JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

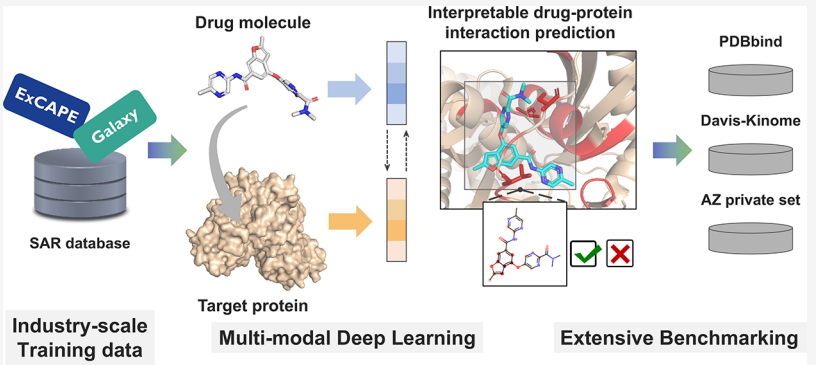
pubs.acs.org/jcim Article

Structure-Aware Multimodal Deep Learning for Drug−Protein Interaction Prediction

Penglei Wang, Shuangjia Zheng, Yize Jiang, Chengtao Li, Junhong Liu, Chang Wen, Atanas Patronov, Dahong Qian,* Hongming Chen,* and Yuedong Yang*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 1308−1317 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Identifying drug−protein interactions (DPIs) is crucial in drug discovery, and a number of machine learning methods have been developed to predict DPIs. Existing methods usually use unrealistic data sets with hidden bias, which will limit the accuracy of virtual screening methods. Meanwhile, most DPI prediction methods pay more attention to molecular representation but lack effective research on protein representation and high-level associations between different instances. To this end, we present the novel structure-aware multimodal deep DPI prediction model, STAMP-DPI, which was trained on a curated industry-scale benchmark data set. We built a high-quality benchmark data set named GalaxyDB for DPI prediction. This industry-scale data set along with an unbiased training procedure resulted in a more robust benchmark study. For informative protein representation, we constructed a structure-aware graph neural network method from the protein sequence by combining predicted contact maps and graph neural networks. Through further integration of structure-based representation and high-level pretrained embeddings for molecules and proteins, our model effectively captures the feature representation of the interactions between them. As a result, STAMP-DPI outperformed state-of-the-art DPI prediction methods by decreasing 7.00% mean square error (MSE) in the Davis data set and improving 8.89% area under the curve (AUC) in the GalaxyDB data set. Moreover, our model is an interpretable model with the transformer-based interaction mechanism, which can accurately reveal the binding sites between molecules and proteins.

INTRODUCTION

The identification of drug−protein interactions (DPIs) lies at the core of *in-silico* drug development. Though experimental assays remain to be the golden standard for determining binding affinities and modes, experimental characterization of every possible drug−protein pair is daunting as there are over 166 billion drug-like compounds[1] and over 5000 potential protein targets.[2] Alternatively, hit compounds could be identified for given protein targets effectively and inexpensively through computational approaches.

Many computational methods have been developed, and these methods could be generally split into two categories: physics-based and machine-learning methods. Physics-based methods like molecular docking apply physics-inspired force fields to simulate the binding of a protein and a molecule at the atomic level and to estimate the binding free energy between them.[3] However, the performance of these methods is often unsatisfactory due to the difficulty in assessing the solvent contributions and conformational entropy. In addition, these physical methods are sensitive to structural fluctuations, which prevents them from dealing well with the flexibility of proteins.[4]

On the other hand, thanks to the rapid increase of protein−ligand binding data and the decrease of computational cost, machine learning-based methods recently gained tremendous progress.[5−7] The general idea for this method is to integrate

Received: January 17, 2022
Published: February 24, 2022

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

structural information from ligands, proteins, and their interactions into a unified framework. In this case, molecules can be characterized by molecular fingerprints, structural descriptors, or topographies, while proteins can be described by sequences or tertiary structures. Their representations are then extracted by the designed neural network to obtain abstract information and are eventually used to predict whether and how they will bind to each other.[5,8−12] Despite a lot of previous efforts, there are a few general caveats among these proposed models:

1. **Using unrealistic data sets with hidden bias.** Although there is a large amount of experimentally reported structure−activity relationship (SAR) data available, data collation and cleaning are quite tedious and laborious. Current deep learning studies either use small data sets, such as the Human and C.elegans data sets[7] which include positive DPI pairs from DrugBank 4.1[13] and Matador,[14] and relatively credible negative samples from a systematic screening framework,[15] or use arbitrary benchmark with expert-defined decoys (i.e., negative samples were generated by fixed rules), including DUD-E, MUV[16] and so on. These data sets unfortunately suffered from obvious chemical biases, therefore overestimating the true accuracy of virtual screening methods. For example, DUD-E was collected with the intention to train structure-based virtual screening with an extremely naïve split according to ligands. As a result, these data sets can be easily separated by ligand information and cannot guarantee that models learn protein information or interaction features. Instead, the key usage of DPI is to identify hit compounds which are unseen for the training set and also nonhomologous to the known actives, such data sets cannot provide fair comparisons of the proposed methods.
2. **Suboptimal representation of protein.** Current works[5,10−12] normally use one-hot encoding vectors to represent residues. Albeit useful, these approaches inherently ignore protein topological information. In fact, the protein topological information is crucial for determining the binding affinity between protein and drug in practice.[17] Although the direct input of 3D structure has been introduced in recent studies,[6,18−21] they have not addressed the issue of 3D transform invariance properly. More recently, several methods[4,22] have shown that the compressed protein structural information like 2D distance map can provide effective signals for DPI prediction. However, the need for crystal structures still limits their application in scenarios where protein crystal structures are not available.
3. **Lack of the high-level associations of instances.** Existing deep learning models mainly focused on the information on the input drug−protein pairs, but weakened the high-level information from protein−protein associations (PPAs) and drug−drug associations (DDAs). The significance of PPAs and DDAs derives from a well-established hypothesis that proteins typically bind with similar drugs,[23] which is key to generalizing DPI predictions. Earlier studies generally considered association by using molecular fingerprinting[24] techniques or BLAST[25] to calculate the similarity of coevolutionary information. However, these approaches were limited in dealing with homologous proteins and had difficulties in dealing with unseen proteins and drugs with novel scaffolds.

To alleviate above problems, we proposed a novel structure-aware multimodal method (coined as STAMP-DPI) for *in-silico* DPI prediction. STAMP-DPI is enabled by the following contributions:

1. We curated a large-scale benchmark GalaxyDB specifically designed for structure-based virtual screening. GalaxyDB was derived from ExcapeDB[26] and consists of 372 common targets with 381 021 confirmed active and 1 634 038 confirmed inactive drug−protein data pairs. The large data size and an unbiased training procedure provide advantages for model building than using a small data set.
2. For informative protein representation, we constructed a structure-aware graph neural network method based on predicted protein contact maps from sequence, which leads to an informative representation of protein and alleviates the inference problem when protein crystal structures are not available.
3. We introduced self-supervised pretrained embeddings for both drugs and proteins in order to strengthen the protein/drug association signals. Our model leverages this high-level information in a unified framework and generates interpretable results with a transformer-based interaction mechanism.

Finally, We provided a comprehensive performance comparison among several state-of-the-art (SOTA) methods. Our results demonstrated that STAMP-DPI has superior performance over these models by two benchmark data sets. More importantly, the model was also proven by prospective predictions on the external data set extracted from the AstraZeneca screening database, a real world industry data set containing 208 958 data points.

METHODS

Data Set Construction. We constructed experiments using the following two benchmark data sets for model building and evaluation. In addition, an external test data set from AstraZeneca was utilized to verify the generalization capability of our model on the industrial data set.

(a) The Davis data set consists of binding affinity information with K_D (dissociation constant) values among 72 drugs and 442 targets. In our experiments, we used SMILES representation of 68 drugs and sequence representation of 442 target proteins from the DeepDTA[11] training/test data set. For the Davis data set, we viewed the DPI prediction task as a regression task that predicts the K_D values for each DPI pair. This small data set is used to initially verify that our model can effectively deal with the DPI prediction problem. When we used the Davis data set for prospective validation, we assigned the data points in the Davis data set to two class according to the criterion of $K_D \geq 6$ and viewed the DPI prediction task as a classification task.

