Drug Target Interaction Prediction using Multi-task Learning and Co-attention

Yuyou Weng, Chen Lin
Department of Computer Science
Xiamen University
Xiamen, China

Xiangxiang Zeng
Department of Information
Hunan University
Changsha, China

Yun Liang
Department of Information
South China Agricultural University
Guangzhou, China

***Abstract*—Various machine learning models have been proposed as cost-effective means to predict Drug-Target Interactions (DTI). Most existing researches treat DTI prediction either as a classification task (i.e. output negative or positive labels to indicate existence of interaction) or as a regression task (i.e. output numerical values as the strength of interaction). However, classifiers are more prone to higher bias and regression models tend to overfit the training data to generate large variance. In this paper, we explore to balance the bias and variance by a multi-task learning framework. We propose an architecture to both predict accurate values of strength of interaction and decide correct boundary between positive and negative interactions. Furthermore, the two tasks are performed on a shared feature representation, which is learnt using a co-attention mechanism. Comprehensive experiments demonstrate that the proposed method significantly outperforms state-of-the-art methods.**

***Index Terms*—Drug Target Interaction Prediction, Multi-task Learning, Co-attention, Deep Neural Network**

I. Introduction

Drug-Target Interaction (DTI) prediction is one of the most important step in drug discovery and drug repurposing. Identifying the biological origin of a disease, and the potential targets for intervention, enables efficient and effective drug development. Naturally, DTI has been extensively researched in the bioinformatics community [1]. Particularly, considerable research attention has recently been devoted to computational DTI systems [2]–[6] to replace traditional biochemical experimental systems.

The majority of computational DTI systems are based on machine learning methods [5], [6]. Machine learning methods have the advantages of being time-saving, labor-efficient, and scalable. With the increasing amount of public available data, machine learning based DTI methods have been even more promising.

The input of DTI learners usually include a chemical compound sequence for a drug and an amino acid sequence for a protein. Feature representations are extracted and processed to generate numerical or categorical predictions. Existing DTI learners usually handle one task only, i.e. either output numerical values as strength of the interaction [4], [5], or output binary values as positive or negative interaction [6], [7]. From the perspective of machine learning, the former type of DTI learners implements a numerical regression task, the latter implements a classification task.

The problem of most existing DTI learners is the bias-variance trade-off. On one hand, numerical regression models are capable of tuning to individual values. However, we will possibly encounter large variance due to over-fitting of the numerical values. On the other hand, classifiers can capture class segmentations, at the expense of missing fine-grained numerical analysis in the process of discretization. Thus, high bias is expected on unseen test data. The bias-variance trade-off is more severe since evaluation metrics of the DTI predictors are sometimes conflicting. For example, commonly adopted evaluation metrics include **M**ean **S**quare **E**rror (MSE) and **A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic curve (AUROC). However, a classifier is likely to perform well on classification metrics, such as AUROC [4], but poorly on regression metrics, such as MSE [5].

In this work, we explore to balance the bias and variance by a multi-task learning framework. We propose a DTI system that can both predict accurate values of strength for all pairs of drug-target interactions and decide correct boundary between positive and negative interactions. We experimentally show that, by combining the two tasks, the DTI prediction performance can be boosted in terms of a set of common evaluation metrics, such as MSE and AUC.

The proposed DTI framework performs both tasks on a shared feature representations space. Previously the feature representations are hand-crafted, e.g. several kinds of hand-crafted features are combined in [4], including occurrence statistics of drugs and targets, PageRank scores on homogeneous networks and so on. This approach is obviously expertise-driven. Nowadays, feature representation in numerous domains has benefitted from the recent advances of deep neural networks [5], which learn and optimize task-specific feature representation during training time. In the literature of DTI prediction, CNN [5] , GNN [6] and GCN [8] are adopted for feature representation learning. However, it is difficult for CNN and RNN to capture long-distance dependencies in chemical compound sequences and amino acid sequences.

In this paper, we utilize a co-attention mechanism. The long-distance dependencies are encoded by putting more emphasis (i.e. attention) on relevant tokens in the whole sequence.

Corresponding author: chenlin@xmu.edu.cn. Chen Lin is supported by the Natural Science Foundation of China (no.61472335, 61972328).Yun Liang is supported by the National Natural Science Fund(61772209) and the Science and Technology Planning Project of Guangdong Province (2016A050502050)

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Compared with CNN or RNN, the attention component requires significantly less time to train. Furthermore, we let the drug sequence attend to the target sequence, while the target sequence attend to the drug sequence simultaneously (i.e. co-attention).

