CSCL-DTI: predicting drug-target interaction through cross-view and self-supervised contrastive learning

Xuan Lin[†,‡], Xi Zhang[†], Zu-Guo Yu[‡], Yahui Long[#,*], Xiangxiang Zeng[⊤] and Philip S. Yu[§]
[†] College of Computer Science, Xiangtan University, China
[‡] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, China
[#] Singapore Immunology Network (SIgN), Agency for Science, Technology and Research(A*STAR), Singapore
[⊤] College of Information Science and Engineering, Hunan University, China
[§] Department of Computer Science, University of Illinois Chicago, USA
[*]Corresponding authors
long_yahui@immunol.a-star.edu.sg

***Abstract*—Accurately predicting drug-target interactions (DTI) is a critical step in drug discovery. Existing methods of DTI prediction primarily employ Simplified Molecular-Input Line-Entry System (SMILES) sequences or molecular graphs to learn drug representations. However, the features learned by such single-view approach is prone to incomplete. While some multi-view methods that consider the views of both SMILES sequences and molecular graphs have been developed, these methods often fall in short in capturing potential interactions between views. In this work, we propose a novel dual contrastive learning framework CSCL-DTI for DTI prediction. First, we design a contrastive-enhanced cross-view representation learning (CVRL) to learn representations for drugs. In this module, Transformer-based and graph convolutional network (GCN)-based encoders are separately adopted to learn view-specific representations, followed by contrastive learning to enrich the representations by accounting for the potential interplay between local chemical context and topological structure. Second, we combine Transformer with self-supervised contrastive learning (SSCL) to learn representations for targets by modelling protein amino acids sequences. The scheme allows to effectively preserve the intrinsic characteristics of the sequences. Finally, we introduce a bilinear attention network to obtain an integrated representation by adaptively incorporating drug and target representations. Benchmarking experiments on three datasets demonstrated that CSCL-DTI[1] outperforms seven state-of-the-art methods.**

***Index Terms*—Drug discovery, Drug-target interactions, Graph neural network, Transformer, Contrastive learning.**

I. Introduction

Predicting drug-target interactions (DTI) is a crucial step in drug discovery and repurposing [1], [2]. In order to develop a new drug, it is essential for discovering which proteins the drug targets. Traditional methods rely on high-throughout screening experiments to examine a drug's affinity toward its targets. However, these methods face significant challenges because of the large-scale search space of potential drug and protein candidates, resulting in high cost and long period. Consequently, the development of computational methods for predicting drug-target interactions is urgently needed.

[1]https://github.com/xiiiz/CSCL-DTI

Over the past decade, some computational methods have been proposed for DTI prediction [3]. We can divide these methods into two main groups, namely *molecular docking simulation (MDS)-based methods* and *machine learning (ML)-based methods*. MDS-based methods focus on simulating the drug binding mechanism to target proteins [4]. This is accomplished by predicting receptor-ligand complex structures. However, these methods face challenges when dealing with numerous proteins that lack 3D structures, as these structures are essential for the simulation process. Besides, docking simulations tend to be time-consuming and thus, inefficient for large-scale applications.

To address these challenges, machine learning-based methods have been proposed and widely applied for DTI prediction [5]. For instance, a deep learning method called DeepDTA was proposed by Öztürk et al. [6] for predicting the binding affinities of drug-target interactions using sequencing information of both drugs and targets. Subsequently, Lee et al. [7] present a convolutional neural network (CNN) [8] based model called DeepConv-DTI to predict DTI using protein sequences and drug fingerprints. Inspired by the successful application of Transformer [9] model in Natural Language Processing (NLP), Chen et al. [10] released TransformerCPI for DTI prediction, which applies the Transformer architecture to represent drug Simplified Molecular-Input Line-Entry System (SMILES) [11] strings and protein amino acid sequences. To take into account the sub-structural nature of DTI, Huang et al. [12] proposed another Transformer-based method named MolTrans for DTI prediction. MolTrans improves the prediction performance by enhancing the extraction of semantic relations among subsequence structures of drugs and targets. However, these models primarily focus on sequence information of drugs and targets, overlooking molecular structural topology, which is essential for enhancing prediction accuracy.

To alleviate this limitation, recently, graph representation methods for molecular graph-based DTI prediction have been developed [13]. Nguyen et al. [14] introduced GraphDTA, a deep learning model that represents drugs SMILES as graphs.

Jahromi et al. [15] presented AttentionsiteDTI, a graph representation learning method for DTI prediction. These methods take into account topological structures, leading to improved performance. However, a common limitation is that most depend solely on single-view information, either sequence or molecular graph, without incorporating insights from both of these crucial perspectives.

For this purpose, multi-view methods have been developed for DTI prediction that simultaneously considers chemical context and topological structures [16]. For example, DeepGS [17], a deep learning model proposed for drug-target binding affinity prediction from our lab, attempted to combine amino acids sequence with molecular graph to improve the prediction. Similarly, Cheng et al. [18] introduced IIFDTI, a deep learning-based method for DTI prediction. It incorporates independent features of drug molecular graphs and target sequences. Despite these advancements, existing multi-view methods face certain limitations. First, although they utilize multi-view information, these methods often treat such information independently, overlooking potential interrelations between cross-views. Second, few of these methods simultaneously consider multi-view information of the same node type (e.g., drugs or targets).

