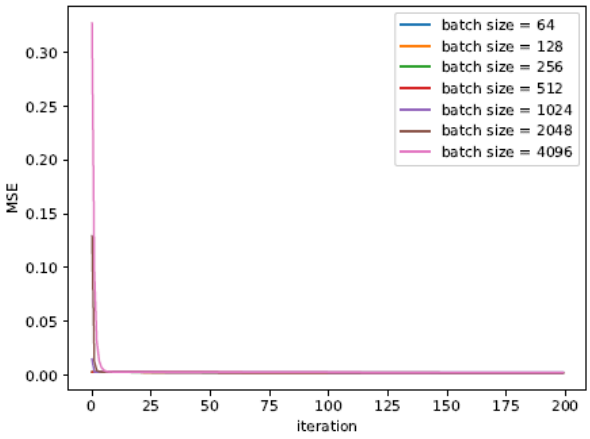

(d)

From Figure 3, we find that as $\delta$ increases, the MSE and AUROC result generally becoming worse and worse. Even the Multi-DTI model only performs a regression task, the MSE and AUROC performance getting the worst point When $\delta = 1.0$. BCE's trend is non-monotonic, so is AUPR. The highest AUPR and lowest BCE is obtained when $\delta = 0.6$. Therefore, $\delta = 0.6$ is regarded as the most suitable value of $\delta$.

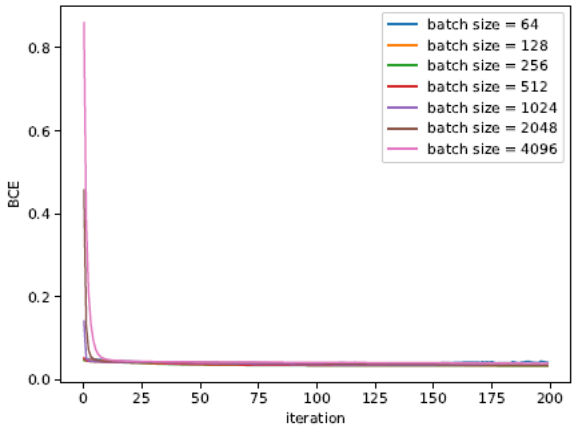
Next, we find out what effects will be caused when batch size changing. Batch size is a parameter which is very critical in the process of training. From 64 to 4096, several different batch sizes are set to test. We plot the curve of performance of MSE, BCE, AUROC, and AUPR at each epoch.

As illustrated in Figure 4, we can discover that there is no distinct tendency in MSE and BCE. However, a bigger batch size leads to slower convergence in AUROC and AUPR. When batch size is 256, the model quickly converges to the best AUROC and AUPR results.

**Figure 4** Convergence of Multi-DTI with different batch size. (a) MSE (b) BCE (c) AUROC (d) AUPR

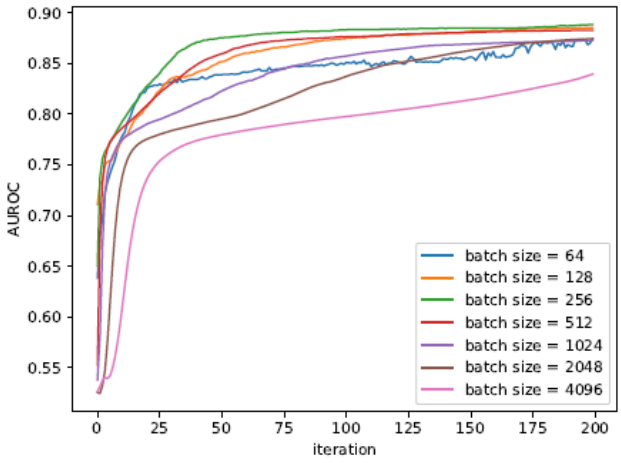



(a)

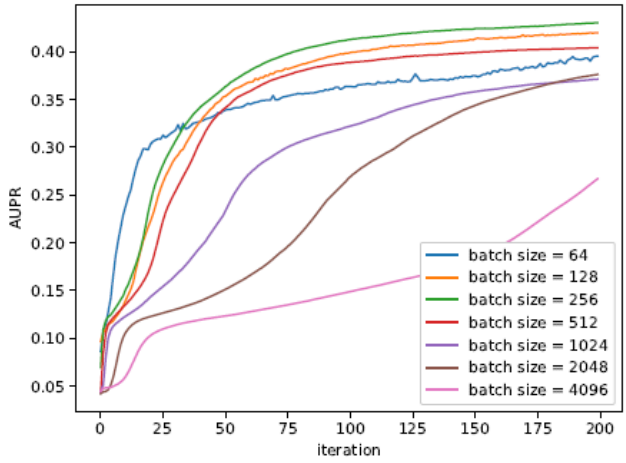



(b)

**Figure 4** Convergence of Multi-DTI with different batch size. (a) MSE (b) BCE (c) AUROC (d) AUPR (continued)




(c)




(d)

## *4.4 Alternative metrics*

Besides MSE, BCE, AUROC, and AUPR, some new metrics, CI and $r_m^2$, are introduced to measure the performance of Multi-DTI.