In summary, the contributions of this work are two-fold. (1) We introduce a new DTI framework to combine numerical, interaction strength prediction and binary, interaction classification. (2) We present to apply co-attention for representing drug/target sequences. We conduct extensive experiments to validate that the proposed framework outperforms state-of-the-art methods.

The rest of the paper is organized as follows. We briefly overview related work in Section II. The DTI framework is presented in Section III. We evaluate the framework and analyze the experimental results in Section IV. Finally, we conclude this work in Section V.

II. Related Work

We briefly review two lines of closely related studies.

A. DTI Prediction

DTI is fundamental to drug discovery and design. As biochemical experimental methods for DTI identification are extremely costly and time-consuming, computational DTI prediction methods have received a growing popularity in literature. Traditional computational methods to predict DTIs mainly include ligand-based methods [9] and molecule docking methods [10]. Ligand-based methods are ineffective when target proteins have little binding ligands , while molecular docking methods are computationally costly and fail to offer accurate predictions when 3D structures of target proteins are not available [11]. To overcome these problems, many machine learning-based methods have been proposed for inferring DTI. There are two major types of DTI learners.

The first type treats DTI prediction as a binary classification task, where known DTIs are labeled as positive and unknown DTIs are labeled as negative [12] or unlabeled (i.e. PU Learning) [13]. A recent work [7] considers unknown DTIs as missing labels. Traditional regression models such as random forrest (RF) [2], [14] and support vector machine (SVM) [15] are adopted. The second type attempts to predict drug-binding affinity, which is a numerical value. Regression models include gradient boosting method [4], and most recently, deep neural networks that apply a regression loss [16], [17].

B. Representation Learning

Machine learning based methods, including regression and classification methods, operate on feature representations of drugs and targets. Prior representations are heavily depended on domain expertise, e.g. molecule docking and descriptors [10], [17]. Thanks to the great success of deep learning, there are some network descriptors applied for drug and target representations. Most of them focus on extracting topological similarity from drug-target pairs. For example, DBN [18] constructs a stack of Restricted Boltzmann Machine (RBM [29]), DeepWalk [19] calculates similarities within a linked tripartite network. Convolutional Neural Network (CNN) [20] is a network structure that works well with grid data. CNN has been successfully applied in many computer vision tasks. As DTI prediction also involves grid-like data such as a molecular graph, CNN has been adopted in a variety of deep CTI predictors, such as CNN scoring function [21], DeepDTA [5], OnionNet [22] and so on. Furthermore, DeepCPI [6] utilizes Graph Neural Network (GNN) [23].

It is challenging for CNN and GNN to capture long-distance dependencies in sequences, due to their poor scaling properties. Self-attention mechanism, which relates different positions of a single sequence in order to compute a representation of the same sequence, addresses this challenge. Self-attention has generated promising performance in many natural language processing models, such as transformer [24]. An improvement of self-attention is to apply attention jointly on two sequences, which becomes the co-attention mechanism [25]. Co-attention models can be coarse-grained or fine-grained [26]. Coarse-grained models compute attention on each input, using an embedding of the other input as a query. In this work, we adopt the co-attention mechanism to preserve topology information in drug and protein sequences.

III. Method

In this section, we introduce a novel model, Multi-DTI, for drug-target interaction prediction under a multi-task learning framework. Multi-DTI operates in a supervised manner, i.e. the model is fed with **S**implified **M**olecular-**I**nput **L**ine-**E**ntry **S**ystem(SMILES) representations of the chemical compound sequences from drugs and amino acid from proteins, as well as the supervision signals. In our multi-task setting, supervision signals include the strength of interaction (i.e. numerical values) and the interaction labels (i.e. binary variables). We first describe how to construct the supervision signals for the convenience of Multi-DTI. Then, we describe in detail the network architecture and each of its component.

In the remaining of the paper, we use lower-case letters for indices, upper-case letters for scalars and functions, lower-case bold-face letters for vectors, and upper-case bold-face letters for matrices.