(b) GalaxyDB. We curated a large-scale DPI benchmark, GalaxyDB, based on the ExCAPE-ML,[27] a collection of protein−ligand entries complied from ExCAPE-DB. ExCAPE-ML contained 955 386 compounds, covering 526 distinct target proteins for a total of 49 316 517

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

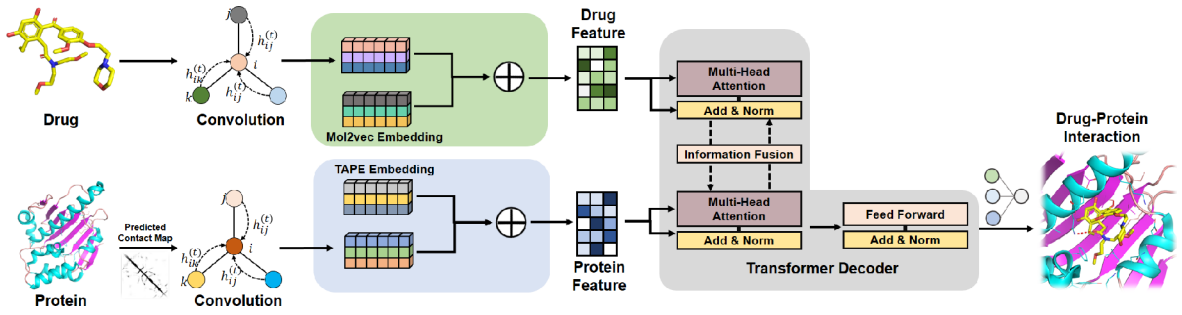


Figure 1. Architecture of STAMP-DPI. It first processes the molecule and protein features in parallel and, then, fuses the embedding of molecule and protein by Interaction Decoder for the DPI prediction.

structure−activity relationship (SAR) data points. For classification tasks, the data points were assigned to two classes (i.e., inactive, active) according to their log-transformed activity values (pXC_{50} values). A compound−target record was defined to be activated if it fulfilled the criterion of $pXC_{50} \geq 6$ (activity ≤ 1 μM). Since the ExCAPE-ML data set contains a large number of (more than 45 million) data points with a pXC_{50} value of 3.101 that are expert-defined negative samples with low confidence, we excluded these data points to form a relatively balanced and high confident benchmark set. Subsequently, the data set was trimmed down by removing target proteins with a sequence length longer than 750 in order to reduce the computational cost during the calculation of the contact map and coevolution features for proteins and the processing of protein features. Finally, a benchmark data set containing 2 015 059 DPI data points, corresponding to 632 459 compounds and 372 distinct target proteins was constructed. We selected this benchmark data set for the training and evaluation of our proposed model.

(c) External test data set. To test whether the model is capable of performing real-world virtual screening tasks, we have made prospective prediction with AstraZeneca in-house SAR data. In particular, for targets seen in our train set, we selected the top 30 targets according to the performance of our model in the GalaxyDB data set and required that each target has at least more than 100 data points. Additionally, we also randomly selected 10 targets that are not included in our training set. In total, we constructed an external test set which is composed of 208 958 data points, including 172 768 data points for 30 targets seen in the training set and 36 190 data points for 10 unseen targets.

Representations of Protein and Molecule. The representations of protein and molecule lie at the core of the DPI task. In this section, we described the initial feature representations of target proteins, followed by the feature representations of molecules.

Protein Representation. The protein was represented from the perspectives of structure and sequence features, respectively. For the structural features, we used a graph to represent spatial relations between residues, which has been proven effective for predicting protein solubility in our previous study.[28] In this model, residues were regarded as nodes and the predicted contact map from sequence was used as the adjacency matrix. Node features were represented by the Hidden Markov Matrix (HMM), position-specific scoring matrix (PSSM), and structural features predicted from SPIDER3.[29] The PSSM and HMM features are evolutionary information that contains the motifs related to protein properties in protein sequences,[30] where the PSSM profile was generated by PSI-BLAST v2.7.1[25] with the UniRef90 sequence database after three iterations, and the HMM profile was generated by HHBLITS v3.0.3 in aligning the UniClust30 profile HMM database[31] with default parameters. The structural features include 14 features to reflect the secondary structure of proteins predicted by SPIDER3. The list of protein node features can be found in the Supporting Information Table S1. For the contact map of proteins, we made the protein contact map by SPOT-Contact,[32] which takes the protein sequence-based and evolutionary coupling-based information as input to predict the contact probability of all residue pairs in one protein. Finally, we obtained a protein graph as $G_\alpha = (V_\alpha, A_\alpha)$, where $V_\alpha \in R^{n\times f}$ is the set of n amino acid nodes, each node represented by f-dimension features vector composed of HMM, PSSM, and structural features, $A_\alpha \in R^{n\times n}$ is the adjacency matrix (contact map) for the protein graph.

For the protein sequence feature, we also considered using the high-level representation learned from a large collection of unlabeled protein sequences provide by TAPE.[33] TAPE is a language model for protein representation, and it encodes each amino acid into an embedding vector. For each embedding vector, it is contextual and includes the sequence information from the input protein sequence, so we embedded the protein sequence to tape embedding with the pretrained BERT[34] model in TAPE.