CI, also called concordance index, is used to measure the prediction of models. We can calculate the metric by

$$CI = \frac{1}{N} \sum_{y_i > y_j} H\left(\hat{y}_i - \hat{y}_j\right) \tag{8}$$

where $\hat{y}_i$ and $\hat{y}_j$ is corresponding predicted value of $y_i$ and $y_j$ respectively. And $H(x)$ is a step function as follow:

$$H(a-b)=\begin{cases}1, & if\ a>b;\\ 0.5, & if\ a=b;\\ 0, & if\ a<b\end{cases} \tag{9}$$

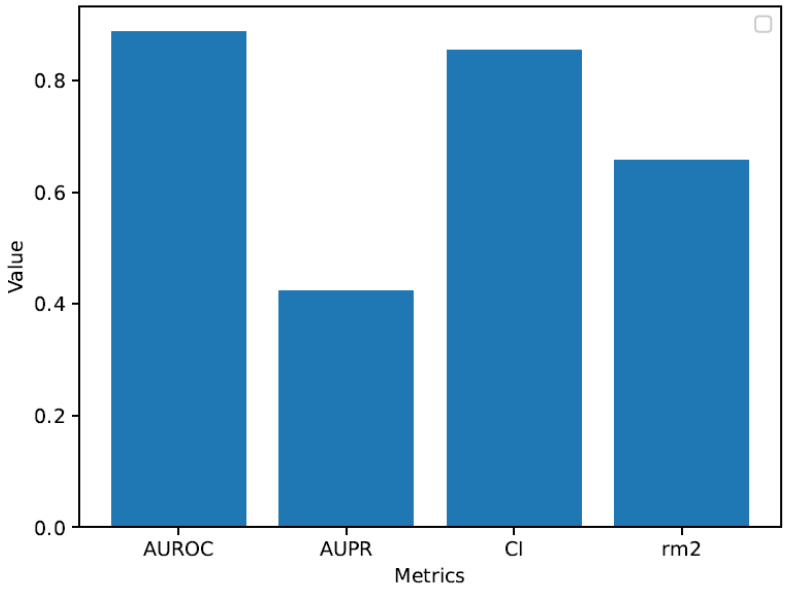
$r_m^2$ is used to characterise the quality of a fit through changes in the data. The computation of $r_m^2$ is shown as follow.

$$r_m^2=r^2\times\left(1-\sqrt{r^2-r_0^2}\right) \tag{10}$$

where $r^2$ is the squared correlation coefficient values between the observed real data and predicted values with intercept, while $r_0^2$ means the squared correlation coefficient values without intercept.

A more significant CI and $r_m^2$ both stand for a better performance of the model. We plot the metrics of Multi-DTI in Figure 5.

**Figure 5** The performance of Multi-DTI in different metrics




## 5 Conclusion

Multi-DTI is proposed as a new DTI predictor, which aims to solve the trade-off between bias and variance. The feature representations is based on a shared network. The model learned about drug and target feature representations and made an innovation by adding a

co-attention mechanism into traditional CNN blocks. Besides, the model primarily based on multi-task learning, which tries to perform both the regression and classification loss. With 5-fold cross-validation experiments, we prove that Multi-DTI performs better than state-of-the-art DTI prediction methods. In our future work, we are looking forward to improving the prediction result of DTI, in terms of multi-task, multi-view, and multi-modality learning.

## Acknowledgements

The project is supported by the Natural Science Foundation of China (grant nos. 61472335, 61972328).

## References

Chen, X. et al. (2016) 'Drug-target interaction prediction: databases, web servers and computational models', *Briefings Bioinform*, Vol. 17, No. 4, pp.696–712.

Cheng, A.C. et al. (2007) 'Structure-based maximal affinity model predicts small-molecule druggability', *Nature Biotechnology*, Vol. 25, No. 1, pp.71–75.

Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S. (2013) 'Similarity-based machine learning methods for predicting drug-target interactions: a brief review', *Briefings Bioinform*, Vol. 15, No. 5, pp.734–747.

Fan, F., Feng, Y. and Zhao, D. (2018) 'Multi-grained attention network for aspect-level sentiment classification', *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.3433–3442.

Guney, E., Menche, J. and Vidal, M. et al. (2016) 'Network-based in silico drug efficacy screening', *Nature Communications*, Vol. 7, pp.1–13.

He, T., Heidemeyer, M., Ban, F.Q., Cherkasov, A. and Ester, M. (2017) 'SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines', *Journal of Cheminformatics*, Vol. 9, No. 24, pp.1–14.

