OXFORD

Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring

Tri Minh Nguyen, Thin Nguyen and Truyen Tran

Corresponding author. Tri Minh Nguyen, Applied Artificial Intelligence Institute, Deakin University, Victoria, Australia. Tel: (061) 416429415;
E-mail: minhtri@deakin.edu.au

Abstract

Predicting the drug-target interaction is crucial for drug discovery as well as drug repurposing. Machine learning is commonly used in drug-target affinity (DTA) problem. However, the machine learning model faces the cold-start problem where the model performance drops when predicting the interaction of a novel drug or target. Previous works try to solve the cold start problem by learning the drug or target representation using unsupervised learning. While the drug or target representation can be learned in an unsupervised manner, it still lacks the interaction information, which is critical in drug-target interaction. To incorporate the interaction information into the drug and protein interaction, we proposed using transfer learning from chemical–chemical interaction (CCI) and protein–protein interaction (PPI) task to drug-target interaction task. The representation learned by CCI and PPI tasks can be transferred smoothly to the DTA task due to the similar nature of the tasks. The result on the DTA datasets shows that our proposed method has advantages compared to other pre-training methods in the DTA task.

Keywords: protein–protein interaction, chemical–chemical interaction, drug-target affinity, transfer learning

Introduction

Predicting the drug-target interaction is an important task in drug discovery and drug repurposing [40]. Experimental assays provide a precise but expensive tool to determine the binding affinity. On the other hand, computational methods have gained attraction due to their low cost and reasonable performance [17].

Over the years, many machine learning-based drug-target affinity (DTA) prediction methods [9, 30–32] have been proposed. However, these computational methods face the cold-start challenge where the model performance drops in novel drugs or targets, which are common in drug discovery or drug repurposing.

Pre-training is an effective method to handle the cold-start problem. Pre-training helps the model to learn a robust and generalized representation by tapping into a huge amount of unlabeled and labeled data from other relevant tasks. Because both chemicals and proteins can be represented as sequences, language modeling is one of the common pre-training tasks. Thanks to the huge available unlabelled dataset, the model can learn the internal structure arrangement, or in short, the grammar of molecules and proteins by predicting the masked tokens in the sequences. Other pre-training methods

such as pre-training graph neural networks, contrastive learning can be either share the same principle as the language model or use different schemes such as mutual information. All the unsupervised pre-training methods share the common strategy that exploits the relationship among components of the structure or between structure classes. These components can vary significantly across atoms, residues or functional groups. These relationships between components can help the model to learn the meaningful representation of each token as well as the whole sequence.

Even though the unsupervised pre-training can model the intra-molecule interaction within the molecule or protein to provide the contextual information in the representation, it still lacks the inter-molecule interaction information. By saying inter-molecule interaction, we mean the interaction between the molecule or protein with other entities. Because the essence of the drug-target interaction is in the inter-molecule interaction, it raises the question of whether the intra-molecule interaction information learned by the language model is sufficient for the DTA task.

To incorporate the inter-molecule interaction into the protein or molecule, we propose a transfer learn-

Tri Minh Nguyen is a PhD student at Deakin University. His research interest is applying machine learning in studying the protein structure and function. He is exploring the application of energy-based models in protein structure and function.

Thin Nguyen is a Senior Research Fellow with the Applied Artificial Intelligence Institute (A2I2), Deakin University, Australia. He graduated with a PhD in Computer Science from Curtin University, Australia. His current research topic is inter-disciplinary, bridging large-scale data analytics, pattern recognition, genetics and medicine. His research direction is to develop machine learning methods to discover functional connections between drugs, genes and diseases.

Truyen Tran is an associate professor at Deakin University. He is member of Applied Artificial Intelligence Institute where he leads the work on deep learning and its application on health, genomics, software and materials science. His other research topics include probabilistic graphical models, recommender systems, learning to rank, anomaly detection, multi-relational databases, model stability and mixed-type analysis.
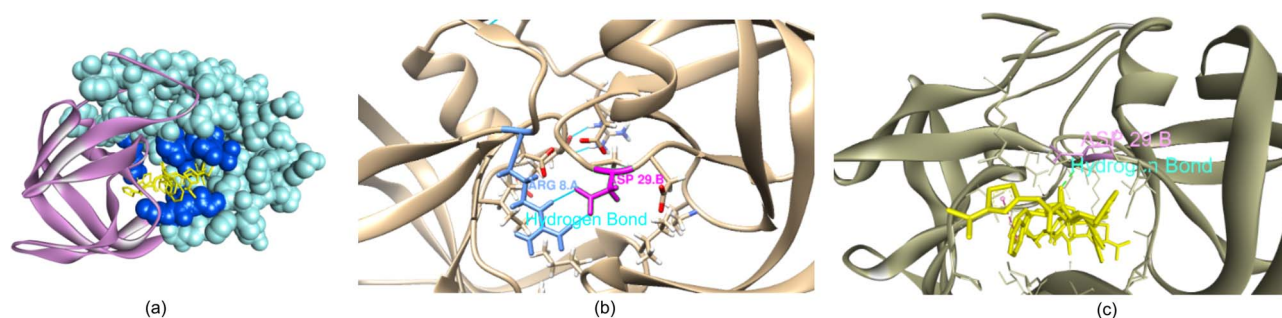
Figure 1. Example of how information from PPI task can be transferred to DTA task. (**A**) Crystal structure of the complex of resistant strain of HIV-1 protease (v82a mutant) with Ritonavir. (**B**) The hydrogen bond in PPI at the protein interface. (**C**) The binding site of Ritonavir in the proximity of protein interface.

ing framework called **C**hemical-**C**hemical **P**rotein-**P**rotein Transferred DTA (**C2P2**). First, C2P2 transfers the inter-molecule interaction knowledge learned from chemical–chemical interaction (CCI) and protein–protein interaction (PPI). Then we combine the inter-molecule interaction with the intra-molecule interaction knowledge to learn the drug-target interaction space.

PPI is the physical interaction between two or more protein macro-molecules. This interaction is the result of the electrostatics forces, hydrogen bonding or hydrophobic effect of the residues at the protein interface [22]. The properties of the protein interface such as size and shape, complementary between surfaces, residue interface propensities, hydrophobicity, segmentation, secondary structure and structure flexibility [22]. Even though the protein interface is usually viewed as large, flat, featureless and usually described as undruggable [4, 5, 19], the PPI can reveal the effective drug-target binding mode [15]. Previous works have taken advantage of PPI in drug discovery [2, 6, 15]. In addition, the distribution of the protein interface can indicate the distribution of ligand-binding pocket. Previous work [16] shows that in the protein–protein complex, the majority of ligand binding pockets are with 6 Amstrong (Å) of the protein interface. Looking at Figure 1, the hydrogen bond between ARG8 and ASP29 in the protein–protein complex (Figure 1b) also exists in the binding configuration with Ritonavir. Therefore, the information from the protein–protein can be beneficial for the drug-target interaction.

