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

Tri Minh Nguyen\textsuperscript{1,*}, Thin Nguyen\textsuperscript{1} and Truyen Tran\textsuperscript{1}

\textsuperscript{1}Applied Artificial Intelligence Institute, Deakin University, Victoria, Australia

*To whom correspondence should be addressed.

Abstract

Motivation: 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, 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.

Results: 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 drug-target affinity datasets shows that our proposed method has advantages compared to other pretraining methods in the DTA task.

Availability: The source code is available at https://github.com/ngminhtri0394/C2P2

Contact: minhtri@deakin.edu.au

1 Introduction

Predicting the drug-target interaction is an important task in the drug discovery and drug repurposing (Thafar et al., 2019). 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 (Gilson et al., 2016).

Over the years, many machine learning-based DTA prediction methods (Cichonska et al., 2017; Özünk et al., 2018; Nguyen et al., 2020, 2021) 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 which 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
Fig. 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 protein-protein interaction at the protein interface (c) The binding site of Ritonavir in the proximity of protein interface

Fig. 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 (Kuhn et al., 2008) database. The CCI report is generated by STITCH database web server tool (Kuhn et al., 2008).

2 Related works
2.1 Learning protein representation
2.1.1 Sequence representation
Recent developments (Devlin et al., 2019; Liu et al., 2019) 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 (Rao et al., 2019) learns the protein...
embedding using language model Transformer (Devlin et al., 2019) with thirty-one million sequences from the Pfam dataset (El-Gebali et al., 2019). Rives et al. (Rives et al., 2021) train the language model varying in size in the same manner as TAPE on 250 million sequences of UniRef (Suzek et al., 2015) dataset. ProfTrans (Elnaggar et al., 2021) uses autoregressive models (Transformer-XL, XLNet) and auto-encoder models (BERT, Albert, Electra, T5) to learn the protein embedding from 2.1 billion protein sequences.

2.1.2 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 (Jumper et al., 2021; Kim et al., 2004) to compute high-resolution protein structures. The predicted 3D structure can be used to construct the detailed protein surface using point cloud (Dai and Bailey-Kellogg, 2021) or multi-scale graph structure (Sonnath et al., 2021). 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, more simple representation can be beneficial.

2.1.3 Protein graph representation

To balance between 3D structural information and simplicity, 2D representation via attributed graph can be used. Previous works (Nguyen et al., 2021; Jiang et al., 2020) 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 (Jiang et al., 2020) or an embedding vector of the residue obtained from the language model (Nguyen et al., 2021).

2.2 Learning molecule representation

2.2.1 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 (Winters et al., 2019; Chithrannada et al., 2020) uses LSTM and Transformer to learn the SMILES sequence representation of chemical space from over 77 million SMILES sequences of PubChem dataset (Kim et al., 2019). Chemical SMILES language modeling is essentially an atom level pre-training where the model can learn the intra-interaction of the molecule.

2.2.2 Graph representation

Graph is the natural representation of the molecule in which the atoms are nodes and bonds are edges. 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, Wehua et al. (Hu et al., 2020) 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 (Sun et al., 2020) 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.

3 Methods

Drug-target binding affinity (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:

$$A = \mathcal{F}_\theta (P, D).$$

where $\theta$ is model parameters of predicting function $\mathcal{F}$.

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. 3.1, 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. 3.4 introduces the combination of the inter-molecule and intra-molecule interaction to predict the binding affinity.

3.1 Overall framework

The overall framework is presented in Fig. 3 and 4. 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 two folds: 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 task. 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.

3.2 Learning chemical inter-molecule interaction space

In this section, we propose the framework to learn the chemical inter-molecule interaction via the chemical-chemical interaction (CCI) prediction task. The overall framework consists of two main steps: 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. 3.2.1) or language model representations (Sec. 3.2.2). Then both representations of $D_{s1}$ and $D_{s2}$ are joined for chemical-chemical interaction. By learning the chemical-chemical interaction, our goal is pre-training the molecule encoder to encode the interaction imbued molecule representation.

3.2.1 Graph representation of drug molecule

Fig. 5 shows the architecture of CCI task with graph neural network. Our CCI framework takes the graph structure $G_1$ and $G_2$ of two molecules. The molecule graph structure has nodes representing the atoms and edges representing the bonds.

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

where $X = p_1, \ldots, p_n$ is node feature matrix of $N$ nodes and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix that describes the graph structure.