To overcome these limitations, we propose a novel contrastive learning framework for predicting DTI, named CSCL-DTI. CSCL-DTI is designed to comprehensively utilize the sequence and topological structures of drugs, as well as their interrelations. With SMILES sequence of a given drug, we apply the Frequent Common Subsequence (FCS) algorithm [12] and the RDKit package [19] to derive its subsequence and molecular graph. Then we introduce a dual-encoder framework, a Transformer encoder for the subsequence and a Graph Convolutional Network (GCN) [20] encoder for the molecular graph. This framework enables us to obtain distinct yet complementary sequence and graph representation. To capture inherent correlation between these two views, cross-view contrastive learning strategy is combined with the duel-encoder framework, leading to refined drug representation. Finally, a bilinear attention aggregation module is utilized to adaptively incorporate view-specific representation to obtain final drug representation. For target representation, we employ a self-supervised contrastive learning scheme. This scheme can effectively capitalizes on the intrinsic properties of the amino acid sequence. Comprehensive experiments on three datasets demonstrated that our CSCL-DTI consistently outperformed state-of-the-art methods.

Overall, our main contributions are summarized as follows.

- We proposed a novel contrastive learning framework for DTI prediction called CSCL-DTI, which fully exploits sequence and topological structures, along with their interrelation, to enhance the prediction.
- To capture potential interplay between different views of drugs and intrinsic characteristics in the target sequence, we introduced cross-view and self-supervised contrastive learning strategies to learn representations for drugs and targets, respectively, resulting in refined representations.
- We incorporated a bilinear attention mechanism to effectively learn drug representations by adaptively integrating features from different views.
- Comprehensive experiments demonstrated the proposed CSCL-DTI model outperformed seven state-of-the-art methods. Ablation study validated the contribution of each component to the overall performance of the model.

II. METHODOLOGY

In this section, we introduce the CSCL-DTI model for DTI prediction. Fig. 1 illustrates the entire architecture of CSCL-DTI, which is comprised of three main components: a contrastive-enhanced cross-view representation learning (CVRL) module for drug representation, a self-supervised contrastive learning (SSCL) module for target representation, and a classifier. First, we design a contrastive-enhanced cross-view representation learning module to learn drug representations, specifically leveraging drug relevant information from dual views, i.e., SMILES sequence and molecular graph. Second, we use a SSCL module to effectively learn target representation by fully exploiting the intrinsic properties of amino acid sequence. Third, concatenating the learnt drug and target representations, they are given into the classifier to predict DTI. Next, we elaborate on each component of the model in detail.

A. Contrastive-enhanced cross-view representation learning for drug representation

For a given drug, data from different perspectives offers unique and complementary features, which are crucial for the improvement of drug representation learning. Inspired by the great success of contrastive learning in the computer vision domains such as SimCLR [21] and MoCo [22], we design a CVRL module for drug representation learning. This module integrates SMILES sequence with molecular graph for the representation learning. We utilize both Transformer and GCN encoders to acquire sequence and graph representations from these respective views. To effectively capture relationships between these two views, we introduce a cross-view contrastive learning strategy to enhance representation learning. Eventually, the sequence and graph representations are adaptively aggregated using a bilinear attention method, deriving the final drug representation.

1) Sequence representation: To acquire sequence representations from SMILES sequence data, we utilize a Transformer-based encoder. Prior to encoding, we preprocess the SMILES sequences using the FCS algorithm [12], renowned for its ability to capture essential biomedical semantics. Specifically, FCS decomposes SMILES sequence into atomic symbols or shorter sub-sequences, and then maps the decomposed sub-sequences via a preset dictionary into corresponding embedding vectors. For a given drug i, FCS yields a content embedding $E^d_{cont_i}$ and a positional embedding $E^d_{pos_i}$. The initial embedding E^d_i of drug i is formulated by summing these content and positional embeddings, effectively integrating both the structural and sequential context of the drug representation.