CCI is the interaction between two chemical entities. The interaction can be derived from various ways such as pathway databases, text mining, structure or activities similarity [27]. The DTA model can benefit from CCI information in many ways. Reaction pathway can describe how closely two molecules are related in a successive reaction chain and their association. Structure and activities similarity between two or more molecules can reveal the core structure and their roles in the binding. Ligand sharing the same pharmacological action is usually predicted to share the same target. CCI can provide information for many related tasks such as toxicity, combination therapies effect, biological functions and drug-target bindings [28] to speed up the drug discovery process [7]. In addition, amino acid alone is also a molecule. We
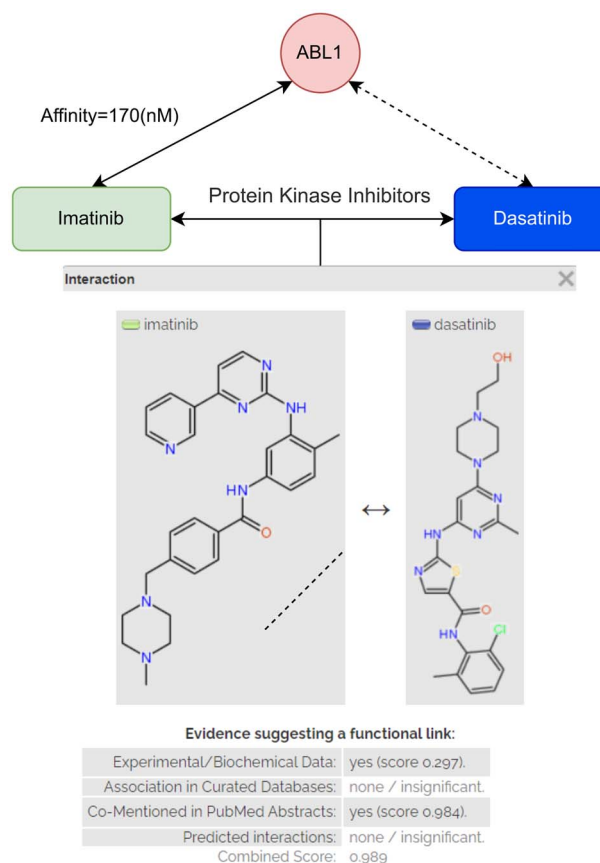

Figure 2. Chemical–chemical interaction provides external information for drug-target binding. Both Imatinib and Dasatinib share the MeSH pharmacological action 'Protein Kinase Inhibitors' reported in the experimental data of STITCH [27] database. The CCI report is generated by STITCH database web server tool [27].

can formulate the residue–ligand interaction as a CCI in which the interaction is the hydrogen bonding, Van der Waals force or electrostatics (Figure 2). The physical interaction between molecules, non-covalent or covalent, may suggest their interaction with amino acids. In this case, the information from the CCI task can be beneficial for learning the residue-ligand interaction, thus protein-ligand interaction.

Our contribution is 2-fold. First, we propose enhancing the drug-target interaction prediction framework with not only inter-molecule interaction learned from

language modeling but also intra-molecule interaction learned from related tasks such as PPI and CCI. We integrate the intra-molecule interaction information into unsupervised pre-training to enhance the representation in (DTA) task where understanding interaction is the key factor. Second, we provide different ways to integrate the learned intra-molecules information into sequence representation and graph representation.

Related works
Learning protein representation
Sequence representation

Recent developments [12, 29] in natural language processing allow the learning model to capture the contextual relationship between tokens in the sequence from a large amount of unlabeled sequence data to achieve state-of-the-art performance on many tasks. The success of the language modeling approach is transferred to protein sequence modeling. TAPE [34] learns the protein embedding using language model Transformer [12] with 31 million sequences from the Pfam dataset [13]. Rives et al. [35] train the language model varying in size in the same manner as TAPE on 250 million sequences of UniRef [38] dataset. ProtTrans [14] uses auto-regressive models (Transformer-XL, XLNet) and auto-encoder models (BERT, Albert, Electra, T5) to learn the protein embedding from 2.1 billion protein sequences. In addition to the language model, dilated-CNN and BiLSTM are also used to perform the sequence encoding [33].

3D structure representation

In the sequential representation, the structure information is lost. Another way to represent the protein is using the exact 3D structure information, meaning using the 3D coordinate to represent each residue. However, acquiring the protein-folding information through experimental methods such as X-ray can be time-consuming or expensive. Therefore, several computational methods are proposed [23, 24] to compute high-resolution protein structures. The predicted 3D structure can be used to construct the detailed protein surface using point cloud [10] or multi-scale graph structure [36]. However, predicting the atom's coordinate with high accuracy requires large computational resources. In addition, encoding the whole protein structure to the atom level may lead to sparse representation and inefficient computational resource usage. Therefore, a more simple representation can be beneficial.

Protein graph representation

To balance between 3D structural information and simplicity, 2D representation via attributed graph can be used. Previous works [21, 31] have been using protein structure graph representation for DTA prediction. The contact/distance map is used as the adjacency matrix of an attributed graph where each node represents a residue and edge represents the contact/distance between residues. The node attribute can be simply a one-hot encoding of residue type [21] or an embedding vector of the residue obtained from the language model [31].

Learning molecule representation
Sequence representation

The molecules can be represented as SMILES sequence. Therefore, we can apply language modeling to learn the embedding of the molecules. Recent works [8, 44] uses LSTM and Transformer to learn the SMILES sequence representation of chemical space from over 77 million SMILES sequences of PubChem dataset [25]. Chemical SMILES language modeling is essentially an atom level pre-training where the model can learn the intra-interaction of the molecule. The molecule SMILES sequence representation can also be merged with structural information like fingerprint to have both motifs and context dependency information [33].

Graph representation

Graph is the natural representation of the molecule in which the atoms are nodes and bonds are edges. The pre-training method on graph neural network allows the model to capture the robust representation at atom level and molecules level. On node level pre-training, Weihua et al. [20] propose both node-level pre-training via attribute masking and context prediction task and graph-level pre-training via transfer learning from graph attribute and graph structure prediction. On graph level pre-training, InfoGraph [37] maximizes the mutual information between supervised and unsupervised representation. Node level pre-training can help the model to learn the intra-interaction and internal structure at atom level while graph level pre-training allows the model to learn a robust representation of graph structure within the same molecule class.

Methods

DTA problem is predicting the binding affinity A between a drug compound D and a protein P. Mathematically, the DTA prediction problem can be formulated as a regression task, minimizing the loss function of the predicted affinity value F_θ of drug-target pair (P_i, D_i) and the actual affinity value Y:

$$\mathcal{L}(P_i, D_i, Y_i) = f_L(\mathcal{F}_\theta(P_i, D_i), Y_i), \qquad (1)$$

where θ is model parameters of predicting function $\mathcal{F}$ and f_L is implemented loss function.

The cold start in DTA prediction is inferring the binding affinity of drugs and proteins, which do not appear in the training set. Formally, we define the cold start problem for drugs (cold-drug) as follows. During the training time, we train the model with the set of proteins X_p and drugs X_d. During the testing time, we are given a set of new
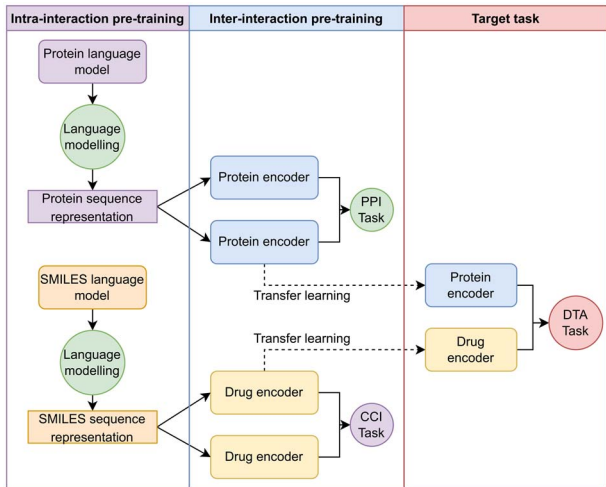
Figure 3. The framework architecture of the C2P2 model. First, the protein encoder and drug encoder are trained with PPI task and CCI task, respectively. Then pre-trained encoders are used for drug and target encoding in the DTA model.

drugs $X_{nd} = \{x_{nd_1}, ..., x_{nd_n}\}, x_{nd_i} \notin X_d$ while the protein set remains the same. Cold-target is similar but the model is tested with new protein set $X_{np} = \{x_{np_1}, ..., x_{np_n}\}, x_{np_i} \notin X_p$.