Suppose there are M drugs, denoted as $\mathbf{D} = [\mathbf{D}_1, \cdots, \mathbf{D}_M]$, and N targets, denoted as $\mathbf{T} = [\mathbf{T}_1, \cdots, \mathbf{T}_N]$. The supervision signals are represented as $\mathbf{Y} \in \mathcal{R}^{M \times N}$. As mentioned in section I, previously there are two approaches to construct $\mathbf{Y}$. For a classification task, $Y_{i,j} \in \{0, 1\}$, where $Y_{i,j} = 1$ represents a positive interaction, otherwise, $Y_{i,j} = 0$. For a regression task, $Y_{i,j} \in (0, +\infty)$, i.e. $Y_{i,j} = R_{i,j}$, where $R_{i,j}$ represents the normalized binding affinity between a drug and a target. It is worthy to point out that, drug and target pairs without known interactions are usually discarded.

In our settings, we combine the classification and regression tasks. Hence, we can construct the supervision signals as follows.

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

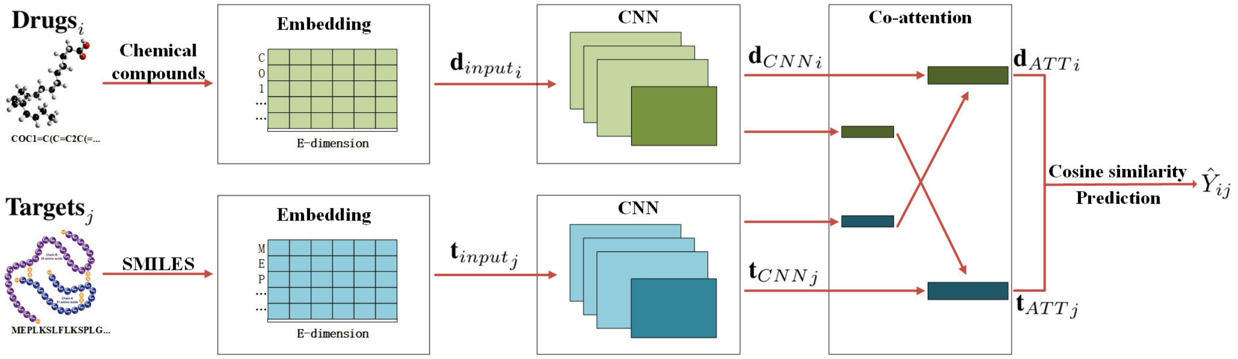


Fig. 1. Network architecture of model Multi-DTI

We can see that, the binding affinity values $\mathbf{R}$ are reserved in Y, because they indicate the strength of interaction of a drug on a target. We retain the normalized strength, i.e. each R_{ij} is divided by the maximal value in $\mathbf{R}$ for computational convenience. Meanwhile, we mark a zero if the affinity is unknown, to incorporate more information.

As shown in figure 1, the embeddings $\mathbf{d}_i, \mathbf{t}_j$ flow through an embedding layer, a feature representation component and a prediction layer. The goal of Multi-DTI is to generate $\hat{\mathbf{Y}}_{ij}$, given the model parameters $\mathbf{\Theta}$, i.e. $\hat{Y}_{ij} = F(\mathbf{d}_i, \mathbf{t}_j|\mathbf{\Theta})$, to approximate $\mathbf{Y}_{ij}$.

A. Embedding Layer

We first use integer/label encoding to represent categorical information in inputs. Similar with [5], for drugs we scan approximately 2M SMILES sequences that are collected from Pubchem and compile 64 labels, e.g. letters “C”, “N”, “=”, etc.. We represent each label by a unique integer, e.g. “C”:1, “=”:22, “N”:3 etc. For example, the label encoding for the SMILES sequence “CN=C=O” is given below.

$$[C\ N\ =\ C\ =\ O] = [1\ 3\ 22\ 1\ 22\ 5]$$

Then, we use an embedding function Ξ to transform the categorical sequence above to a dense E-dimensional float vector, i.e. $\Xi : V \leftarrow \mathcal{R}^E$, where V denotes the integer/label set for all labels. The embedding function is learned during the training phase. After a look up operation to obtain separate integer/label embeddings, all integer/label embeddings in a drug sequence are concatenated in rows. For the convenience of proceeding operations in CNN, we construct a matrix for each drug i. Hereafter, without ambiguity, we will omit subscript index i and j, and use $\mathbf{D}_{input} \in \mathcal{R}^{D\times E}$ to denote the input of CNN, where D is the maximal length of a drug sequence.