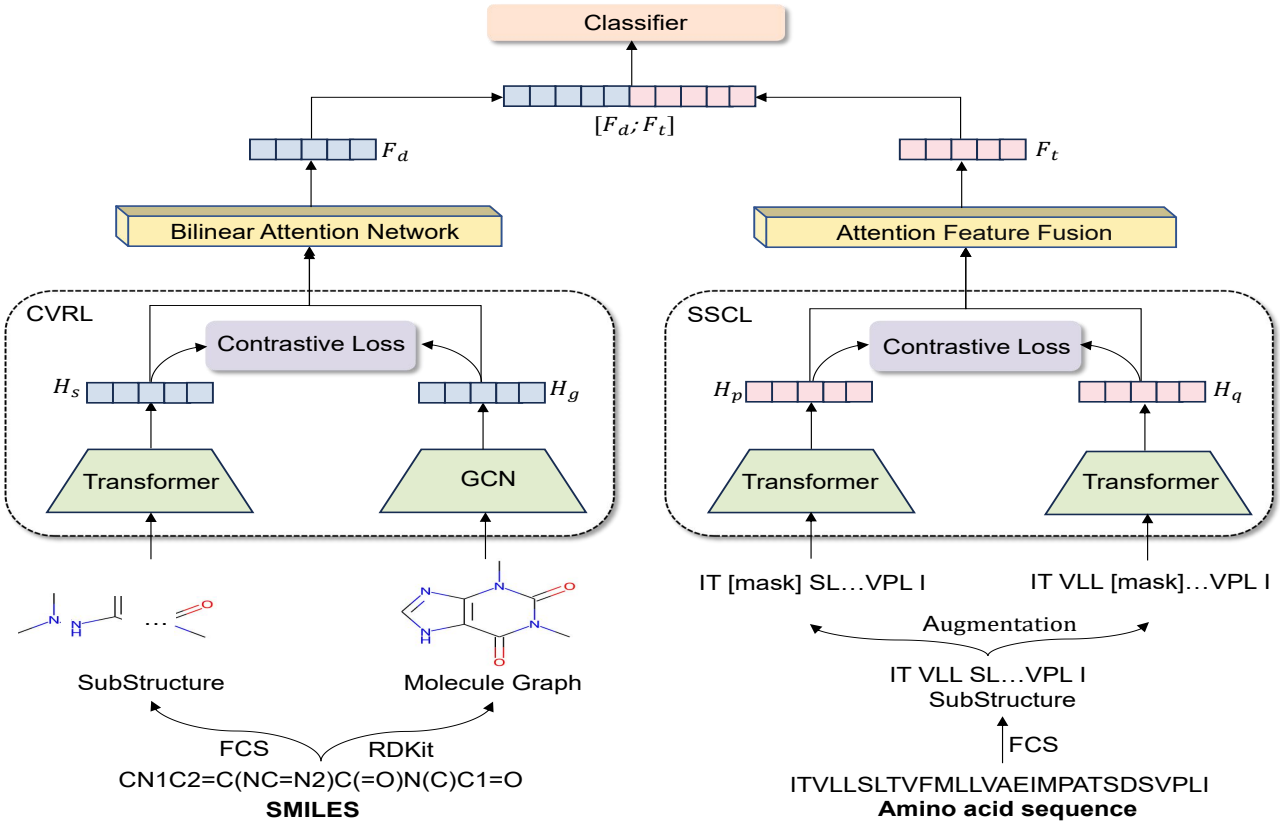


Fig. 1. CSCL-DTI workflow for drug-target interaction prediction.

$$E_i^d = E_{cont_i}^d + E_{pos_i}^d. \tag{1}$$

The initial embedding $E^d \in \mathbb{R}^{N_d \times d_1}$, N_d and d_1 denote the number of drugs and the dimension of initial embedding respectively. However, this representation lacks the preservation of chemical relationships or contextual information among these sub-structures. To address this, we utilize a Transformer model to map the initial embedding into a latent space. The Transformer model has demonstrated powerful ability in learning contextual information [9]. By taking the initial embedding E_i^d as its input, the Transformer model can be formulated as follows.

$$(Q, K, V) = E_i^d * (W_Q, W_K, W_V), \tag{2}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_2}}\right)V, \tag{3}$$

where $Q \in \mathbb{R}^{N_d \times d_2}$, $K \in \mathbb{R}^{N_d \times d_2}$, and $V \in \mathbb{R}^{N_d \times d_2}$ represent the query, key, and value vectors respectively. d_2 denotes the dimension of feature vectors. $*$ represents the matrix dot product operation. $W_Q \in \mathbb{R}^{d_1 \times d_2}$, $W_K \in \mathbb{R}^{d_1 \times d_2}$, and $W_V \in \mathbb{R}^{d_1 \times d_2}$ are trainable projection weights initialized by a neural network. With initial embeddings as inputs, the Transformer model outputs sequence representation $H_s \in \mathbb{R}^{N_d \times d_2}$ for drugs. The Transformer model enables the modeling of contextual dependencies among the sub-structures, thereby enriching the drug representation.

2) Graph representation: Each SMILES sequence is transformed into a molecular graph using RDKit [19] tool. Assuming that the molecular graph is denoted as $G(\nu, \varepsilon)$, where ν is the set of atoms in the molecular graph, and ε is the set of chemical bonds connecting these atoms. We denote $A \in \mathbb{R}^{N_a \times N_a}$ as the adjacent matrix of this graph, where the entity A_{ij} is equal to 1 if there is a bond between atoms i and j, 0 otherwise. N_a is the number of atoms. The initial features of atoms are defined by physicochemical properties and are denoted as $X \in \mathbb{R}^{N_a \times d_3}$ with d_3 representing the dimension of features (The details of atom features can be found in the **Appendix I**[1]). The rationale behind GCN lies in learning node representation by aggregating information from neighbor nodes. Formally, the propagation mechanism of GCN can be defined as follows.

$$Z^{(l+1)} = \sigma(\tilde{A} Z^{(l)} W^{(l)}), \tag{4}$$

where $\widetilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ denotes the normalized symmetrical adjacent matrix and D is a diagonal matrix with diagonal elements being $D_{ii} = \sum_{j=1}^{N_a} A_{ij}$. $Z^{(l)}$ represents the representation at the l-th layer, and we set Z^0 to the initial atom features X. σ is a non-linear activation function such as $ReLU$. $W^{(l)} \in \mathbb{R}^{N_a \times d_2}$ is a trainable parameter matrix for the convolution transformation of the current layer. d_2

denotes the dimension of representation. It should be noted that the dimension of graph representation is set to the same as the sequence representation for benefiting downstream cross-view contrastive learning. We set $Z = Z^{(L)}$ as the final representation of the molecular graph of drug i and L is set as 2 in our experiment.