In this section, we present our framework to combine the intra-molecule interaction from language modeling with the inter-molecule interaction knowledge learned from PPI and CCI tasks. In Sec. Overall framework , we present the overall framework of C2P2, followed by learning inter-molecule and intra-molecule interaction with language modeling, CCI, and PPI task. Then Sec. Integrating inter-molecule interaction into DTA model introduces the combination of the inter-molecule and intra-molecule interaction to predict the binding affinity.

Overall framework

The overall framework is presented in Figure 3. The goal is to transfer the interaction learned from the source domain, which is PPI and CCI task, to the target domain DTA task. First, the protein and drug encoder is pre-trained with PPI and CCI tasks. The benefits of pre-training the protein and drug encoder with PPI and CCI tasks are 2-fold: better generalization representation and interaction-oriented representation. By better generalization representation, we mean that the encoder can learn from a large amount of drug and protein samples from PPI and CCI tasks. Interaction-oriented representation means that the encoder can learn the binding interaction of many different drugs and proteins. Then the pre-trained drug and target encoders are transferred to the target domain DTA task to extract the drug and target interaction-oriented representation. Finally, both drug and target representation are combined to predict the binding affinity.

Learning chemical inter-molecule interaction space

In this section, we propose the framework to learn the chemical inter-molecule interaction via the CCI

Table 1. Molecule feature vector

Feature	Feature length
Element types	43
Degree	10
Number of hydrogens	10
Implicit valence	10
Aromatic	1

prediction task. The overall framework consists of two main stsvg: learning molecule representation and interaction inference. Our CCI model takes two chemical SMILES sequences D_{s1} and D_{s2} as the inputs. The molecule representations of two SMILES sequences can be either graph representations (Sec. Graph representation of drug molecule) or language model representations (Sec. Molecule SMILES representation by language modeling). Then both representations of D_{s1} and D_{s2} are joined for CCI. By learning the CCI, our goal is pre-training the molecule encoder to encode the interaction imbued molecule representation.

Graph representation of drug molecule

Figure 4 shows the architecture of CCI task with a graph neural network. Our CCI framework takes the graph structure $\mathcal{G}_1$ and $\mathcal{G}_2$ of two molecules. The molecule graph structure $\mathcal{G}_1$ has nodes representing the atoms and edges representing the bonds.

$$\mathcal{G}_1 = (\mathcal{X}_1, \mathcal{A}_1) \qquad (2)$$

where $\mathcal{X}_1 = [p_1, ..., p_N]$ is the node feature matrix of N nodes where each node is represented by vector p_i and $\mathcal{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix that describes the graph structure. The molecule graph structure $\mathcal{G}_2$ is also constructed in the same manner.

The atom node feature $\mathcal{X}$ is its element type, degree, number of Hydrogens, and implicit valence. The detail of the feature vector of the molecule graph node is shown in Table 1. The graph representation is learned using graph isomorphism network (GIN) [45]. The graph neural network updates the node feature vector by:

$$p_v^{(k)} = MLP^{(k)}((1 + \epsilon^{(k)}) \cdot p_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} p_u^{(k-1)}) \in \mathbb{R}^{C^{(k)}} \quad (3)$$

where $\epsilon^{(k)}$ is a trainable parameter, MLP is a multi-layer perceptron, $p_v^{(k)} \in \mathbb{R}^{C^{(k)}}$ is the kth layer feature vector of vth node, $C(k)$ is the feature vector dimension at k^{th} layer.

After kth GIN layers, we have the $\mathcal{P}'_d = \{p'_i \in \mathbb{R}^{h_1} \mid i \in 1, 2, ..., S\}$ as node features of molecule graph, where S is the number of nodes in the drug graph, h_1 is the dimension of the node feature vector. Then we use the max pooling operation followed
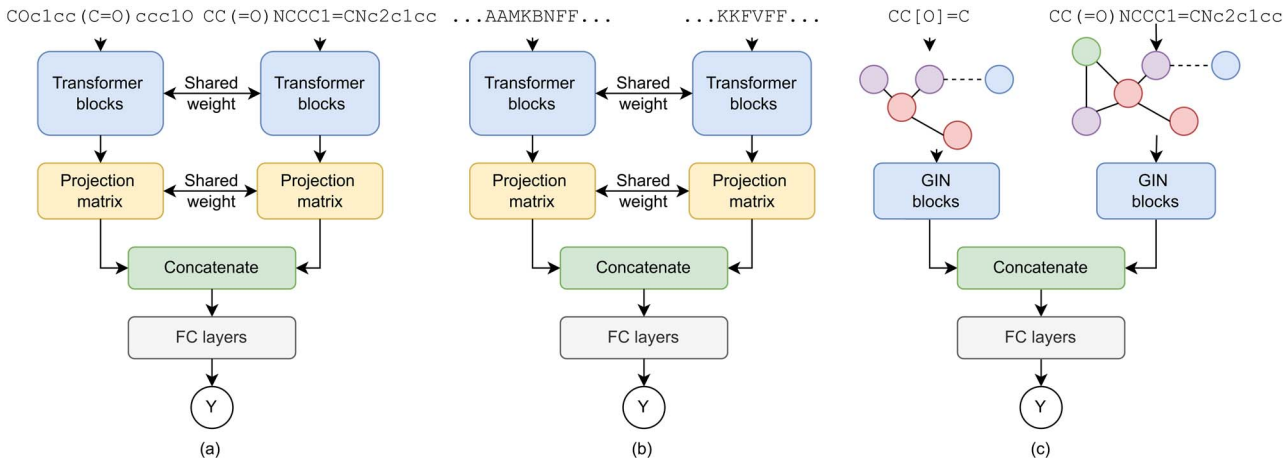
Figure 4. Learning and enhancing the drug and protein representation from (**A**) SMILES sequence encoder, (**B**) protein sequence and (**C**) molecule graph with interaction from CCI and PPI tasks.

by linear layers for feature projection:

$$p'_{\max} = \text{MaxPool}(\mathcal{P}'_d), \tag{4}$$

$$x_d = (W_0 p'_{\max} + b_0)W_1 + b_1. \tag{5}$$

where W_0, b_0 and W_1 b_1 are trainable weight and bias of two linear layers. Finally, we obtain x_d as the feature vector of the drug molecule.

Molecule SMILES representation by language modeling

Figure 4 shows the architecture of enhancing the molecule representation learned from the language model with the interaction information. As the language model tends to learn the internal arrangement (grammar structure) which is essentially the internal interaction. To enhance the language model representation with molecule inter-molecule interaction information, we fine-tune the language model on the CCI task.

Given the SMILES sequence D_s with length n, SMILES sequence representation is extracted using the pre-trained Transformer blocks. We use the BERT language model named ChemBERTa pre-trained on SMILES sequence [8].

$$X_s = \text{BERT}(D_s), X_s \in \mathbb{R}^{n \times d} \tag{6}$$

where d is the dimension of the embedding vector. ChemBERTa uses the RoBERTa [29] Transformer implementation. Transformer [41] is attention-based architecture commonly used in language modeling. The language modeling pre-training task is predicting the masked character in the SMILES sequence. ChemBERTa is trained on 10 million SMILES sequences from the PubChem dataset. The ChemBERTa language model is directly used as sequence feature extraction without any fine-tuning. Then the sequence feature vector x_s is the

average along feature vector:

$$x_s = \text{AVG}(X_s), x_s \in \mathbb{R}^d \tag{7}$$

Then the sequence representation x_s is projected into lower dimension using linear layer:

$$x_d = (W_{\theta_d} x_s + b), x_d \in \mathbb{R}^{d'}, d' < d \tag{8}$$

where W_{θ_d} and b are trainable weight and bias of linear layer. The goal of the linear layer is to learn to extract important features from the sequence representation and reduce noise. The Transformer and projection matrices in both branches are shared weight to reduce the number of parameters.