The atom node feature $X$ 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) (Xu et al., 2018). The graph neural network updates the node feature vector by:
3.2.2 Molecule SMILES representation by language modeling

Fig. 6 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 (Chithrananda et al., 2020).

$$X_s = BERT(D_s), X_s \in \mathbb{R}^{n \times d}$$  \hspace{1cm} (8)

where $d$ is the dimension of the embedding vector. The pre-training task is predicting the masked character in SMILES sequence. Then the sequence feature vector $x_s$ is the average along feature vector:

$$x_s = \text{AVG}(X_s), x_s \in \mathbb{R}^{d}$$  \hspace{1cm} (9)

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

$$x_d = (W_d x_s + b). x_d \in \mathbb{R}^{d',d' < d}$$  \hspace{1cm} (10)

The goal of the linear layer is to learn to extract important features from the sequence representation and reduce noise. The Transformer and projection matrix in both branches are shared weight to reduce the number of parameters.

3.2.3 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 graph neural network (Sec. 3.2.1) or pre-trained language model (Sec. 3.2.2). Then both chemical representations are joined with a simple concatenate operator:

$$x_{d} = [x_{d1}; x_{d2}]$$  \hspace{1cm} (11)

Finally, the interaction is predicted with a classifier:

$$y = \text{sigmoid}(\text{RELU}(W x_d + b))$$  \hspace{1cm} (12)
Mitigating cold start problems

3.3 Learning protein inter-molecule interaction space

3.3.1 Protein sequence representation by language modeling

Fig. 7 presents the protein-protein interaction 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 encoding \( X_p \) is extracted by a protein language model named ESM (Rives et al., 2021):

\[
X_p = ESM(D_p), X_p \in \mathbb{R}^{n \times d}
\]  

where \( d \) is the embedding dimension. ESM is pre-trained with predicting masked tokens in the protein sequence. The protein sequence embedding is averaged along dimension \( d \):

\[
x_p = AVG(X_p), x_p \in \mathbb{R}^{d}, x_p \in \mathbb{R}^{d'}, d' < d
\]  

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

\[
x_p = (W_p \cdot x_p + b).
\]  

3.3.2 Protein-protein interaction prediction

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

\[
x_{pj} = [x_{p1}; x_{p2}]
\]

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

\[
y = \text{sigmoid}(\text{RELU}(Wx_{pj} + b))
\]

3.4 Integrating inter-molecule interaction into DTA model

After being pre-trained with CCI (Sec. 3.2) and PPI task (Sec. 3.3), the drug encoder \( f(D_s, \theta_s) \) and protein encoder \( f(D_p, \theta_p) \), where \( \theta_s \) and \( \theta_p \) are model parameters, are used to encode the protein and drug for DTA task:

\[
x_p = f(D_p, \theta_p)
\]

\[
x_d = f(D_s, \theta_s)
\]

The protein-drug joint representation is:

\[
x_{pd} = [x_p; x_d]
\]

Finally, the binding affinity is predicted by:

\[
y_u = (W_0 x_{pd} + b_0) W_1 + b_1
\]

4 Experiments

4.1 Dataset

We use the STRING dataset (Szklarczyk et al., 2021) 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, 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 (Kuhn et al., 2008). The dataset contains over 0.5 million chemicals with over 1.6 billion interaction. 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 (Davis et al., 2011) and Kiba (Tang et al., 2014). In the DTA task, we test our proposed method in cold-start settings, including cold-drug and cold-target.

### 4.2 Benchmark

We use four benchmark methods to evaluate the performance of extra-interaction transfer learning on different representations. First, we compare our proposed method with the previous SOTA method GraphDTA (Nguyen et al., 2020). GraphDTA uses CNN as protein encoder and graph neural network as drug encoder. Then the second benchmark method is ESM/DTA which replaces the CNN protein encoder with protein representation pre-trained with protein language model ESM (Rives et al., 2021). The third benchmark is Chemberta/DTA which replaces the graph encoder with SMILES sentences language model representation (Churtnanada et al., 2020). Finally, to evaluate with other graph pre-training strategies, we compare our method with Infograph pre-training method (Sun et al., 2020).

We evaluate the model performance on the test set using Root Mean Squared Error (RMSE), Pearson, Spearman, and Kendall Concordance Index (CI) (Gönen and Heller, 2005).

### 4.3 Implementation detail

Table 2. Hyper-parameters in the experiments.

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

Our methods are implemented using Pytorch. The hyper-parameters are tuned using the validation set. The hyper-parameters detail reported in Table 2. 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. 10). The model is trained with MSE loss using Adam optimizer for 500 epochs. The number of GIN layers (Sec. 3.2.1) $k = 5$.

### 5 Results and Discussion

#### 5.1 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 information learned from PPI task with protein language model such as Chemberta-CCI and ESM-PPI feature across two datasets. This suggests some degree of incompatibility between Chemberta-CCI and ESM-PPI in the cold-target setting.

In general, cooperating the intra-molecule interaction learned from PPI task with protein language model such as Chemberta or ESM 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 interaction learned from PPI task with 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 enhanced models has 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 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 interaction learned from PPI task with protein language model such as ESM benefits the DTA task performance.