After obtaining the representation of molecular graph, the graph representation $H_g \in \mathbb{R}^{N_d \times d_2}$ is obtained through an averaging pooling function. Specifically, it can be formulated as follows.

$$H_g = -\frac{1}{N_a} \sum_{i=1}^{N_a} Z_i \tag{5}$$

3) Cross-view contrastive learning for representation refinement: Motivated by the intuition that the sequence and graph representations of the same drug are intrinsically more similar to each other than those of different drugs, we further incorporate a novel cross-view contrastive learning scheme to refine drug representation. Unlike traditional contrastive learning typically relying on data augmentation [23], this scheme is augmentation-free and focuses on aligning representations of the same drug while distinguishing representations between different drugs.

In our model, we employ mini-batches for model training. For the contrastive learning, the definition of positive and negative samples is required. Departing from previous works, we propose a new positive selection strategy. Specifically, for a given drug i, we treat its sequence representation h_i^s and graph representations h_i^g as positive sample. In contrast, the sequence and graph representations h_j^s and h_j^g of any other drug j ($j \in N_{-i}$) within the same batch are considered as negative samples.

The objective function of the cross-view contrastive learning loss for drugs is defined as follows.

$$\mathcal{L}_{CL}^{d} = \mathcal{L}_{CL}^{s} + \mathcal{L}_{CL}^{g}, \tag{6}$$

$$\mathcal{L}_{CL}^{s} = -\frac{1}{2} \sum_{i=1}^{N_d} \log \frac{\exp\left(\text{sim}(h_i^s, h_i^g)/\tau\right)}{\sum_{j=0}^{n} \left(\exp\left(\text{sim}(h_i^s, h_j^g)/\tau\right) + \exp\left(\text{sim}(h_i^s, h_j^s)/\tau\right)\right)}, \tag{7}$$

$$\mathcal{L}_{CL}^{g} = -\frac{1}{2} \sum_{i=1}^{N_d} \log \frac{\exp\left(\text{sim}(h_i^g, h_i^s)/\tau\right)}{\sum_{j=0}^{n} \left(\exp\left(\text{sim}(h_i^g, h_j^s)/\tau\right) + \exp\left(\text{sim}(h_i^g, h_j^g)/\tau\right)\right)}, \tag{8}$$

where $j \neq i$ and n is batch size. τ denotes temperature parameter and $\text{sim}(u, v)$ denotes cosine similarity.

4) Bilinear attention for representation aggregation: Following previous study [24], we utilize a Bilinear Attention Network (BAN) to effectively integrate the sequence and graph representations H_s and H_g, taking into account the intrinsic relationship between the SMILES sequence and molecule graph. The BAN is comprised of two main layers: a bilinear correlation map layer and a bilinear pooling layer. The correlation map layer is specifically designed to capture pairwise attention weights, while the pooling layer is applied over the correlation map to extract a unified representation.

$$I = \left((1 \cdot q^T) \circ \sigma((H_s)^T U)\right) \cdot \sigma(V^T H_g), \tag{9}$$

where $q \in \mathbb{R}^{d_4}$ is a learnable weight vector. $U \in \mathbb{R}^{N_d \times d_4}$ and $V \in \mathbb{R}^{N_d \times d_4}$ are learnable weight matrices for the drug sequence and graph representations, respectively. $q \in \mathbb{R}^{d_4}$ is a learnable weight vector, and $\mathbf{1} \in \mathbb{R}^{d_2}$ is a fixed vector of all-ones. $\circ$ represents the Hadamard product (element-wise). $\cdot$ represents matrix multiplication operation. The elements in I indicate the correlation relationships between the sequence and graph representations.

Over the correlation map I, we add a bilinear pooling layer to generate the joint representation $f_d \in \mathbb{R}^{d_4}$. In particular, the following formula is used to get the k-th element of $f_d \in \mathbb{R}^{d_4}$.

$$f_d^k = \sigma\left((H_s)^T U\right)_k^T \cdot I \cdot \sigma\left((H_g)^T V\right)_k, \tag{10}$$

where U and V are learned weight matrices shared with the previous correlation map layer to reduce the amount of parameters and alleviate overfitting. We additionally use a sum pooling on the joint representation vector to reduce dimension and produce a compact feature map.

$$F_d = \textit{SumPool}(f_d, s), \tag{11}$$

where $\textit{SumPool}(u, v)$ is a non-overlapping, one-dimensional sum pooling process with a stride of s. $F_d \in \mathbb{R}^{N \times d_5}$ is the joint representation of drugs with d_5 denoting the dimension of representation.

B. Self-supervised contrastive learning for target representation

1) Target representation: To learn target representation, we employ a Transformer-based encoder, augmented with SSCL. In a process analogous to that used for drugs, we preprocess the amino acid sequences of targets using the FCS algorithm before encoding. The FCS algorithm decomposes amino acid sequences into subsequence structures, generating a content embedding $E_{cont_i}^t$ and a positional embedding $E_{pos_i}^t$ for each target i. The initial embedding $E_i^t \in \mathbb{R}^{N_t \times d_1}$ (N_t is the number of targets) of the target i is then derived by summing these content and positional embedding.

$$E_i^t = E_{cont_i}^t + E_{pos_i}^t. \tag{12}$$

As previously mentioned, the initial embedding primarily captures information from independent subsequences but is insufficient in preserving the chemical relationships between them. To address this limitation, we employ the Transformer model, which is defined in Section II-A1, to map the initial embedding E^t into a latent space, deriving target representation $H^t \in \mathbb{R}^{N_t \times d_2}$.