Chemical–chemical prediction

The SMILES sequences from two chemical $D_s 1$ and $D_s 2$ are encoded into x_{d1} and x_{d2} by either the graph neural network (Sec. Graph representation of drug molecule) or the pre-trained language model (Sec. Molecule SMILES representation by language modeling). Then both chemical representations are joined with a simple concatenate operator:

$$x_{dj} = [x_{d1}; x_{d2}] \tag{9}$$

Finally, the interaction is predicted with a classifier:

$$y = sigmoid(RELU(Wx_{df} + b)) \tag{10}$$

where Wx_{df} and b are trainable weight and bias of linear layer.

Learning protein inter-molecule interaction space

Protein sequence representation by language modeling

Figure 4 presents the PPI prediction model. The goal is to enhance the protein sequence representation learned by the language model with the protein interaction. Given two protein sequences D_{p1} and D_{p2} length n, the protein

sequence embedding X_p is extracted by a protein language model named ESM [35].

$$X_p = ESM(D_p), X_p \in \mathbb{R}^{n \times d} \tag{11}$$

where d is the embedding dimension. ESM is an attention-based Transformer [12] language model. ESM is pre-trained with predicting masked tokens in the protein sequence. ESM language model is pre-trained on UniRef50 dataset [38]. Similar to SMILES sequence representation in Eq. 6, we also directly use ESM protein language model as a protein sequence feature extraction without fine-tuning the language model. The protein sequence embedding is averaged along dimension d:

$$x_p = AVG(X_p), x_p \in \mathbb{R}^d, x_p \in \mathbb{R}^{d'}, d' < d \tag{12}$$

The protein sequence representation x_p is projected into lower dimension using linear layer:

$$x_p = (W_{\theta_p} x_s + b). \tag{13}$$

Protein–protein interaction prediction

Given the two protein sequence representations x_{p1} and x_{p2} of two input protein sequences $p1$ and $p2$, the joint representation is:

$$x_{pj} = [x_{p1}; x_{p2}] \tag{14}$$

where $[;]$ is the concatenate operator. The $p1 - p2$ interaction is predicted by:

$$y = sigmoid(RELU(Wx_{pf} + b)) \tag{15}$$

where Wx_{pf} and b trainable weight and bias of linear layer.

Integrating inter-molecule interaction into DTA model

After being pre-trained with CCI (Sec. Learning chemical inter-molecule interaction space) and PPI task (Sec. Learning protein inter-molecule interaction space), the drug encoder $f(D_s, \theta_d)$ and protein encoder $f(D_p, \theta_p)$, where θ_p and θ_d are model parameters, are used to encode the protein and drug:

$$x_p = f(D_p, \theta_p) \tag{16}$$

$$x_d = f(D_s, \theta_d) \tag{17}$$

The protein-drug joint representation is:

$$x_{pdj} = [x_p; x_d] \tag{18}$$

Table 2. The train, valid and test set split for Davis and PDBBind v2019.

Dataset	Setting	Split	Number of sample
Davis	Cold-target	Train	15708
		Valid	3877
		Test	4964
	Cold-drug	Train	19006
		Valid	4862
		Test	6188
PDBBind v2019	Cold-target	Train	9134
		Valid	2282
		Test	2595
	Cold-drug	Train	8927
		Valid	2256
		Test	2828

Finally, the binding affinity is predicted by:

$$y_a = (W_0 x_{pdj} + b_0)W_1 + b_1 \tag{19}$$

where W_0, b_0 and W_1 b_1 are trainable weight and bias of two linear layers.

Experiments
Dataset

We use the STRING dataset [39] for the PPI task. The STRING dataset is the protein–protein network database from over 67.6 million proteins with over 20 billion protein–protein pairs. The protein–protein association includes text mining from literature, interaction experiments, computational experiments and systematic interaction transferring. As we only need the protein physical interaction, we filter out other types of protein–protein association such as text mining.

For the CCI task, we use the STITCH dataset [27]. The dataset contains over 0.5 million chemicals with over 1.6 billion interactions. The chemical–chemical associations are built from the experimental results from pathway dataset, text mining from literature, structural similarity and activities similarity. The drug encoder is pre-trained by either full STITCH dataset or only experimental association.

For the DTA task, we conduct our experiments on two popular DTA datasets: Davis [11] and PDBBind v2019 [42, 43]. In the DTA task, we test our proposed method in cold-start settings, including cold-drug and cold-target. We follow previous works [1, 47] on cold start splitting process. In the cold drug setting, all drugs in the validation and test set are absent from the training set. In cold target setting, all targets in the validation and test set are absent from the training set. The train/valid/test split of Davis and PDBBind v2019 are provided in Table 2

Benchmark

We use four benchmark methods to evaluate the performance of extra-interaction transfer learning on different representations. First, we compare our

Table 3. Hyper-parameters in the experiments.

Hyper-parameters	Value
Learning rate	[0.0005:0.005]
Batch size	[128; 256; 512; 1024]

proposed method with the previous SOTA method GraphDTA [30]. GraphDTA uses CNN as protein encoder and graph neural network as drug encoder. Then the second benchmark method is ESMDTA which replaces the CNN protein encoder with protein representation pre-trained with protein language model ESM [35]. The third benchmark is ChemBERTaDTA that replaces the graph encoder with SMILES sequences language model representation [8]. Finally, to evaluate with other graph pre-training strategies, we compare our method with Infograph pre-training method [37]. We evaluate the model performance on the test set using Root Mean Squared Error (RMSE), Pearson [3], Spearman [48] and Concordance Index (CI) [18].

Implementation detail

Our methods are implemented using Pytorch. The source code and data is available at https://github.com/ngminhtri0394/C2P2. The hyper-parameters are tuned using the validation set. The hyper-parameters detail reported in Table 3. The results are reported on the independent test set. The protein language model ESM embedding dimension is $d = 768$, which is later projected to $d' = 128$ (Eq. 12). The ChemBERTa embedding dimension $d = 768$ is projected to $d' = 128$ (Eq. 8). The model is trained with MSE loss using Adam optimizer for 500 epochs. The number of GIN layers (Sec. Graph representation of drug molecule) $k = 5$.

Results and Discussion
Inter-molecule interaction knowledge benefits the DTA task

We demonstrate the advantages of transferring the inter-molecule interaction learned from PPI and CCI tasks to the DTA tasks in cold-drug and cold-target settings across two benchmark datasets with balance distribution (PDBBind dataset) and long-tail distribution (Davis dataset).

In the cold-target setting, we group the proposed methods by the drug encoder and compare the performance between models with and without PPI transfer learning. Overall, the models with PPI transfer learning show advantages compared with the models without transfer learning. With the graph-based drug encoder (GIN, GIN-CCI and Infograph), PPI enhanced models have better overall performance compared to model using only ESM feature. Looking at the language model-based drug encoder, the combination of ChemBERTa as drug encoder and ESM–PPI as protein encoder consistently outperforms the model with only ESM as protein encoder. However, combining ChemBERTa-CCI with

ESM feature outperforms ESM–PPI feature across two datasets. This suggests some degree of incompatibility between ChemBERTa-CCI and ESM–PPI in the cold-target setting. In the end, in general, cooperating the intra-molecule information learned from PPI task with a protein language model such as ESM benefits the DTA task performance.