Drug Molecular Representation. We represented the drug molecule as a graph to get more accurate structure information for the molecule. In this sense, a molecular graph can be formulated as $G_c = (V_c, A_c)$, where $V_c \in R^{n\times f}$ is the set of n atom nodes with each node represented by f-dimension features vector composed of atomic properties. Here we used f-dimension atomic features that are detailed in the Supporting Information Table S2. $A_c \in R^{n\times n}$ is the set of edges represented by the adjacency matrix for the molecular graph. The existence of edges in the adjacency matrix depends on whether the corresponding atoms in the molecule directly have a covalent chemical bond. Besides, we also used mol2vec[35] features at graph level as a high-level representation for a molecule to capture the DDAs.

We believe that the additional high-level representation from pretrained embedding for proteins and molecules could provide implicit information to make the model distinguish different proteins and molecules. The high-level representation provides the global similarity information for DPI models, which describes the protein−protein association and drug−

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

drug associations (PPAs and DDAs). The DPI models could leverage the global similarity to measure the associations between the seen and unseen proteins and molecules and make full use of the features of existing data to improve the performance of DPI prediction.

Model Architecture of STAMP-DPI. The overview framework of our proposed STAMP-DPI network is shown in Figure 1. The input information includes multilevel representation for proteins and ligands. As shown in Figure 1, our model consists of three main modules: a graph representation network for proteins (Protein GNN Encoder), a graph representation network for molecules (Molecular GNN Encoder), and an interactive network with a transformer decoder for message interaction (Interaction Decoder).

Protein GNN Encoder. In our model, graph representation of proteins and the pretrained feature encode by TAPE were input to the Protein GCN Encoder to learn structure and sequence representations of proteins at the same time. The Protein GCN Encoder includes two aspects: the first is the GCN encoder which encodes the structure information on protein graphs. The second is an information fusion unit to fuse the embedding information from the GCN encoder and the high-level representation from the pretrained model. The protein graph is represented as a combination for the contact map and node features as input for the GCN encoder, then the GCN encoder learns node-level outputs for the protein graph. The GCN can be used to effectively process the graph structure data. The propagation rule can be represented in the normalized form as eq 1:

$$H^{(l+1)} = \sigma(\tilde{D}^{1/2}\tilde{A}\tilde{D}^{-1/2}H^{l}W^{l}) \tag{1}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix of the graph with added self-connection and $\tilde{D} = \sum_i A_{ii}$ is the diagonal node degree matrix. W^l and H^l are the learnable parameters in GCN and the output of lth layer, respectively. $\sigma(\cdot)$ is an activation function such as ReLU. For protein graph, $H^0 = V_a$, $A = A_a$. To further extract the high-level features for protein, we used CNN with Conv1D and gated linear unit (GLU) to fuse the different node embedding. In addition to the structure information from the protein graph, we also used dense layers to encode the extra protein sequence embedding information generated from the TAPE model. Finally, the protein graph and sequence information were concatenated to form the protein feature for the following Interaction Decoder module.

Molecular GNN Encoder. Similarly, the GCN was used to encode the molecular graphs. In particular, we used the isolated GCN but similar architecture as in the Protein GNN Encoder to learn the node-level features for molecules and obtain the molecule structure embedding. Considering the different sizes in molecular graphs, we padded the graphs to keep the consistent size for different molecules and masked the padded information in our model. On the other hand, the dense layers were used to encode the high-level representation information from the mol2vec embedding, then the structure and mol2vec embedding information were concatenated to form the molecule feature for the following Interaction Decoder module.

Interaction Decoder. This module is inspired by the TransformerCPI,[22] which provides a method to fuse the embedding features of molecule and protein. A transformer[36] decoder was leveraged in our Interaction Decoder module to combine the information on proteins and molecules. The Interaction Decoder here served as a fusion unit to capture features useful for the interaction between molecule and protein. The decoder mainly consists of a multihead self-attention layer and feed-forward layer. The multihead self-attention layer employed the multiple self-attention mechanisms to extract interaction information and it can be represented as