#### Table 3. 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 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.

| Drug encoder   | Protein encoder | RMSE  | Pearson | Spearman | CI    |
|----------------|-----------------|-------|---------|----------|-------|
| GIN(Nguyen et al., 2020) | CNN(Nguyen et al., 2020) | 0.6974 | 0.5725  | 0.4838  | 0.7586 |
| GIN | ESM | 0.6753 | 0.5881  | 0.5241  | 0.7805 |
| GIN-CI | ESM-PPI | 0.6793 | 0.595 | 0.4928  | 0.763  |
| Chemberta | ESM-PPI | 0.6672 | 0.6089  | 0.5361  | 0.7873 |
| Chemberta-CCI | ESM-PPI | 0.6651 | 0.6009  | 0.515  | 0.7768 |
| ESM-PPI | Infograph | 0.6697 | 0.5966  | 0.4923  | 0.7632 |
| ESM-PPI | Infograph | 0.6852 | 0.5728  | 0.4621  | 0.7459 |

#### Table 4. 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 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.

| Drug encoder   | Protein encoder | RMSE  | Pearson | Spearman | CI    |
|----------------|-----------------|-------|---------|----------|-------|
| GIN(Nguyen et al., 2020) | CNN(Nguyen et al., 2020) | 1.597 | 0.595  | 0.5949  | 0.7116 |
| GIN | ESM | 1.5385 | 0.6667  | 0.6753  | 0.7472 |
| GIN-CI | ESM-PPI | 1.4320 | 0.6970  | 0.6823  | 0.7499 |
| Chemberta | ESM-PPI | 1.4429 | 0.6946  | 0.681  | 0.7507 |
| Chemberta-CCI | ESM-PPI | 1.3815 | 0.7152  | 0.7071  | 0.7606 |
| ESM-PPI | Infograph | 1.3937 | 0.7177  | 0.7095  | 0.7599 |
| ESM-PPI | Infograph | 1.3739 | 0.7241  | 0.7154  | 0.7634 |
| ESM-PPI | Infograph | 1.3698 | 0.7175  | 0.7051  | 0.7598 |
| ESM-PPI | Infograph | 1.3086 | 0.7137  | 0.7022  | 0.7579 |
| ESM-PPI | Infograph | 1.5029 | 0.6884  | 0.6862  | 0.7528 |
| ESM-PPI | Infograph | 1.4027 | 0.7143  | 0.7094  | 0.7616 |
Mitigating cold start problems

Table 5. 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 model is X and pre-trained with Y task. PPI, CCI, and Infograph are pre-training with PPI, CCI task, or Infograph unsupervised training.

| Protein encoder | Drug encoder | RMSE | Pearson | Spearman | CI |
|-----------------|--------------|------|---------|----------|----|
| CSN((Nguyen et al., 2020)) | GIN | 0.9485 | 0.413 | 0.3998 | 0.6903 |
| GIN | 0.9614 | 0.4919 | 0.4964 | 0.7413 |
| Infograph | GIN | 0.9853 | 0.3579 | 0.4416 | 0.7088 |
| GIN-CI | 0.9755 | 0.5757 | 0.5034 | 0.743 |
| Chemberta | 0.9169 | 0.5174 | 0.3974 | 0.6909 |
| CCI | 0.9146 | 0.5259 | 0.4885 | 0.7171 |
| ESM-PPI | GIN | 0.9637 | 0.4535 | 0.4206 | 0.7007 |
| Infograph | GIN | 0.9489 | 0.4202 | 0.4027 | 0.692 |
| GIN-CI | 0.8841 | 0.5564 | 0.4741 | 0.7299 |
| Chemberta | 0.9032 | 0.4935 | 0.3449 | 0.6645 |
| CCI | 0.9171 | 0.4906 | 0.4216 | 0.7034 |

5.2 Protein-protein interaction knowledge enhances protein language model representation

Looking back to the complex of resistant strain of HIV-1 protease (v82a mutant) with Ritonavir in Sec. 1, we compare the performance of model using only ESM and model with PPI transfer learning and ESM (ESM-PPI). The results in Table 8 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.

Table 8. 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 |

5.3 Integrating different types of CCI improves the DTA prediction model performance

The Chemical-chemical interaction in STITCH dataset (Kuhn et al., 2008) 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 task. The results in Table 9 and 10 show that pre-training with all types of CCI outperforms pretraining 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.
Fig. 8. The T-sne plot of protein embedding of (a) ESM (b) ESM-PPI. Proteins are annotated with druggability which is black text for non-druggable and green text for druggable protein.

Table 9. 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 10. 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  |

6 Conclusions and Future works

In conclusion, migrating the cold-start problem in drug-target affinity 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 protein-protein interaction and chemical-chemical interaction. The protein-protein interaction can provides knowledge regarding protein surface, activity, druggability. The chemical-chemical interaction 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 cold-start problem. In addition, interactions curated from different resources such as text mining are also useful for learning interaction knowledge.