Similar to the cold-target setting, for the cold-drug setting, we group the proposed models by protein encoder and compare the performance of models with and without CCI transfer learning. Among graph-based drug encoders, pre-training graph neural network with CCI task outperforms Infograph pre-training and training from scratch across two datasets and two types of protein encoder. In case of language model-based drug encoder, while pairing with ESM protein encoder, models with CCI pre-training have better performance than models without pre-training. However, ChemBERTa-CCI and ESM–PPI show a certain degree of incompatibility shown in lower performance than ChemBERTa and ESM–PPI pair. Overall, integrating CCI information into DTA models enhances the DTA model performance, especially in graph representation.

It is worth noting that the impact of pre-training encoder using auxiliary task CCI in cold-target is minimal and vice versa. In the cold-target scenario, the drugs in the test set are also in the training set. The encoder has already learned the representation as well as the interaction information of the test set's drugs. As the result, the external information from auxiliary task CCI is redundant. This is also the case for PPI pre-training in the cold-drug setting. To further verify this point, we conduct our experiment in a warm setting in which drugs and proteins in the test set are also in the training set. The results from Tables 8 and 9 show the similarity in the performance level of pre-trained and non-pretrained models.

The performance of encoder architecture design is reported in Table 10. For the CCI task, we follow the evaluation process of previous work [46] and report on CCI700, CCI800 and CCI900 dataset [46]. The performance of sequence encoder ChemBERTa is quite similar to graph encoder GIN in CCI700 and CCI800 and slightly better in CCI900. However, it is difficult to conclude the correlation between the performance in auxiliary tasks and the downstream tasks as many factors are interfering with the performance such as data distribution of pre-training and downstream dataset, the DTA model design.

Protein–protein interaction knowledge enhances protein language model representation

Figure 5 shows the t-SNE plot of protein embedding with ESM encoder and ESM–PPI encoder using PDBBind cold-target test set. We also annotate the plot with druggability obtained from 'NonRedundant dataset of Druggable and Less Druggable binding sites' (NRDLD) dataset [26]. In the PDBBind cold-target setting test set, the Glucarate

Table 4. The performance of the different drug and protein encoder combinations on Davis dataset with the cold-target setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI and Infograph are pre-training with PPI, CCI task or Infograph unsupervised training. In this experiment, we compare the protein encoder ESM with (ESM–PPI) and without PPI pre-training (ESM) in the same drug encoder setting to demonstrate the effectiveness of PPI pre-training in cold-target scenario. The numbers in bold indicate the top performance within the same drug encoder.

Drug encoder	Protein encoder	RMSE	Pearson	Spearman	CI
GIN [30]	CNN [30]	0.696 (0.009)	0.548 (0.013)	0.439 (0.025)	0.733 (0.013)
GIN [30]	ESM [35]	0.708 (0.011)	0.579 (0.012)	0.493 (0.017)	0.764 (0.01)
	ESM-PPI (Ours)	**0.676** (0.008)	**0.589** (0.014)	**0.506** (0.014)	**0.771** (0.007)
GIN-CCI (Ours)	ESM [35]	0.741 (0.009)	0.565 (0.016)	0.454 (0.017)	0.742 (0.01)
	ESM-PPI (Ours)	**0.684** (0.009)	**0.583** (0.012)	**0.492** (0.015)	**0.763** (0.008)
ChemBERTa [8]	ESM [35]	0.784 (0.01)	0.54 (0.01)	0.41 (0.013)	0.718 (0.007)
	ESM-PPI (Ours)	**0.675** (0.01)	**0.589** (0.013)	**0.497** (0.02)	**0.765** (0.011)
ChemBERTa-CCI (Ours)	ESM [35]	0.733 (0.004)	0.557 (0.004)	0.493 (0.007)	0.763 (0.004)
	ESM-PPI (Ours)	**0.686** (0.006)	**0.581** (0.009)	**0.508** (0.009)	**0.772** (0.005)
Infograph [37]	ESM [35]	0.718 (0.006)	0.58 (0.01)	0.481 (0.011)	0.757 (0.006)
	ESM-PPI (Ours)	**0.67** (0.01)	**0.601** (0.01)	**0.517** (0.015)	**0.777** (0.009)

Table 5. The performance of the different drug and protein encoder combinations on PDBBind dataset with the cold-target setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI and Infograph are pre-training with PPI, CCI task or Infograph unsupervised training. In this experiment, we compare the protein encoder ESM with (ESM–PPI) and without PPI pre-training (ESM) in the same drug encoder setting to demonstrate the effectiveness of PPI pre-training in cold-target scenario. The numbers in bold indicate the top performance within the same drug encoder.

Drug encoder	Protein encoder	RMSE	Pearson	Spearman	CI
GIN [30]	CNN [30]	1.638 (0.034)	0.576 (0.02)	0.575 (0.025)	0.704 (0.009)
GIN [30]	ESM [35]	1.702 (0.045)	0.614 (0.057)	0.642 (0.035)	0.732 (0.016)
	ESM-PPI (Ours)	**1.397** (0.012)	**0.708** (0.006)	**0.699** (0.006)	**0.757** (0.003)
GIN-CCI (Ours)	ESM [35]	1.473 (0.013)	0.686 (0.009)	0.682 (0.007)	0.747 (0.003)
	ESM-PPI (Ours)	**1.394** (0.007)	**0.715** (0.003)	**0.703** (0.005)	**0.759** (0.002)
ChemBERTa [8]	ESM [35]	1.487 (0.023)	0.689 (0.015)	0.684 (0.012)	0.748 (0.006)
	ESM-PPI (Ours)	**1.461** (0.013)	**0.695** (0.009)	**0.688** (0.008)	**0.750** (0.004)
ChemBERTa-CCI (Ours)	ESM [35]	1.395 (0.01)	0.709 (0.004)	0.697 (0.004)	0.756 (0.002)
	ESM-PPI (Ours)	**1.390** (0.007)	**0.712** (0.005)	**0.700** (0.005)	**0.758** (0.002)
Infograph [37]	ESM [35]	1.597 (0.01)	0.634 (0.008)	0.639 (0.01)	0.729 (0.004)
	ESM-PPI (Ours)	**1.395** (0.015)	**0.710** (0.007)	**0.699** (0.008)	**0.758** (0.004)

Dehydratase (PDB:1ec9) is labeled as undruggable [26]. We can observe the clear distribution of druggable and undruggable protein in the embedding space of ESM–PPI protein encoder. We hypothesize that the knowledge learned from PPI task can assist the druggability prediction. To verify this hypothesis, we use the ESM or ESM–PPI as the input for simple SVM model for druggability classification. We use the NRDLD dataset [26] for training

Table 6. The performance of the different drug and protein encoder combinations on Davis dataset with the cold-drug setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI and Infograph are pre-training with PPI, CCI task or Infograph unsupervised training. In this experiment, we compare different types of drug encoders using the same protein encoder (ESM and ESM–PPI) to demonstrate the effectiveness of CCI pre-training in cold-drug scenario. We also group the models based on drug encoder type (graph based GIN and sequence based ChemBERTa) to further investigate the impact of CCI pre-training on molecule graph representation as well as SMILES sequence representation. The numbers in bold indicate the top performance within the same protein encoder and same drug encoder representation type (graph-based GIN and sequence-based ChemBERTa).