To capture intrinsic characteristics in the amino acid sequences, we further introduce SSCL mechanism to refine target representation. Specifically, we implement data augmentation on the decomposed subsequence structures by randomly

masking subwords with a given ratio r (by default, set to 0.1), resulting in two augmented views p and q for each target. For a given target i, its two augmented views form positive pair while both of them form negative pairs with the augmented views from other targets in the same batch. The purpose of the contrastive learning is to pull intra-view representations together while pushing inter-view representations away. The following formula represents the goal function of contrastive learning for targets.

$$\mathcal{L}_{CL}^{t} = \mathcal{L}_{CL}^{p} + \mathcal{L}_{CL}^{q}, \tag{13}$$

$$\mathcal{L}_{CL}^{p} = -\frac{1}{2}\sum_{i=1}^{N_t} \log \frac{\exp\left(\text{sim}(h_i^p, h_i^q)/\tau\right)}{\sum_{j=0}^{n}\left(\exp\left(\text{sim}(h_i^p, h_j^q)/\tau\right) + \exp\left(\text{sim}(h_i^p, h_j^p)/\tau\right)\right)}, \tag{14}$$

$$\mathcal{L}_{CL}^{q} = -\frac{1}{2}\sum_{i=1}^{N_t} \log \frac{\exp\left(\text{sim}(h_i^q, h_i^p)/\tau\right)}{\sum_{j=0}^{n}\left(\exp\left(\text{sim}(h_i^q, h_j^p)/\tau\right) + \exp\left(\text{sim}(h_i^q, h_j^q)/\tau\right)\right)} \tag{15}$$

where $i, j \in (1, n)$ and $j \neq i$, τ denotes the temperature parameter, n is the batch size, and $\text{sim}(u, v)$ denotes cosine similarity.

2) Attentive representation aggregation: To acquire the final target representation, the Attentional Feature Fusion (AFF) mechanism is designed to combine view-specific representations. This mechanism adopts a weighted averaging strategy that assigns weights to different representation vectors based on their importance. We first perform initial integration on the view-specific representations, and then evaluate the importance of each view using the sigmoid activation function, resulting in a learned weight vector in the range of 0 to 1. For two feature maps H_p and H_q, AFF can be represented as follows.

$$F_t = M(H_p \oplus H_q) \odot H_p + (1 - M(H_p \oplus H_q)) \odot H_q, \tag{16}$$

where $F_t \in \mathbb{R}^{N_t \times d_5}$ is the fused feature, the initial feature integration is denoted by $\oplus$. To initiate integration, we opt for an element-wise summation as the primary step. The symbol $\odot$ represents the element-wise product operation. M is the weight matrix calculated by a Sigmoid function, where $M \in (0, 1)$. Note that the fusion weights $M(H_p \oplus H_q)$ are made up of real numbers ranging from 0 to 1. Similarly, $1 - M(H_p \oplus H_q)$ are also real numbers between 0 and 1, enabling the network to conduct soft selection or weighted averaging between H_p and H_q.

C. Drug-Target Interaction Prediction

The DTI prediction task is approached in this study as a binary classification problem. Utilizing the acquired representations, the objective is to predict whether a given drug-target pair is interactive. With the representations of drugs and targets F_d and F_t, the classification problem is formulated as follows.

$$\hat{X} = \delta(W_{\text{out}}(F_d; F_t) + b_{\text{out}}), \tag{17}$$

where δ denotes non-linear activation function (i.e., Sigmoid). W_{out} and b_{out} are a learnable weight matrix and a learned bias vector respectively, and $\hat{X}$ is predicted label. The cross-entropy loss function in the classifier is defined as follows.

$$\mathcal{L}_{CLS} = -\frac{1}{N}\sum_{i=1}^{N}\left(x_i \cdot \log\left(\hat{x}_i\right) + (1 - x_i) \cdot \log\left(1 - \hat{x}_i\right)\right) + \frac{\lambda}{2}\|\theta\|_2^2, \tag{18}$$

where x_i and $\hat{x}_i$ denote known and predicted labels respectively. N is the number of training samples. θ denotes the set of model parameters, and λ represents the coefficient for L_2 regularization. The model is jointly trained by the classification loss and contrastive loss.

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{CLS} + \beta \cdot \mathcal{L}_{CL}^{d} + \gamma \cdot \mathcal{L}_{CL}^{t}, \tag{19}$$

where α, β, and γ are hyper-parameters controlling the influences of different losses on model training.

III. Results and Discussion

This section begins with an introduction to the experimental setups including datasets and baseline methods. Next, we compare our proposed CSCL-DTI model with a variety of baselines. Additionally, an ablation study is conducted to demonstrate the performance of the model. Finally, we test several important parameters.

A. Dataset

To evaluate the performance of our proposed CSCL-DTI, We adopt three widely-used benchmark datasets, including GPCR [10], Human [25], and DrugBank [18] datasets. The details of these three datasets are summarized in Table I. The detailed dataset partition on three datasets are all arranged in earlier research [18].