For protein sequences, we scan 550K protein sequences from UniProt and extract 25 labels. Similarly, protein sequences are first encoded using integer/label encodings, and concatenated in rows to construct a matrix representation $\mathbf{T}_{input} \in \mathcal{R}^{T\times E}$ as the input of CNN.

Both target and protein sequences have varying lengths. Hence, in order to create an effective representation form, we decided to choose a length limit, i.e. D for drugs and T for targets. The sequences that are longer than the maximum length are truncated, whereas shorter sequences are padded with zeros.

B. Feature Representation Component

In this component, we first apply CNN networks for $\mathbf{D}_{input}$ and $\mathbf{T}_{input}$ to encode sequential information. For each CNN block, we use three consecutive 2D-convolutional layers, followed by a max-pooling layer. The filters of each 2D-convolutional layer is doubled.The result is $\mathbf{D}_{CNN}$ and $\mathbf{T}_{CNN}$.

Next, we apply co-attention mechanism on $\mathbf{D}_{CNN}$ and $\mathbf{T}_{CNN}$ respectively, and output $\mathbf{D}_{ATT}$ and $\mathbf{T}_{ATT}$. The underlying assumption is that, some drugs influence more on certain targets, and the pattern can be captured by the attention weights.

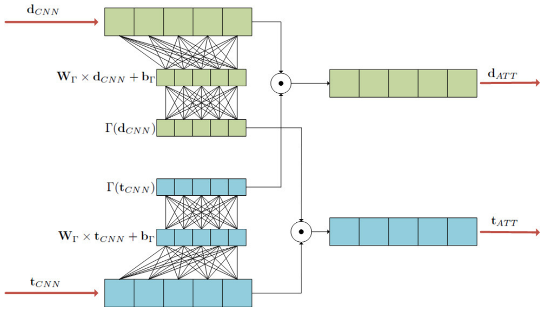

$\mathbf{d}_{CNN}$
$\mathbf{W}_\Gamma \times \mathbf{d}_{CNN} + \mathbf{b}_\Gamma$
$\Gamma(\mathbf{d}_{CNN})$
$\mathbf{d}_{ATT}$
$\Gamma(\mathbf{t}_{CNN})$
$\mathbf{W}_\Gamma \times \mathbf{t}_{CNN} + \mathbf{b}_\Gamma$
$\mathbf{t}_{CNN}$
$\mathbf{t}_{ATT}$

Fig. 2. Illustration of parallel coarse-grained co-attention.

Specifically, given a drug vector $\mathbf{d}_{CNN}$ and a target vector $\mathbf{t}_{CNN}$, the co-attention layer generates the drug output vector $\mathbf{d}_{ATT}$ by multiplying each element in the input drug vector with its attention weight to input target vector:

$$\mathbf{d}_{ATT} = \mathbf{d}_{CNN} \bigodot \Gamma(\mathbf{t}_{CNN}), \tag{2}$$

where Γ calculates the attention, i.e. the importance of $\mathbf{t}_{CNN}$ for $\mathbf{d}_{CNN}$. To derive the attention weights, a single layer feed forward network is adopted.

$$\Gamma(\mathbf{t}_{CNN}) = A(\mathbf{W}_\Gamma \times \mathbf{t}_{CNN} + \mathbf{b}_\Gamma), \tag{3}$$

where $\mathbf{W}_{\Gamma}$ and $\mathbf{b}_{\Gamma}$ are the weight matrix and the bias vector respectively for the feed forward network. A is the softmax function to normalize the attention weights over elements in each vector.

A similar structure of co-attention is also implemented on each target vectors $\mathbf{t}_{CNN}$ extracted by CNN. The co-attention mechanism is illustrated in Figure 2.

C. Prediction Layer

The feature representation component transforms the sequences of drugs and targets to low-dimensional vectors in latent feature space. In the prediction layer, we measure the similarity between two feature vectors, i.e. $\mathbf{d}_{ATTi}$ and $\mathbf{t}_{ATTj}$ by computing the cosine similarity. Intuitively, if a drug feature vector is close to a target feature vector, their binding affinity should be large. Therefore, the output $\hat{Y}_{ij}$ is defined as:

$$\hat{Y}_{ij} = cosine(\mathbf{d}_{ATTi}, \mathbf{t}_{CNNj}) = \frac{\mathbf{d}_{ATTi}^{T}\mathbf{t}_{ATTj}}{||\mathbf{d}_{ATTi}|| \cdot ||\mathbf{t}_{ATTj}||} \quad (4)$$

To combine the regression and classification tasks, we adopt the **N**ormalized **C**ross **E**ntropy(NCE) loss [28]. Given Y_{ij} the true label and $\hat{Y}_{ij}$ the predicted result,

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

Note that Equation 5 resembles in form of the conventional Binary Cross Entropy, in which $Y_{ij} \in \{0, 1\}$. Therefore, the NCE loss will encourage a classifier to assign positive labels on instances that are actually more confident, i.e. drug-target pairs that are closely binded.