Protein encoder	Drug encoder	RMSE	Pearson	Spearman	CI
CNN [30]	GIN [30]	0.905	0.480	0.428	0.705
		(0.024)	(0.03)	(0.035)	(0.017)
ESM [35]	GIN [30]	1.011	0.475	0.407	0.695
		(0.07)	(0.081)	(0.088)	(0.043)
	Infograph [37]	0.970	**0.530**	0.392	0.688
		(0.021)	(0.024)	(0.046)	(0.022)
	GIN-CCI (Ours)	**0.927**	0.501	**0.436**	**0.710**
		(0.028)	(0.042)	(0.03)	(0.015)
	ChemBERTa [8]	1.048	0.433	0.358	0.671
		(0.031)	(0.071)	(0.053)	(0.027)
	ChemBERTa-CCI (Ours)	**0.982**	**0.502**	**0.441**	**0.712**
		(0.032)	(0.015)	(0.021)	(0.01)
ESM-PPI (Ours)	GIN [30]	0.985	0.496	0.416	0.699
		(0.039)	(0.049)	(0.087)	(0.043)
	Infograph [37]	0.949	0.488	0.44	0.712
		(0.019)	(0.034)	(0.05)	(0.025)
	GIN-CCI (Ours)	**0.907**	**0.526**	**0.463**	**0.723**
		(0.028)	(0.044)	(0.041)	(0.021)
	ChemBERTa [8]	1.02	0.431	0.342	0.663
		(0.033)	(0.048)	(0.026)	(0.013)
	ChemBERTa-CCI (Ours)	**0.971**	**0.506**	**0.448**	**0.716**
		(0.044)	(0.027)	(0.033)	(0.016)

Table 7. The performance of the different drug and protein encoder combinations on PDBBind dataset with the cold-drug setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI and Infograph are pre-training with PPI, CCI task or Infograph unsupervised training. In this experiment, we compare different types of drug encoders using the same protein encoder (ESM and ESM–PPI) to demonstrate the effectiveness of CCI pre-training in cold-drug scenario. We also group the models based on drug encoder type (graph based GIN and sequence based ChemBERTa) to further investigate the impact of CCI pre-training on molecule graph representation as well as SMILES sequence representation. The numbers in bold indicate the top performance within the same protein encoder and same drug encoder representation type (graph-based GIN and sequence-based ChemBERTa).

Protein encoder	Drug encoder	RMSE	Pearson	Spearman	CI
CNN [30]	GIN [30]	1.495	0.643	0.631	0.728
		(0.024)	(0.015)	(0.015)	(0.006)
ESM [35]	GIN [30]	1.622	0.588	0.611	0.718
		(0.061)	(0.036)	(0.012)	(0.006)
	Infograph [37]	1.599	0.610	0.617	0.722
		(0.017)	(0.013)	(0.013)	(0.006)
	GIN-CCI (Ours)	**1.443**	**0.683**	**0.667**	**0.742**
		(0.015)	(0.007)	(0.004)	(0.002)
	ChemBERTa [8]	1.446	0.683	0.664	0.741
		(0.013)	(0.009)	(0.007)	(0.004)
	ChemBERTa-CCI (Ours)	**1.389**	**0.695**	**0.677**	**0.747**
		(0.011)	(0.01)	(0.005)	(0.004)
ESM-PPI (Ours)	GIN [30]	1.629	0.587	0.613	0.719
		(0.046)	(0.022)	(0.016)	(0.007)
	Infograph [37]	1.591	0.609	0.617	0.721
		(0.015)	(0.009)	(0.009)	(0.004)
	GIN-CCI (Ours)	**1.438**	**0.686**	**0.669**	**0.743**
		(0.02)	(0.01)	(0.007)	(0.004)
	ChemBERTa [8]	1.423	0.689	0.668	0.743
		(0.011)	(0.006)	(0.005)	(0.002)
	ChemBERTa-CCI (Ours)	**1.387**	**0.692**	**0.678**	**0.747**
		(0.01)	(0.005)	(0.005)	(0.002)

Table 8. The performance of the different drug and protein encoder combinations on Davis dataset with the **warm** setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI and Infograph are pre-training with PPI, CCI task or Infograph unsupervised training. In this experiment, we compare the protein encoder ESM with (ESM–PPI) and without PPI pre-training (ESM) in the same drug encoder setting to demonstrate the effectiveness of PPI pre-training in cold-target scenario. The numbers in bold indicate the top performance within the same drug encoder.

Drug encoder	Protein encoder	RMSE	Pearson	Spearman	CI
GIN [30]	CNN [30]	0.506 (0.005)	0.825 (0.003)	0.69 (0.006)	0.883 (0.003)
GIN [30]	ESM [35]	0.476 (0.004)	0.848 (0.002)	0.706 (0.003)	0.895 (0.002)
	ESM-PPI (Ours)	0.476 (0.005)	0.847 (0.003)	0.703 (0.004)	0.893 (0.002)
GIN-CCI (Ours)	ESM [35]	0.477 (0.004)	0.847 (0.003)	0.701 (0.006)	0.892 (0.004)
	ESM-PPI (Ours)	0.477 (0.003)	0.847 (0.002)	0.697 (0.005)	0.89 (0.003)
ChemBERTa [8]	ESM [35]	0.483 (0.005)	0.844 (0.003)	0.7 (0.005)	0.891 (0.003)
	ESM-PPI (Ours)	0.481 (0.003)	0.844 (0.002)	0.7 (0.004)	0.891 (0.002)
ChemBERTa-CCI (Ours)	ESM [35]	0.481 (0.004)	0.843 (0.003)	0.703 (0.005)	0.893 (0.003)
	ESM-PPI (Ours)	0.479 (0.004)	0.845 (0.003)	0.704 (0.005)	0.893 (0.003)
Infograph [37]	ESM [35]	0.476 (0.004)	0.848 (0.002)	0.705 (0.005)	0.895 (0.003)
	ESM-PPI (Ours)	0.473 (0.003)	0.85 (0.002)	0.706 (0.004)	0.895 (0.002)

Table 9. The performance of the different drug and protein encoder combinations on PDBBind dataset with the **Warm** setting. The X-Y drug or protein encoder means that the base model is X and pre-trained with Y task. PPI, CCI, and Infograph are pre-training with PPI, CCI task, or Infograph unsupervised training. In this experiment, we compare the protein encoder ESM with (ESM-PPI) and without PPI pre-training (ESM) in the same drug encoder setting to demonstrate the effectiveness of PPI pre-training in cold-target scenario. The numbers in bold indicate the top performance within the same drug encoder.

Drug encoder	Protein encoder	RMSE	Pearson	Spearman	CI
GIN [30]	CNN [30]	1.6469 (0.013)	0.601 (0.006)	0.595 (0.008)	0.713 (0.003)
GIN [30]	ESM [35]	1.521 (0.014)	0.671 (0.005)	0.68 (0.006)	0.748 (0.002)
	ESM-PPI (Ours)	1.515 (0.015)	0.671 (0.01)	0.682 (0.008)	0.749 (0.004)
GIN-CCI (Ours)	ESM [35]	1.519 (0.013)	0.679 (0.005)	0.687 (0.004)	0.751 (0.002)
	ESM-PPI (Ours)	1.509 (0.011)	0.683 (0.004)	0.69 (0.003)	0.753 (0.002)
ChemBERTa [8]	ESM [35]	1.584 (0.018)	0.657 (0.007)	0.662 (0.006)	0.741 (0.003)
	ESM-PPI (Ours)	1.586 (0.013)	0.656 (0.006)	0.665 (0.005)	0.743 (0.002)
ChemBERTa-CCI (Ours)	ESM [35]	1.499 (0.015)	0.676 (0.007)	0.678 (0.007)	0.747 (0.003)
	ESM-PPI (Ours)	1.502 (0.016)	0.672 (0.007)	0.675 (0.007)	0.746 (0.003)
Infograph [37]	ESM [35]	1.534 (0.015)	0.669 (0.008)	0.681 (0.007)	0.748 (0.003)
	ESM-PPI (Ours)	1.527 (0.014)	0.666 (0.005)	0.68 (0.004)	0.748 (0.002)

and validation. The result (Table 11) indicates that the knowledge learned from PPI task can help the model learn the druggability of protein, thus assisting the DTA task.