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_0, \mathrm{head}_1, ..., \mathrm{head}_n) \tag{2}$$

where the Q, K, and V are the queries, keys, and values in the Transformer. The term $head_i$ is the output of the ith self-attention layer, $W \in R^{nd_v \times nd_m}$ is a learnable parameter for fusing attention information from different heads. n and d_m is the number of heads and the dimension of the hidden state, respectively. The self-attention in each head calculates the attentions by

$$\mathrm{attention}(QW^Q, KW^K, VW^V) = \mathrm{softmax}\left(\frac{QW^Q(KW^K)^T}{\sqrt{d_k}}\right)VW^V \tag{3}$$

where the projection matrices $W^K \in R^{d_m \times d_k}$, $W^K \in R^{d_m \times d_k}$, $W^V \in R^{d_m \times d_v}$ are learnable parameters. Compared with previous methods of directly concatenating the protein and molecule embedding information, we believe that this architecture can more effectively capture the interaction between protein and molecule embedding. Finally, we obtained the interaction features between protein and molecule and we could calculate the molecular interaction with a protein as follows:

$$h = \sum_i^n \mathrm{softmax}(\|X\|_2^2)x_i \tag{4}$$

where X is the output matrix of transformer decoder and composed of a set of interaction vectors x_1, x_2, ..., x_n. The $\|X\|_2^2$ represent the l_2 norm for x_i in interaction matrix X. Here n is the number of a set of interaction vectors from the Interaction Decoder. Finally, the interaction feature h is fed into a fully connected layer and a sigmoid function and obtains the predicted interaction probability $p(\tilde{y})$ between protein and molecule. The model would be trained by maximizing the likelihood of regressing the training data, which means minimizing the binary cross-entropy loss as follows:

$$\arg\min_\theta -(\tilde{y}\cdot\log(p(\tilde{y})) + (1-\tilde{y})\cdot\log(1-p(\tilde{y}))) \tag{5}$$

where θ are the learnable parameters of the model.

Model Training and Evaluation. Our model took protein graphs, protein evolutionary and predicted structural features, molecule graphs, and molecular substructure features as input, where we converted the SMILES representation for the molecule to graph representation through RDKit.[37] The model was implemented in Pytorch and trained on RTX 2080Ti. And the training details for our model as follows: The hidden state size d_m are set to 64 and 256 for molecule embedding and protein embedding, respectively. The number of graph convolution iterations is set to 3, and the kernel size in CNN for protein embedding is 7. For Interaction Decoder, the number of decoder layers is set to 3, and the number of heads in the multihead layer is set to 8. We tuned these hyperparameters by grid searching, and the value list is shown in Supporting Information Table S4. Apart from all the hyperparameters mentioned above, the maximum number of

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

epochs during the training process in our model is set to 50 and the batch size equals 32 in every epoch. For the classification task, the training process will early stop when the performance no longer improves after 5 epochs in the validation data set. For the regression task on Davis data set, we adopted 5-fold cross-validation to evaluate the model performance. The data were randomly divided into 5-fold according to protein targets, and then one fold was used as the test set and the rest as the training set. The model was trained with fixed 100 epoch, and finally we used the mean of metrics on 5-fold data set as the evaluation results. Dropout is applied in CNN and the Interaction Decoder module, the dropout rate is set to 0.2. For the optimizer in our model, we use the LookAhead[38] optimizer combined with RAdam[39] optimizer, in which the learning rate is set to 1×10^{-4} and weight decay is set to 1×10^{-4}.

In order to evaluate the performance of our model, we divided the Davis and GalaxyDB data sets to obtain the training, validation and test sets, respectively. In particular, we divided the GalaxyDB data set according to the proteins so that the target proteins in validation and test set were not seen in the training set. For the Davis data set, the targets which have the same sequence in the training, validation, and test sets were filtered out which results in 361 targets. Table 1 summarizes the split data set in detail.

Table 1. Detailed Information for the Split Dataset (Davis is a regression dataset)

type	proteins	pos	neg	pairs
		Davis		
train	231	–	–	15708
valid	57	–	–	3876
test	73	–	–	4964
		GalaxyDB		
train	298	305702	1295867	1601569
valid	38	43825	197666	241491
test	36	31494	140505	171999

For the regression task on the Davis data set, the performances of models are evaluated using root mean square error (RMSE), mean square error (MSE), and Pearson and Spearman metrics. The main metric for evaluating the prediction performance in the classification task is the area under the receiver operation characteristic curve (ROC-AUC), which can reflect the ability of the model to correctly discriminate the active compounds and inactive compounds. And, the AUC metric is also the condition for early stopping during model training. Additionally, we measure the accuracy, recall, precision, and F1 score metrics for evaluating the performance of the model prediction. It is worth noting that we determined the threshold for the above four metrics in the test set by finding the best threshold in the validation set. We set the search threshold in the range of 0.0−0.9 and search with 0.001 steps to find the best threshold in the validation set according to the F1 score.