The loss function is further integrated with **M**ean **S**quare **E**rror(MSE).

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

Where M, N represents the number of proteins and drugs.

Finally, the loss in Multi-DTI is defined as follows:

$$Loss = \delta \times MSE + (1 - \delta) \times NCE \quad (7)$$

where δ controls proportion of the two kinds of loss.

Discussion. Given a shared feature space for drug and target representations, how to proceed to output the prediction is still an open question. It is worthy to point out that, most previous researches [5], [6], [8] implement a multi-layer perception on a concatenation of drug and target representations. Instead, in this paper, our proposed Multi-DTI directly computes cosine similarity between drug and target representation vectors. On one hand, it is convenient for multi-task learning. On the other hand, it requires less model parameters, and thus speeds up training.

IV. Experiment

In this section, we study the following research questions. **RQ1**: How does the proposed Multi-DTI model perform, compared with state-of-the-art methods? **RQ2**: How do the parameters impact the performance of Multi-DTI? [1]

A. Experimental Setup

We use the **K**inase **I**nhibitor **B**io**A**ctivity (KIBA) dataset[2] to validate our model. The dataset consists of drug-target bioactivity strength, which is an integration of Kd, Ki and IC50 scores. The basic summary of the used datasets is shown in Table I. We used 5-fold cross-validation and report average results in the cross-validation.

TABLE I
Statistics of the KIBA dataset

	#Drugs	#Targets	# Pairs
KIBA	$2,008$	185	$92,706$

In the experiment, unless otherwise stated, we fix a length limit of maximal 100 characters for SMILES and 1000 for protein sequences. According to [5], the maximum length covers at least 80% of the proteins and 95%of the compounds. The embedding size $E = 128$. We use a batch size of 256 samples. The optimizer is Adam, convergence is declared for 200 Epochs with learning rate $1e - 5$.

B. Comparative Study

We compare our proposed method against the baseline algorithms listed as follows.

(1) PUDTI [13]: an SVM-based optimization model that is trained on negative samples extracted based on positive-unlabeled learning. Each DTI input can be described based on PaDEL-Descriptors of drugs and domains, PAACs and PSSM of target proteins. PUDTI model optimizes the Hinge loss function.

(2) DTINet [27]: a regression model that predicts drug-target interactions from a constructed heterogeneous network, which integrates diverse drug-related information. The feature is extracted by **R**andom **W**alk with **R**estart(RWR). With default drug feature P and target feature Q, the feature space mapping Z is learned by MSE loss function $min_Z(\hat{Y} - PZQ)^2$.

(3)DeepCPI [6]: an end-to-end deep neural network. The drug representation is learnt via a graph neural network (GNN) module. The protein representation is learnt via a CNN module. A neural attention mechanism is adopted to predict DTI based on concatenation of drug and protein representations. The loss function is BCE.

(4) DeepDTA [5]: another deep neural structure with two separate CNN blocks to learn the features from SMILES

[1] The code and data used in Multi-DTI are available at: https://github.com/XMUDM/Multi-DTI

[2] https://pubs.acs.org/doi/suppl/10.1021/ci400709d

TABLE II
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

strings and sequences, respectively. For prediction, a fully-connected feed-forward layer is conducted on concatenation of drug and protein representations. The loss function is MSE.

We also compare Multi-DTI with a variant to testify the impact of co-attention and multi-task learning.

(5) CNN-basic: As a variant of Multi-DTI, we conduct basic CNN modules on protein and drug sequences, and compute cosine similarity. The loss function is BCE.

(6) CNN-Multi: Another variant is to employ multi-task loss (i.e. Equation 7) on CNN representations.

We evaluate the models in terms of MSE, BCE, AUROC and AUPR.