Looking back to the complex of resistant strain of HIV-1 protease (v82a mutant) with Ritonavir in Sec. Introduction, we compare the performance of model using only ESM and model with PPI transfer learning and
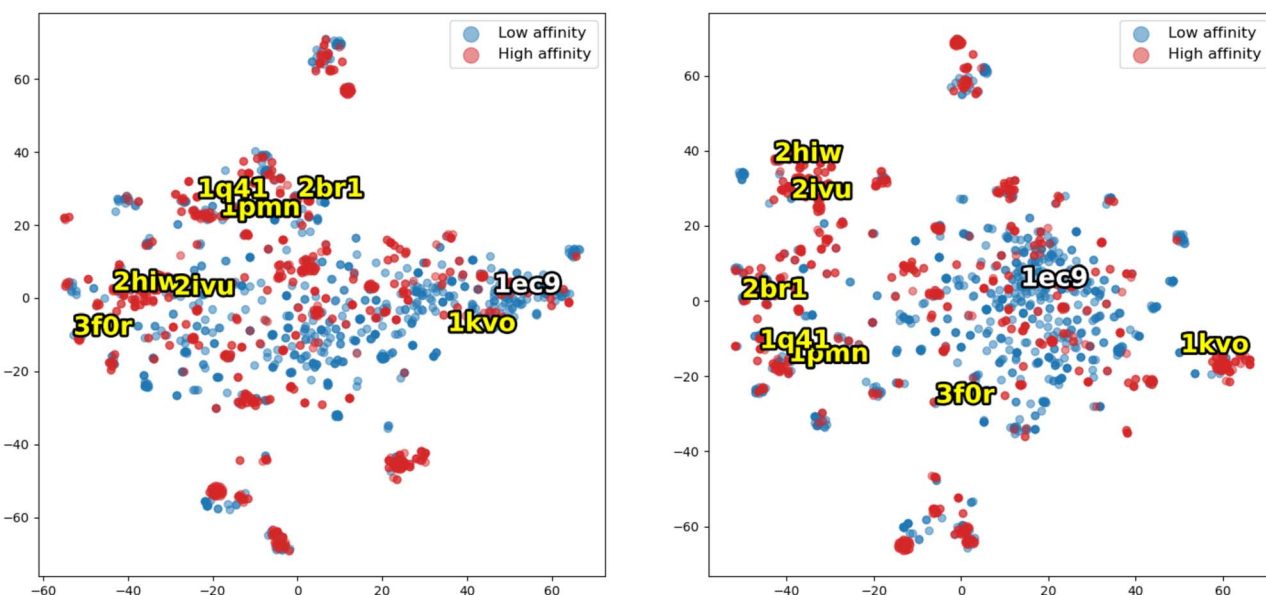
Figure 5. The T-sne plot of protein embedding of (**A**) ESM (**B**) ESM–PPI. Proteins are annotated with druggability, which is white text for non-druggable and yellow text for druggable protein. The druggability information is obtained from NRDLD dataset [26]. Each dot represents a protein in a drug-target pair in the PDBBind dataset. The low-affinity drug-target pair is in blue and the high-affinity pair is in red.

Table 10. The performance of the drug encoder and protein encoder architecture on chemical-chemical interaction and protein-protein interaction tasks.

Task	Dataset	Encoder	Accuracy	AUC	AUPR
CCI	CCI700[46]	GIN	0.964	0.993	0.993
		ChemBERTa	0.962	0.993	0.992
	CCI800[46]	GIN	0.944	0.985	0.984
		ChemBERTa	0.95	0.988	0.986
	CCI900[46]	GIN	0.868	0.935	0.922
		ChemBERTa	0.885	0.952	0.943
PPI	STRING[39]	ESM	0.758	0.878	0.725

Table 11. The result of druggability classification on NRDLD dataset [26] using ESM and ESM-PPI features with a simple SVM model. The result shows that ESM-PPI clusters the druggability, thus improving SVM model performance.

Protein encoder	Precision	Recall	F1	Accuracy
ESM	0.6803	0.8028	0.7349	0.6434
ESM-PPI	**0.6979**	**0.8742**	**0.7733**	**0.6869**

Table 12. The prediction of ESM and ESM-PPI model for the resistant strain of HIV-1 protease (v82a mutant) with Ritonavir.

Protein encoder	Predicted affinity	Error
ESM	7.2532	1.1532
ESM-PPI	6.9038	0.8038

ESM (ESM–PPI). The results in Table 12 shows that model with PPI transfer learning has a lower error rate than the model without PPI transfer learning. This implies that knowledge of protein interface and PPI integrates well into the DTA model.

Integrating different types of CCI improves the DTA prediction model performance

The CCI in STITCH dataset [27] consists of not only interaction from experimental data but also interaction in a sense of similarity between activities or structure and literature text co-occurrence. The number of experimental data is only a small proportion of full CCI data. We hypothesize that not only the experimental interaction but also other types of interaction are useful for pre-training tasks. The results in Table 13 and 14 show that pre-training with all types of CCI outperforms pre-training with only experimental data by a large margin. This suggests drug structure and activities similarity, as well as text co-occurrence can also provide useful information for DTA task.

Conclusions and Future work

In conclusion, migrating the cold-start problem in DTA prediction requires external knowledge from labeled and unlabeled data. Unsupervised learning such as language modeling learns the intra-molecule interaction and internal structure representation of the proteins and drugs from unlabeled data. The drugs and proteins representation are then imbued with inter-molecule interaction learned from similar tasks such as PPI and CCI. The PPI can provides knowledge regarding protein surface, activity, druggability. The CCI provides common pharmacological action, similarity in structure and targets. Combining both intra-molecule interaction and inter-molecule interaction information allows more robust drug and protein representation to deal with the cold-start problem. In addition, interactions curated from different resources such as text mining are also useful for learning interaction knowledge.

Table 13. The performance of the DTA model on Davis dataset with drug encoder pre-trained with only experimental interaction CCI and drug encoder pre-trained with all types of interaction available in the stitch STITCH dataset.

Protein encoder	Drug encoder	Pretrain	RMSE	Pearson	Spearman	CI
ESM	GIN-CCI	Full	**0.8755**	**0.575**	**0.5034**	**0.743**
		Exp	0.98	0.3588	0.4275	0.707
	ChemBERTa-CCI	Full	**0.9146**	**0.5259**	**0.4485**	**0.7171**
		Exp	1.07	0.346	0.3664	0.6769
ESM-PPI	GIN-CCI	Full	**0.8841**	**0.5564**	**0.4741**	**0.7299**
		Exp	1.0398	0.3595	0.3706	0.6753
	ChemBERTa-CCI	Full	**0.9171**	**0.4906**	**0.4216**	**0.7034**
		Exp	0.9181	0.4774	0.4087	0.6956

Table 14. The performance of the DTA model on PDBBind dataset with drug encoder pre-trained with only experimental interaction CCI and drug encoder pre-trained with all types of interaction available in the stitch STITCH dataset.