We compared our model with the following baselines:

1. SGDRegressor is a linear model fitted by minimizing a regularized empirical loss with Stochastic Gradient Descent (SGD). We used it for Davis data set in the regression task. We experimented on the concatenated molecule and protein features. Here the molecule feature is Morgan Fingerprint calculated by RDKit,[37] and the protein feature is the average tape embedding which suggests that taking the mean values at the amino acid level for original tape embedding.
2. L2-logistic regression (LR) applied a logistic regression model on the Morgan Fingerprint and tape embedding concatenated feature vectors, we used it for our GalaxyDB data set in the classification task.
3. TransformerCPI[8] modified the transformer architecture with a self-attention mechanism to address sequence-based DPI classification task, we followed the default parameter settings in TransformerCPI and the same training and evaluating strategies as STAMP-DPI.
4. GraphDTA[10] represented molecules as graphs and uses graph neural networks to predict drug-target affinity. Here we compared our model with the GIN[40] in GraphDTA with default parameters. Besides, in order to fit the binary classification task on GalaxyDB data set, we added a sigmoid function for the last layer in the GraphDTA network.
5. MolTrans[9] is an end-to-end biological-inspired deep learning-based framework that models the DPI process. We followed the same hyper-parameter setting described in the paper and compared our model with the MolTrans on our data set.

RESULTS AND DISCUSSION

Performance on the Davis and GalaxyDB Data Sets. In order to validate the effectiveness of our model, we first tested our model on a well-defined small data set, Davis. As shown in Table 2, our model obtained the best MSE with 0.4317, which is 7.00% lower than the TransformerCPI (0.4642) model, the best performance baseline model. Interestingly, we found that the performance of the complex deep learning models on the Davis data set is not significantly better than other traditional machine learning models, and the strong learning ability of the deep learning model could not be well reflected on the Davis data set. This is likely because the Davis data set consists only of the kinase protein family with relatively small amounts of data. We also compared the performance of models on the PDBbind v2016 data set[41] in Supporting Information Table S3 and observed that our model achieved the best performance to baseline models, which is

Table 2. Cross Validation Performance Comparisons of STAMP-DPI and Baseline Models on Davis Dataset

model	RMSE (Std)	MSE (Std)	Pearson (Std)	Spearman (Std)
SGDRegressor	0.7208 (0.0187)	0.5199 (0.0274)	0.5101 (0.0073)	0.4686 (0.0100)
GraphDTA	0.7409 (0.0149)	0.5492 (0.0221)	0.4981 (0.0172)	0.4087 (0.0185)
MolTrans	0.7688 (0.1237)	0.6063 (0.2127)	0.4591 (0.1006)	0.4215 (0.0321)
TranformerCPI	0.6805 (0.0333)	0.4642 (0.0443)	0.5809 (0.0292)	0.4542 (0.0201)
STAMP-DPI	**0.6569** (0.0148)	**0.4317** (0.0192)	**0.6322** (0.0319)	**0.5113** (0.0326)

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

consistent with the Davis data set. We used these benchmark data sets to tune the model architecture.

We further compared the performance of different methods on the industry-scale large data set, GalaxyDB. As shown in Table 3, the proposed model achieved the best performance in terms of both AUC and F1, two balanced metrics. Specifically, our model achieved an AUC of 0.8011, which is 8.89% higher than the best baseline MolTrans (0.7357) and 12.21% higher than the TransformerCPI (0.7139). Our model has the highest precision and the lowest but reasonable sensitivity. It should be noted that the thresholds of all methods were optimized for F1 values according to the validation set, and the final threshold corresponding to LR, GraphDTA, MolTrans, TransformerCPI, and STAMP-DPI are 0.191, 0.438, 0.360, 0.523, and 0.449, respectively. Figure 2 compared all methods by the receiver operating characteristic (ROC) and precision-recall (PR) curves on the GalaxyDB test set. We could see that our model obtained consistent results and achieved superior performance over the other baseline models. This followed our expectations, as the graph representation for proteins provides richer structure information than sequence representation, our model could provide more abundant information for molecules and proteins by combing the structure and sequence information from pretrained embedding features, which could effectively improve the performance of the DPI prediction. As shown in Supporting Information Figure S3, we plotted the sequence similarity information between the test set and the training set, which was calculated by the Blast tool.[42] We divided the test proteins into those with low similarity to the training set (similarity <0.4) and those with high similarity (similarity ≥0.4). We also plotted the AUC distribution as shown in Supporting Information Figure S2c and calculated the average AUC for low similarity proteins (average AUC: 0.6070) and high similarity proteins (average AUC: 0.6730) in the test set. This results suggest that the high similarity proteins achieve better performance than low similarity proteins. We further plotted the AUC of proteins sorted by the active ligand number for our model in Figure Supporting Information S2d and observed that our method could work well in low data but high similarity targets.