Table. I. The detailed dataset division of GPCR, Human, and DrugBank.

Datasets	Drugs	Targets	Interactions	Positive	Negative
GPCR	5,359	356	15,343	7,989	7,354
Human	1,052	852	6,738	3,369	3,369
DrugBank	2,615	2,932	17,248	7,261	9,987

B. Baseline methods

To demonstrate the efficacy of our model, we compare CSCL-DTI with seven state-of-the-art methods.

- **Vina** [26] adopts the AutoDock Vina tool for molecular docking, selecting the top 1 binding affinity based on its scoring function, and the affinity threshold is set according to a 1:1 ratio of positive and negative samples to convert affinity into the probability of drug-target interaction.
- **DeepDTA** [6] adopts two separate CNN modules to predict the binding affinity between a drug-target pair by extracting features from SMILES and amino acid sequences, respectively.

- **DeepConv-DTI** [7] utilizes the drug molecular fingerprints and protein amino acid sequences as input. It also employs CNN to capture local residual patterns.
- **MolTrans** [12] employs an augmented Transformer to extract the contextual features from both drug SMILES sequences and amino acid sequences of protein.
- **GraphDTA** [14] adopts GCN, Graph Attention Networks (GAT) [27], and Graph Isomorphism Networks (GIN) [28] to extract molecular graph features of drugs, while utilizing CNNs to extract features of amino acid sequences for predicting the binding affinity.
- **TransformerCPI** [10] extracts the useful features from drug SMILES and protein amino acid sequences by employing an attention-based Transformer for predicting compound-protein interaction.
- **IIFDTI** [18] fuses the interactive and independent features between drug-target pairs to predict DTI. On one hand, for drug-target pairs, it extracts interactive features from substructures using a bidirectional encoder-decoder extractor. On the other hand, it separately models the independent features of drugs and proteins by employing GAT and CNN, respectively.

Table. II. The results on all the dataset: AUC, AUPR, recall of the baselines and CSCL-DTI.

Datasets	Methods	AUC	AUPR	Recall
GPCR	Vina	0.519±0.003	0.635±0.005	0.532±0.008
	DeepDTA	0.776±0.006	0.762±0.015	0.712±0.015
	DeepConv-DTI	0.752±0.011	0.685±0.010	0.713±0.021
	MolTrans	0.807±0.004	0.788±0.009	0.762±0.014
	GraphDTA	0.840±0.004	0.836±0.006	0.790±0.006
	TransformerCPI	0.842±0.007	0.837±0.010	0.796±0.015
	IIFDTI	0.845±0.008	0.842±0.007	0.783±0.017
	CSCL-DTI	**0.860±0.008**	**0.862±0.009**	**0.799±0.018**
Human	Vina	0.569±0.006	0.703±0.004	0.570±0.007
	DeepDTA	0.972±0.001	0.973±0.002	0.935±0.017
	DeepConv-DTI	0.967±0.002	0.964±0.004	0.907±0.023
	MolTrans	0.974±0.002	0.976±0.003	0.933±0.022
	GraphDTA	0.972±0.005	0.973±0.005	0.946±0.006
	TransformerCPI	0.970±0.006	0.974±0.005	0.937±0.011
	IIFDTI	0.984±0.003	0.985±0.003	0.947±0.017
	CSCL-DTI	**0.987±0.001**	**0.988±0.001**	**0.951±0.005**
DrugBank	Vina	0.503±0.005	0.436±0.007	0.513±0.009
	DeepDTA	0.784±0.004	0.519±0.007	0.635±0.010
	DeepConv-DTI	0.782±0.005	0.472±0.005	0.626±0.016
	MolTrans	0.501±0.010	0.203±0.006	0.417±0.015
	GraphDTA	0.786±0.006	0.517±0.008	0.638±0.008
	TransformerCPI	0.782±0.005	0.500±0.015	0.660±0.007
	IIFDTI	0.797±0.004	0.527±0.009	0.679±0.008
	CSCL-DTI	**0.808±0.002**	**0.557±0.007**	**0.689±0.010**

C. Comparation to other methods

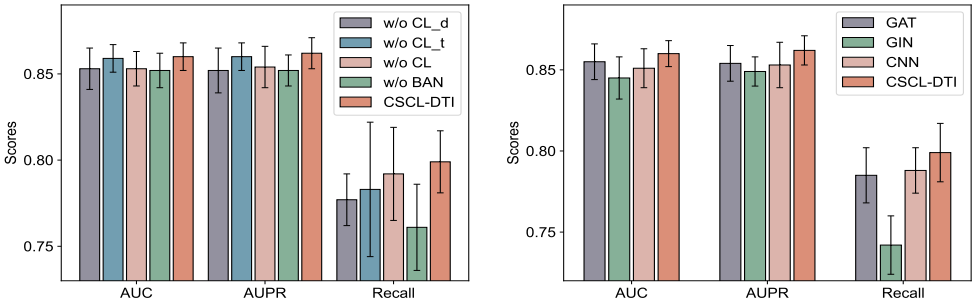
Table II shows the comparison results on GPCR, Human, and DrugBank datasets, respectively. Experimental results of some baseline methods are obtained from the previous work [18]. The metrics with the best results are bolded, while the second-best results are underlined. It is observed that CSCL-DTI consistently achieves better performance compared to baseline methods on three datasets. More specifically, for GPCR dataset, CSCL-DTI outperforms the second-best method by 1.5% and 2%, respectively, in terms of AUC and AUPR. For small public dataset (i.e., Human), CSCL-DTI is slightly better than the second best method, while it achieves at least 1.3% on AUC, 1.2% on AUPR, and 1.4% on Recall higher performance than other methods, respectively. On the DrugBank dataset, the AUC, AUPR, and Recall of CSCL-DTI are about 1.1%, 3%, and 1% higher than those of the second best method (i.e., IIFDTI), and over 2.4%, 3.8%, and 2.9% than those of the other methods, respectively.