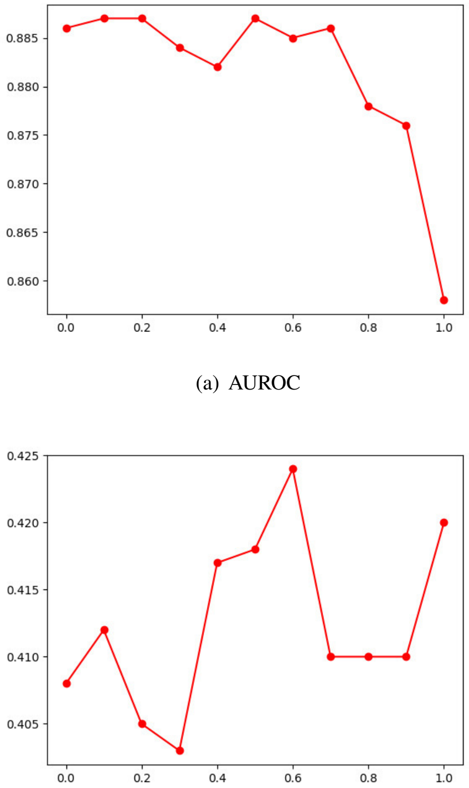
From Table II, we have the following observations. (1) Multi-DTI achieves best performance in terms of MSE, BCE and AUROC. It also achieves a comparable AUPR performance. (2) It is clear that state-of-the-art competitors can handle one task only. For example, DeepCPI adopts a classification loss, and thus, it performs best in terms of BCE, AUROC and AUPR metrics. But it performs poorly in terms of MSE. On the contrary, CNN-Multi and Multi-DTI which adopt multi-task loss, perform well in terms of all evaluation metrics. (3) Attention mechanism boosts the prediction performance. We can see that BCE of Multi-DTI is decreased by 17%, compared with CNN-Multi. (4) Multi-DTI obtains the second highest AUPR, while the best AUPR is obtained by DeepDTA. The possible reason is the trade-off between AUPR and AUROC. In the next section we will study in detail the relationship between δ, AUPR and AUROC.

C. Effect of parameters

We proceed to seek answers for **RQ2** and study the effect of parameters.

We first study the impact of proportion parameter δ, which controls the proportion of MSE in our loss function. 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 report the AUROC and AUPR performance in Figure. 3.

We can see that as δ increases, the AUROC performance generally decreases. When $\delta = 1.0$, the Multi-DTI model performs regression task only, and thus, the AUROC result reaches the lowest point. The tendency of AUPR is not monotonic. When $\delta = 0.6$, the highest AUPR is obtained.



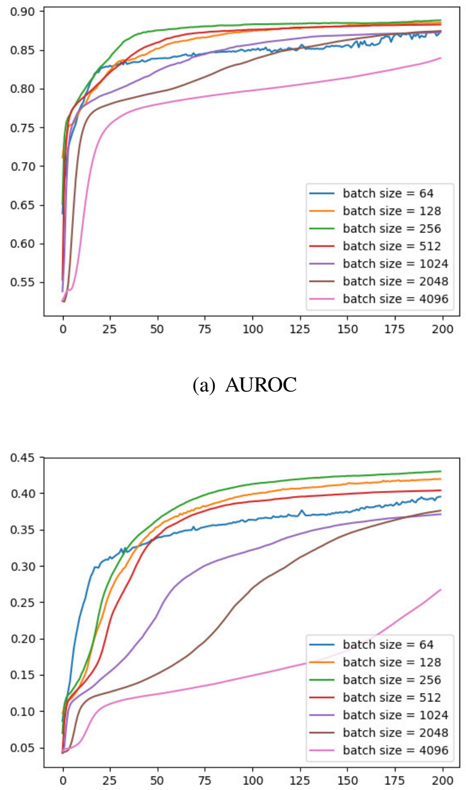
(a) AUROC

(b) AUPR

Fig. 3. Performance with different values of δ

We next study the impact of batch size. Batch size is an importance parameter in training the model. We set different batch sizes from 64 to 4096, and plot the AUROC and AUPR performance at each epoch.

As shown in Figure 4, a larger batch size leads to lower convergence. The most appropriate batch size if 256, with which the model converges fast to the highest AUROC and AUPR performance.