Protein encoder	Drug encoder	Pretrain	RMSE	Pearson	Spearman	CI
ESM	GIN-CCI	Full	**1.3484**	**0.7236**	**0.7025**	**0.7603**
		Exp	1.4053	0.6927	0.6638	0.7441
	ChemBERTa-CCI	Full	**1.3653**	**0.7059**	**0.6798**	**0.7498**
		Exp	1.3816	0.7012	0.6696	0.7454
ESM-PPI	GIN-CCI	Full	**1.3379**	**0.7282**	**0.7039**	**0.7618**
		Exp	1.4789	0.6672	0.6482	0.7351
	ChemBERTa-CCI	Full	1.3735	0.7009	0.6800	0.75
		Exp	**1.3627**	**0.7112**	**0.6835**	**0.751**

PPI is a complex interaction. Our framework focus on protein sequence learned from the language model. Because the protein is represented as a sequence, the information on protein structure and the binding site is lost. Therefore, proteins with multiple interaction sites and binding configurations are not considered during the PPI pre-training. Modeling the exact interaction between two proteins requires surface and structure information reflected in the protein encoding architecture such as graph or cloud points. Learning PPI with more dedicated architecture could potentially benefit not only DTA task but other tasks such as druggability as well. In addition, the number of the high resolution of protein–protein 3D structures is limited. Using solely protein–protein 3D information for pre-training may lower the benefits of pre-training. However, with the advance in structure prediction, e.g. AlphaFold [23], more 3D structure data become available that will open up opportunities for pre-training techniques.

Key Points
- We have proposed a deep learning DTA framework that uses inter-molecule interaction information learned from unsupervised language model and intra-molecule interaction learned from auxiliary tasks to deal with cold start problem.
- The representation learned by unsupervised pre-training tasks can be further enhanced by auxiliary tasks to encourage the model to learn the key features relevant to the task of interest.
- Knowledge regarding protein surface, activity, druggability from protein–protein interaction and common pharmacological action, similarity in structure and targets from chemical–chemical interaction allows more robust drug and protein representation to deal with cold-start problem.
- Interactions curated from different resources such as text mining and experimental results are useful for learning interaction knowledge.

Funding

References

1. Agyemang B, Wu WP, Kpiebaareh MY, *et al.* Multi-view self-attention for interpretable drug-target interaction prediction. *J Biomed Inform* 2020;**110**:103547.
2. Arkin MR, Tang Y, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem Biol* 2014;**21**(9):1102–14.
3. Benesty J, Chen J, Huang Y, *et al. Noise reduction in speech processing*, Vol. **2**. Springer Science & Business Media, 2009.
4. Blundell TL, Burke DF, Chirgadze D, *et al.* Protein-protein interactions in receptor activation and intracellular signalling. *Biol Chem* 2000;**381**(9–10):955–9.
5. Blundell TL, Sibanda BL, Montalvão RW, *et al.* Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos Trans R Soc Lond B Biol Sci* 2006;**361**(1467):413–23.
6. Fry DC. Small-molecule inhibitors of protein-protein interactions: how to mimic a protein partner. *Curr Pharm Des* 2012; **18**(30):4679–84.
7. Chen L, Yang J, Zheng M, *et al.* The use of chemical-chemical interaction and chemical structure to identify new

candidate chemicals related to lung cancer. *Plos One* 2015;**10**(6): e0128696.

8. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. In: *Machine Learning for Molecules Workshop, NeurIPS*, 2020.

9. Cichonska A, Ravikumar B, Parri E, *et al*. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Comput Biol* 2017; **13**(8):e1005678.

10. Dai B, Bailey-Kellogg C. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics* 2021; **37**(17):2580–8.

11. Davis MI, Hunt JP, Herrgard S, *et al*. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**(11):1046–51.

12. Devlin J, Chang M-W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, 4171–86.

13. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**(D1):D427–32.

14. Elnaggar A, Heinzinger M, Dallago C, *et al*. ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell* 2021;**14**(8):1–1.

15. Fry DC. Targeting protein-protein interactions for drug discovery. *Methods Mol Biol (Clifton, NJ)* 2015;**1278**:93–106.

16. Gao M, Skolnick J. The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proc Natl Acad Sci U S A* 2012; **109**(10):3784–9.

17. Gilson MK, Liu T, Baitaluk M, *et al*. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**(D1):D1045–53.

18. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005;**92**(4):965–70.

19. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov* 2002;**1**(9):727–30.

20. Weihua H, Liu B, Gomes J, *et al*. (eds). Strategies for pre-training graph neural networks. In: *The International Conference on Learning Representations*, 2020.

21. Jiang M, Li Z, Zhang S, *et al*. Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv* 2020;**10**(35):20701–12.

22. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci* 1996;**93**(1):13–20.

23. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873): 583–9.

24. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 2004;**32**(Web Server issue):W526.

25. Kim S, Chen J, Cheng T, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**(D1): D1102–9.

26. Krasowski A, Muthas D, Sarkar A, *et al*. DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *J Chem Inf Model* 2011;**51**(11):2829–42.

27. Kuhn M, von Mering C, Campillos M, *et al*. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2008;**36**(Database issue):D684.

28. Kwon S, Yoon S. End-to-end representation learning for chemical-chemical interaction prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(5):1436–47.

29. Liu Y, Ott M, Goyal N, *et al*. RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

30. Nguyen T, Le H, Quinn TP, *et al*. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* 2021;**37**(8):1140–7.

31. Nguyen TMINH, Nguyen T, Le TM, *et al*. GEFA: early fusion approach in drug-target affinity prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**(2):718–28.

32. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* 2018;**34**(17):821–9.

33. Yuqian P, Li J, Tang J, *et al*. Deepfusiondta: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. *IEEE/ACM Trans Comput Biol Bioinform* 2021;1–1.

34. Rao R, Bhattacharya N, Thomas N, *et al*. Evaluating protein transfer learning with TAPE. In: *Advances in Neural Information Processing Systems*, 2019.

35. Rives A, Meier J, Sercu T, *et al*. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2021;**118**(15).

36. Somnath VR, Bunne C, Krause A. Multi-scale representation learning on proteins. In: *Advances in Neural Information Processing Systems*, 2021.

37. Sun F-Y, Hoffmann J, Verma V, *et al*. (eds). InfoGraph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In: *The International Conference on Learning Representations*, 2020.

38. Suzek BE, Wang Y, Huang H, *et al*. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* 2015;**31**(6):926–32.

39. Szklarczyk D, Gable AL, Nastou KC, *et al*. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**(D1):D605–12.

40. Thafar M, Raies AB, Albaradei S, *et al*. Comparison study of computational prediction tools for drug-target binding affinities. *Front Chem* 2019;**7**.

41. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Curran Associates Inc., 2017, 5998–6008.

42. Wang R, Fang X, Yipin L, *et al*. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* 2004;**47**(12):2977–80.

43. Wang R, Fang X, Yipin L, *et al*. The PDBbind database: methodologies and updates. *J Med Chem* 2005;**48**(12):4111–9.

44. Winter R, Montanari F, Noé F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 2019;**10**(6):1692–701.

45. Xu K, Jegelka S, Hu W, *et al*. How powerful are graph neural networks? In: *International Conference on Learning Representations*, 2019.

46. Xu N, Wang P, Chen L, *et al*. Mr-gnn: multi-resolution and dual graph neural network for predicting structured entity interactions. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 3968–3974, California, 2019.

47. Yuan W, Chen G, Chian CY, *et al*. Fusiondta: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Brief Bioinform* 2022;**23**.

48. Zwillinger D, Kokoska S. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press, 1999.