The exceptional performance of CSCL-DTI can be attributed to several reasons as follows. Firstly, compared to traditional molecular docking and simulation methods (e.g., Vina), which utilize computational approaches to simulate and to predict the binding between drugs and proteins, machine learning methods achieve better performance because they can automatically discover more effective features through feature engineering. Secondly, single-view methods (e.g., DeepDTA, GraphDTA and TransformerCPI), which solely leverages drug SMILES sequence or molecule graph for DTI prediction, the multi-view methods, CSCL-DTI and IIFDTI, generally exhibit better performance across all three datasets. This is because multi-view methods (i.e., CSCL-DTI and IIFDTI) can explore both sequence and graph features, while sequence-based methods do not consider the topological features of drug molecular graph and graph-based methods do not consider the context features of drug SMILES. Finally, compared to multi-view method (i.e., IIFDTI), CSCL-DTI also achieved superior results (e.g., 2% and 3% improvement of AUPR on GPCR and DrugBank dataset, respectively). This is because the multi-view embeddings in CSCL-DTI can better extract intricate relationships through hierarchical contrastive learning strategies. Furthermore, we construct imbalanced datasets to assess the robustness of CSCL-DTI, The details of the comparative experiments on the imbalanced dataset can be found in the **Appendix II**[1].

D. Ablation study

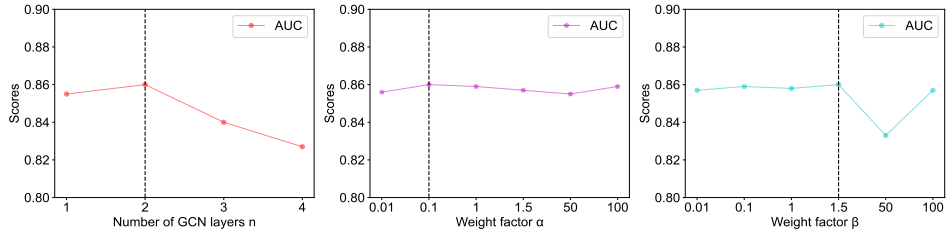
We conducted ablation study on the GPCR dataset from two perspectives: component ablation and model design, to further check the effectiveness of different configurations of CSCL-DTI. The result of each experiment were determined by repeating the experiment 10 times using different seeds.

1) component ablation: We remove relevant network structures, namely contrastive learning (CL) and BAN, to confirm their contribution in performance improvement. On one hand, we design tailored contrastive learning strategies for drug and target representations to extract multi-view features. To study the effectiveness of this central idea, we implement three variants of our model. Specifically, the models "*w/o CL_d*" and "*w/o CL_t*" indicate the removal of corresponding CL strategies for drug and target representations in CSCL-DTI, respectively. And "*w/o CL*" denotes the removal of CL strategies for both drug and target representations in CSCL-DTI. On the other hand, different from previous multi-view based methods that simply concantenate the drug and target features, the proposed CSCL-DTI adopt BAN to align learned representations across different views. The model "*w/o BAN*"



(a) (b)

Fig. 2. Comparison between CSCL-DTI and its variants on GPCR dataset.



(a) (b) (c)

Fig. 3. Parameter sensitivity analysis for CSCL-DTI on GPCR dataset.

denotes the elimination of a BAN module from CSCL-DTI. From Fig. 2a, we can draw the following conclusions: (i) either only using CVRL for drug representation or SSCL for target representation, the Recall result decrease at least by about 1.6% comparing to the full model. This result shows that hierarchical contrastive learning strategies are effective for feature representations. Moreover, the CVRL contributes more than SSCL while multi-view CL strategy can bring relatively more informative features than self-supervised CL strategy. (ii) Comparing the results of "*w/o BAN*" and CSCL-DTI, the BAN module improves the results on AUPR and Recall by nearly 1% and 3.8%, respectively. In summary, our findings highlight the essential contributions of contrastive learning and the BAN module to the effectiveness of CSCL-DTI.

2) model design: We design GCN-based and Transformer-based encoder to learn drug and target representations based on molecular graph and anmino acid sequences, respectively. We separately replaced two encoders in our model with other existing models to validate that these selected modules are effective. Specifically, the models "*GAT*" and "*GIN*" denote that the molecular graph encoder on drug representation is replaced by GAT and GIN, respectively, and the remaining settings remain the same as the full model. Additionally, the model "*CNN*" adopts CNN module to obtain the target representation from target sequences. As illustrated in Fig. 2b, CSCL-DTI shows considerable performance advantage over these variants. The performance of "*GAT*" model is close to the full model, with a difference of just 0.005 in AUC and 0.008 in AUPR, respectively.