Protein-protein interaction is a complex interaction. 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.

References

Arkin, M. R. et al. (2014). Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. Chemistry & Biology, 21(9), 1102–1114.

Benesty, J. et al. (2009). Noise reduction in speech processing, volume 2. Springer Science & Business Media.

Blundell, T. L. et al. (2000). Protein-protein interactions in receptor activation and intracellular signalling. Biological chemistry, 381(9-10), 955–959.

Blundell, T. L. et al. (2006). Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 361(1467), 413–423.

C. Fry, D. (2012). Small-molecule inhibitors of protein-protein interactions: how to mimic a protein partner. Current pharmaceutical design, 18(30), 4679–4684.
Mitigating cold start problems

Chen, L. et al. (2015). The Use of Chemical-Chemical Interaction and Chemical Structure to Identify New Candidate Chemicals Related to Lung Cancer. PLoS ONE, 10(6), e0128696.

Chithrananda, S. et al. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction.

Cichonska, A. et al. (2017). Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. PLoS Computational Biology, 13(8), e1005678.

Dai, B. and Bailey-Kellogg, C. (2021). Protein interaction interface region prediction by geometric deep learning. Bioinformatics, 37(17), 2580–2588.

Davis, M. I. et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology, 29(11), 1046–1051.

Devlin, J. et al. (2019). 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, pages 4171–4186.

El-Gebali, S. et al. (2019). The Pfam protein families database in 2019. Nucleic acids research, 47(D1), D427–D432.

Elmaghrabi, A. et al. (2021). ProfTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(9), 1–1.

Fry, D. C. (2015). Targeting protein-protein interactions for drug discovery. Methods in molecular biology (Clifton, N.J.), 1278, 93–106.

Gao, M. and Skolnick, J. (2012). The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. Proceedings of the National Academy of Sciences of the United States of America, 109(10), 3784–3789.

Gilson, M. K. et al. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44(D1), D1045–D1053.

Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. Biometrika, 92(4), 965–970.

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. Nature reviews. Drug discovery, 3(9), 727–730.

Hu, W. et al. (2020). Strategies for Pre-training Graph Neural Networks. In The International Conference on Learning Representations.

Jiang, M. et al. (2020). Drug–target affinity prediction using graph neural network and contact maps. RSC Advances, 10(35), 20701–20712.

Jones, S. and Thornton, J. M. (1996). Principles of protein–protein interactions. Proceedings of the National Academy of Sciences, 93(1), 13–20.

Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873, 596(7873), 583–589.

Kim, D. E. et al. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research, 32(Web Server issue), W526.

Kim, S. et al. (2019). PubChem 2019 update: improved access to chemical data. Nucleic Acids Research, 47(D1), D1102–D1109.

Krasowski, A. et al. (2011). DrugPred: A structure-based approach to predict protein druggability developed using an extensive nonredundant data set. Journal of Chemical Information and Modeling, 51(11), 2829–2842.

Kuhn, M. et al. (2008). STITCH: interaction networks of chemicals and proteins. Nucleic Acids Research, 36(Database issue), D664.

Kwon, S. and Yoon, S. (2019). End-to-end representation learning for chemical-chemical interaction prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(5), 1436–1447.

Liu, Y. et al. (2019). RobBERTa: A Robustly Optimized BERT Pretraining Approach.

Nguyen, T. et al. (2020). GraphDTA: Predicting drug-target binding affinity with graph neural networks. Bioinformatics.

Nguyen, T. M. et al. (2021). GEFA: early fusion approach in drug-target affinity prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, pages 1–1.

Ottír, H. et al. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829.

Rao, R. et al. (2019). Evaluating Protein Transfer Learning with TAPE. Advances in Neural Information Processing Systems, 32.

Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America, 118(15).

Sommeth, V. R. et al. (2021). Multi-Scale Representation Learning on Proteins. Advances in Neural Information Processing Systems, 34.

Sun, F.-Y. et al. (2020). InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In The International Conference on Learning Representations.

Suzek, B. E. et al. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics (Oxford, England), 31(6), 926–932.

Szkuczyk, D. et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1), D605–D612.

Tang, J. et al. (2014). Making sense of large-scale kinase inhibitor biaactivity data sets: a comparative and integrative analysis. Journal of chemical information and modeling, 54(3), 733–743.

Thafar, M. et al. (2019). Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. Frontiers in Chemistry, 7.

Winter, R. et al. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chemical Science, 10(6), 1692–1701.

Xu, K. et al. (2018). How Powerful are Graph Neural Networks? 7th International Conference on Learning Representations, ICLR 2019.

Zwillinger, D. and Kokoska, S. (1999). CRC standard probability and statistics tables and formulae. CRC Press.