Keiser, M.J. et al. (2007) 'Relating protein pharmacology by ligand chemistry', *Nature Biotechnology*, Vol. 25, No. 2, pp.197–206.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Imagenet classification with deep convolutional neural networks', *Communications of the ACM*, Vol. 60, No. 6, pp.84–90.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) 'ImageNet classification with deep convolutional neural networks', *Communications of the ACM*, Vol. 60, No. 6, pp.84–90.

Li, H.J., Leung, K., Wong, M. and Ballester, P. (2015) 'Low-quality structural and interaction data improves binding affinity prediction via random forest', *Molecules*, Vol. 20, No. 6, pp.10947–10962.

Luo, H., Mattes, W., Mendrick, D.L. and Hong, H.X. (2016) 'Molecular docking for identification of potential targets for drug repurposing', *Current Topics in Medicinal Chemistry*, Vol. 16, No. 30, pp.3636–3645.

Luo, Y., Zhao, X. and Zhou, J. et al. (2017) 'A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information', *Nature Communications*, Vol. 10229, pp.383–384.

Ma, D., Li, S., Zhang, X. and Wang, H. (2017) 'Interactive attention networks for aspect-level sentiment classification', *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp.4068–4074.

Nguyen, T., Le, H. and Venkatesh, S. et al. (2019) 'GraphDTA: prediction of drug – target binding affinity using graph convolutional networks', *bioRxiv preprint*, Doi: 10.1101/684662.

Ni, S., Lin, C. and Zeng, X. et al. (2018) 'Drug target interaction prediction with non-random missing labels', *IEEE Computer Society*, pp.496–501.

Öztürk, H., Özgür, A. and Ozkirimli, E. (2018) 'Deepdta: deep drug-target binding affinity prediction', *Bioinform*, Vol. 34, No. 17, pp.i821–i829.

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A. and Aittokallio, T. (2015) 'Toward more realistic drug-target interaction predictions', *Briefings Bioinform*, Vol. 16, No. 2, pp.325–337.

Palma, G., Vidal, M.E. and Raschid, L. (2014) 'Drug-target interaction prediction using semantic similarity and edge partitioning', *Springer International Publishing*, Vol. 8796, pp.131–146.

Peng, L., Zhu, W. and Liao, B. et al. (2017) 'Screening drug-target interactions with positive-unlabeled learning', *Scientific Reports*, Vol. 7, No. 1, pp.1–17.

Ragoza, M., Hochuli, J. and Idrobo, E. et al. (2017) 'Protein-ligand scoring with convolutional neural networks', *Journal of Chemical Information and Modeling*, Vol. 57, No. 4, pp.942–957.

Shar, P.A., Tao, W.Y., Gao, S., Huang, C., Li, B.H., Zhang, W.J., Shahen, M., Zheng, C.L., Bai, Y.F. and Wang, Y. H. (2016) 'Pred-binding: large-scale protein-ligand binding affinity prediction', *Journal of Enzyme Inhibition and Medicinal Chemistry*, pp.1–8.

Tsubaki, M., Tomii, K. and Sese, J. (2018) 'Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences', *Bioinform*, Vol. 35, No. 2, pp.309–318.

Vaswani, A., Shazeer, N. and Parmar, N. et al. (2017) 'Attention is all you need', *arXiv preprint arXiv:1706.03762.*

Wang, Y. and Zeng, J. (2013) 'Predicting drug-target interactions using restricted Boltzmann machines', *Bioinform*, Vol. 29, No. 13, pp.126–134.

Wen, M., Zhang, Z.M., Niu, S.Y., Sha, H.Z., Yang, R.H., Yun, Y.H. and Lu, H.M. (2017) 'Deep-learning-based drug-target interaction prediction', *Journal of Proteome Research*, Vol. 16, No. 4, pp.1401–1409.

Weng, Y., Lin, C. and Zeng, X. et al. (2019) 'Drug target interaction prediction using multi-task learning and co-attention', *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp.528–533.

Xue, H.J., Dai, X. and Zhang, J. et al. (2017) 'Deep matrix factorization models for recommender systems', *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp.3203–3209.

Ying, R., He R. and Chen, K. et al. (2018) 'Graph convolutional neural networks for web-scale recommender systems', *ACM*, pp.974–983.

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S.Y., Zhu, F., Li, Z.R., Jiang, Y.Y., Chen, W.P. and Chen, Y.Z. (2017) 'A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks', *Briefings Bioinform*, Vol.18, No. 6, pp.1057–1070.

Zheng, L., Fan, J. and Mu, Y. (2019) 'OnionNet: a multiple-layer inter-molecular contact based convolutional neural network for protein-ligand binding affinity prediction', *arXiv preprint arXiv:1906.02418.*

Zong, N., Kim, H. and Ngo, V. et al. (2017) 'Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations', *Bioinform*, Vol. 33, No. 15, pp.2337–2344.

## Notes

1 The code and data set used in Multi-DTI are on hand. Available online at: https://github.com/XMUDM/Multi-DTI

2 Available online at: https://pubs.acs.org/doi/suppl/10.1021/ci400709d