Table 3. Performance Comparisons of STAMP-DPI and Baseline Models on the GalaxyDB Dataset[a]

model	AUC	precision	recall	F1
LR	0.6422	0.1813	**0.8304**	0.2977
GraphDTA	0.7136	0.2580	0.7122	0.3788
MolTrans	0.7357	0.3683	0.6248	0.4634
TranformerCPI	0.7139	0.3268	0.5989	0.4228
STAMP-DPI	**0.8011**	**0.5097**	0.5777	**0.5415**

[a]The precision, recall, and F1 score are calculated with the best threshold for each model.

Ablation Experiments. In order to validate the contribution of each component in our model, we performed ablation experiments by removing coevolution features, predicted contact map, or pretrained embedding features.

First, we evaluated the function of additional pretrained embedding, which includes the sequence information with tape embedding for proteins and substructure information for molecules. Instead of fusing the structure information from the protein graph and sequence information from tape embedding, we only used the structure information of proteins as the input features of the Interaction Decoder. We also used the structure information of molecules only in Interaction Decoder. As shown in Table 4, the removal of pretrained embedding decreases the prediction performance of the model significantly. And this ablation experiment clearly shows the importance of pretrained embedding in the model, which provides high-level protein sequence information and molecular substructure information for model learning.

Second, we evaluated the importance of structure information for protein in our model. In our model, the structure information of protein mainly comes from the graph representation for protein, which includes the amino acid node features HMM/PSSM/structural features and the contact map of protein. In order to comprehensively evaluate the influence of protein structure information on model performance, we conducted ablation experiments on protein node features and contact maps, respectively. For the ablation experiments of the contact map, we avoided utilizing the protein structure information from the contact map and only used CNN to extract and fuse the node features in protein. The results represented in Table 4 suggest that the contact map processed by GCN could provide efficient and rich structure information for proteins, which results in better performance in the DPI prediction task. However, due to the structure of proteins is extremely complex that the predicted contact map cannot accurately reflect the structure of proteins, we still need to introduce additional information such as evolutionary and predicted structural features to compensate for the information loss of the predicted contact map. And the ablation

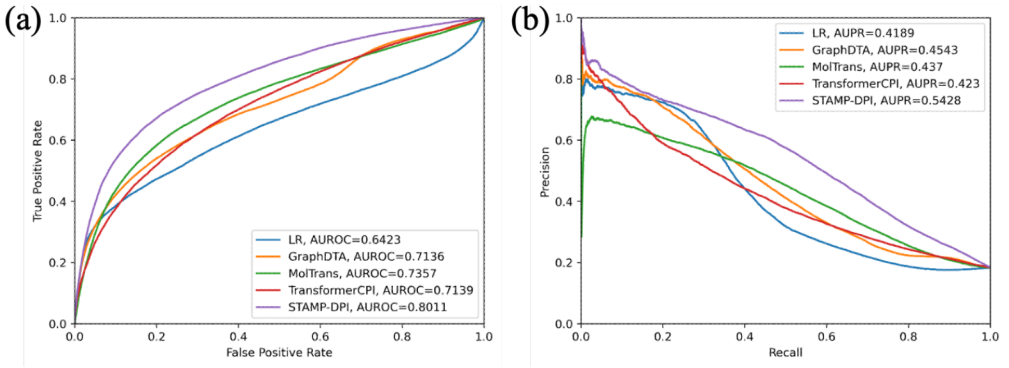


Figure 2. Performance of different methods on the GalaxyDB test set. (a) Receiver operating characteristic (ROC) curves of prediction results. (b) Precision−recall (PR) curves of prediction results.

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

Table 4. Results of Ablation Experiments

network	HMM/PSSM/structure	contact map	pretrained features	AUC	ACC	precision	recall	F1	best threshold
STAMP-DPI	×	×	×	0.7139	0.7006	0.3268	0.5989	0.4228	0.412
	√	×	×	0.7593	0.7448	0.3874	**0.6771**	0.4929	0.329
	×	√	×	0.7282	0.6893	0.3277	0.6623	0.4384	0.392
	×	×	√	0.7590	**0.8451**	**0.6057**	0.4419	0.5110	0.598
	×	√	√	**0.7947**	0.7000	0.3534	**0.7697**	0.4484	0.319
	√	×	√	0.7649	0.8059	0.4766	0.6114	0.5356	0.311
	√	√	×	0.7832	0.8178	0.5021	0.6531	**0.5677**	0.234
	√	√	√	**0.8011**	**0.8209**	**0.5097**	0.5577	**0.5415**	0.449