(a) AUROC

(b) AUPR

Fig. 4. Convergence of Multi-DTI with different batch size

V. Conclusion

We propose Multi-DTI: a novel DTI prediction model based on multi-task learning to address the challenge of bias-variance trade-off. The model learns protein and drug feature representations by adding co-attention mechanism on conventional CNN blocks. Based on the shared feature representations, the model attempts to optimize both the regression and the classification loss. We experimentally demonstrate that Multi-DTI outperforms state-of-the-art computational DTI identification methods. Our future directions include enhancing the DTI prediction performance by multi-view, multi-modality and multi-task learning.

References

[1] Palma G, Vidal M E, Raschid L. Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. The Semantic Web ISWC 2014. Springer International Publishing, 2014.

[2] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, and T. Aittokallio. Toward more realistic drug-target interaction predictions. Briefings in Bioinformatics, 2015, 16(2):325-337.

[3] H. Luo, W. Mattes, D. L. Mendrick, and H. X. Hong. Molecular docking for identification of potential targets for drug repurposing. Current topics in medicinal chemistry, 16(30):36363645, 2016.

[4] T. He, M. Heidemeyer, F. Q. Ban, A. Cherkasov, and M. Ester. Simboost:a read-across approach for predicting drugc-target binding affinities using gradient boosting machines. Journal of Cheminformatics, 9(24):114, 2017.

[5] H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: deep drug-target binding affinity prediction. Bioinformatics, 34(17):i821i829, 2018.

[6] M. Tsubaki, K. Tomii, and J. Sese. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics, 35(2):309318, 2018.

[7] Ni S , Lin C , Zeng X , et al. Drug Target Interaction Prediction with Non-random Missing Labels[C] 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018.

[8] Nguyen Thin, Le Hang, Venkatesh Svetha. et al. GraphDTA: prediction of drugtarget binding affinity using graph convolutional networks. 2019, 10.1101/684662.

[9] Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology, 2007, 25(2):197-206.*

[10] Cheng, A. C. et al. Structure-based maximal a nity model predicts small-molecule druggability. Nature Biotechnology, 2007, 25(1):71-75.

[11] Chen, X. et al. Drug-target interaction prediction: databases, web servers and computational models. Brief. Bioinform. , 2016, 17(4):696.

[12] Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. In Briefings in bioinformatics*, 15(5), 734-747.*

[13] Peng L, Zhu W, Liao B, et al. Screening drug-target interactions with positive-unlabeled learning. In Scientific Reports*, 2017, 7(1): 8087.*

[14] H. J. Li, K. Leung, M. Wong, and P. Ballester. Low-quality structural and interaction data improves binding affinity prediction via random forest. Molecules, 2015, 20(6):10947-10962.

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

[16] Guney E , Menche J , Vidal M , et al. Network-based in silico drug efficacy screening. Nature Communications, 2016, 7:10331.

[17] P. Zhang, L. Tao, X. Zeng, C. Qin, S. Y. Chen, F. Zhu, Z. R. Li, Y. Y. Jiang, W. P. Chen, and Y. Z. Chen. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *Briefings in Bioinformatics, 18(6):10571070, 2017.*

[18] M. Wen, Z. M. Zhang, S. Y. Niu, H. Z. Sha, R. H. Yang, Y. H. Yun, and H. M. Lu. Deep-learning-based drug-target interaction prediction. Journal of proteome research, 16(4):14011409, 2017.

[19] Zong N, Kim H, Ngo V, et al. Deep Mining Heterogeneous Networks of Biomedical Linked Data to Predict Novel Drug-Target Associations. In *Bioinformatics*, 2017, 33(15).

[20] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C] *International Conference on Neural Information Processing Systems. 2012.*

[21] Ragoza M , Hochuli J , Idrobo E , et al. ProteinLigand Scoring with Convolutional Neural Networks. Journal of Chemical Information and Modeling, 2017, 57(4):942-957.

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

[23] Ying R , He R , Chen K , et al. Graph Convolutional Neural Networks for Web-Scale Recommender Systems In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 974-983.

[24] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need *arXiv preprint arXiv:1706.03762, 2017.*

[25] Ma, D., Li, S., Zhang, X., Wang, H. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 4068-4074.

[26] Fan, F., Feng, Y., Zhao, D. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3433-3442. Association for Computational Linguistics.*

[27] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. In Nature Communications*, 2017, 8(1).*

[28] Xue H J , Dai X , Zhang J , et al. Deep Matrix Factorization Models for Recommender Systems. Twenty-Sixth International Joint Conference on Artificial Intelligence. AAAI Press, 2017.

[29] Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines[J]. *Bioinformatics, 2013, 29(13):126-134.*