E. Parameter analysis

The performance of our model is influenced by several significant parameters, such as the number of GCN layers n, weight factor α and β, learning rate p, drop rate λ and temperature parameter T. Note that here all the sensitivity analysis experiments are conducted on the GPCR dataset, and the analysis of the parameter p, λ, and T is provided in the **Appendix III**[1].

1) Impact of the number of GCN layers: We evaluate our model by varying n from 1 to 4 with a step value of 1. Fig. 3a displays how the performance gradually rises and finally falls as n varies, with $n = 2$ achieving its best performance.

2) Impact of weight factor: Weight factor α and β represent the proportions of contrastive loss and cross-entropy loss in the total loss. To evaluate their impact, we choose their values from $\{0.01, 0.1, 1, 1.5, 50, 100\}$. Fig. 3b and Fig. 3c demonstrate that the best performance is achieved when α and β are simultaneously set to 0.1 and 1.5, respectively.

F. Case Study

We conducted a case study on the DrugBank dataset to further validate the effectiveness of our proposed CSCL-DTI. Specifically, we applied CSCL-DTI for *de novo* predictions on the important drug *Diacerein* (DrugBank ID: DB11994) and target *Aspartate aminotransferase* (Uniprot ID: Q2TU84), respectively. Table III presents the top 10 predicted candidate targets for the new drug *Diacerein* predicted by CSCL-DTI among a total of 4,254 targets. From the result we can observe that 5 out of 10 targets were successfully predicted (marked in bold). The details of predicting candidate drugs for the

Table. III. The predicted candidate targets for new drug *Diacerein*.

Rank	Target name	Target Uniprot ID	Evidence
1	**Nitric oxide synthase**	**P35228**	**PMID: 12747270**
2	Retinoic acid receptor gamma	P13631	Unconfirmed
3	Muscarinic acetylcholine receptor	P11229	Unconfirmed
4	**Glutamate receptor**	**Q05586**	**PMID: 24121043**
5	**TGF-beta receptor type-2**	**P37173**	**PMID:10329300**
6	**Cytochrome P450**	**P11510**	**PMID: 34821124**
7	**Bile salt export pump**	**O95342**	**PMID: 29355060**
8	Cholinesterase	P06276	Unconfirmed
9	Estrogen receptor	P03372	Unconfirmed
10	Prostaglandin G/H synthase 1	P23219	Unconfirmed

new target *Aspartate aminotransferase* can be found in the **Appendix IV**[1]. All these results suggested that CSCL-DTI is a beneficial method for accurately predicting interacting candidates for unknown drugs and targets.

IV. Conclusion

In this paper, we introduce CSCL-DTI, a novel dual-channel contrastive learning model for predicting drug-target interactions. Specifically, the proposed CSCL-DTI employs cross-view contrastive learning and self-supervised contrastive learning for drug multi-view representations and target protein amino acid sequence representations, respectively. This effectively captures the intrinsic relationships between different views. Comprehensive experiments demonstrate that CSCL-DTI outperforms seven state-of-the-art methods. This study focuses on utilizing a protein sequence, drug sequence, and molecular graph as input. In the future, we will focus on enhancing the alignment of specific views in CSCL-DTI by incorporating cross-view features for target representation, such as the binding pockets of 3D proteins.

References

[1] Ali Ezzat, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, 20(4):1337–1357, 2019.

[2] Xiangxiang Zeng, Fei Wang, Yuan Luo, Seung-gu Kang, Jian Tang, Felice C Lightstone, Evandro F Fang, Wendy Cornell, Ruth Nussinov, and Feixiong Cheng. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, 2022.

[3] Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585, 2023.

[4] Bo-Wei Zhao, Xiao-Rui Su, Peng-Wei Hu, Yu-An Huang, Zhu-Hong You, and Lun Hu. iGRLDTI: an improved graph representation learning method for predicting drug–target interactions over heterogeneous biological information network. *Bioinformatics*, 39(8):btad451, 2023.

[5] Shuya Li, Tingzhong Tian, Ziting Zhang, Ziheng Zou, Dan Zhao, and Jianyang Zeng. PocketAnchor: Learning structure-based pocket representations for protein-ligand interaction prediction. *Cell Systems*, 14(8):692–705, 2023.

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

[7] Ingoo Lee, Jongsoo Keum, and Hojung Nam. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[8] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020.

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

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

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

[12] Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

[13] Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023.

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

[15] Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Sudipta Seal, and Ozlem Ozmen Garibay. AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics*, 23(4):bbac272, 2022.

[16] Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350, 2023.

[17] Xuan Lin, Kaiqi Zhao, Tong Xiao, Zhe Quan, Zhi-Jie Wang, and Philip S Yu. DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction. In *European Conference on Artificial Intelligence (ECAI)*, pages 1301–1308. IOS Press, 2020.

[18] Zhongjian Cheng, Qichang Zhao, Yaohang Li, and Jianxin Wang. IIFDTI: predicting drug–target interactions through interactive and independent features based on attention mechanism. *Bioinformatics*, 38(17):4153–4161, 2022.

[19] Greg Landrum et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

[20] Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017.

[21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. Proceedings of Machine Learning Research, 2020.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[23] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

[25] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229, 2015.

[26] AutoDock Vina. Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading trott, oleg; olson, arthur j. *J. Comput. Chem*, 31(2):455–461, 2010.

[27] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

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