experiments have also shown that our model could significantly improve the performance by combing the contact map and other additional protein information. To validate the function of protein node features in our model for DPI prediction, we used word2vec embedding in TransformerCPI to replace the protein node features as HMM, PSSM and structural features. We observed that when we used the word2vec embedding as the protein node feature, the performance has slightly decreased. This suggests that the HMM, PSSM, and structural features used in our model can provide better protein information at the amino acid level compared to word2vec, which is beneficial to the accurate prediction of the DPI prediction task.

Performance on the Prospective Validation. In order to verify the generalization ability of our model, we used the model trained on the GalaxyDB set to evaluate the Davis data set and external test data set from AstraZeneca.

For the Davis data set in prospective validation, the experimental results were shown in Figure 3. In general, our

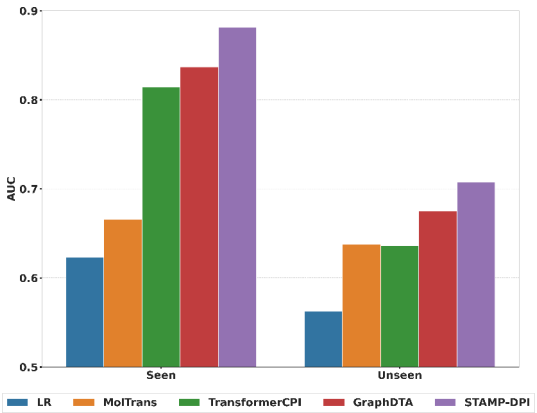


Figure 3. Performance comparisons of STAMP-DPI and baseline models on seen and unseen protein targets on Davis data set.

model consistently performed well on the test set. When the tested proteins were observed in the training set, STAMP-DPI achieved an AUC of 0.8813, which is 5.28% higher than the best baseline GraphDTA (0.8371) and 8.23% higher than the TransformerCPI (0.8143). For the unseen proteins, our model also achieved the best AUC with 0.7073, which is 4.77% higher than GraphDTA (0.6751) and 11.21% higher than TransformerCPI (0.6360).

For the external test data set in prospective validation, the data distribution and classification performance of our model were given in Figure 4. As shown in Figure 4, for the seen proteins, the average AUC on a total of 30 targets is 0.7724, and 70% of the seen targets reached AUC ≥ 0.7. For the unseen proteins, the average AUC on a total of 10 targets is 0.7264 and 50% of the seen targets reached AUC ≥ 0.7. We further plotted the protein sequence similarity information between external data set and the training set as shown in Supporting Information Figure S4, we found that the protein targets usually have poor performance when the data points are insufficient and the similarities to training proteins are low.

In general, our model achieved reasonable performance on both seen and unseen proteins, indicating that the STAMP-DPI trained on GalaxyDB generalizes well to independent virtual screening tasks. However, performance gaps between seen and unseen proteins were observed both on the Davis data set and external test data set. We argue that there might be two potential reasons for these performance gaps. The first one is that the chemical space for seen and unseen target proteins is different, which makes the knowledge of chemical spatial distribution for seen proteins learned by our model unable to be effectively applied to unknown chemical space for unseen proteins. In order to validate this point, we further utilized the t-SNE method[43] to visualize the distribution of protein targets on Davis data set, which based on the tape embedding, coevolution, and predicted structural features. As shown in Figure S5, there are some differences in the main distribution regions of data points between seen and unseen proteins, which suggests the differences in chemical space between them. The second is that the ability of our model to learn unseen protein representations is still somewhat deficient. The predicted contact maps and protein pretrained embeddings have helped us to improve our predictions for unseen proteins, but there is still much room for improvement.

Model Interpretability. Benefiting from the Interaction Decoder architecture module, our model is able to analyze the interaction mechanism between the protein and molecule. The positions focused on the self-attention mechanism can provide a reasonable explanation for the binding activity prediction and also help to quickly locate the key interaction sites between the protein and molecule when performing further activity analysis.

To exemplify this, we selected two complexes from the RCSB Protein Data Bank (PDB)[44] as the representatives, where the proteins were presented in the test of GalaxyDB. We took the attention weight calculated with the molecular feature as the Query and the protein feature as the Key in the last decoder layer of the Interaction Decoder and then calculated the mean of attention weight at the molecular dimension to obtain the attention information on proteins. In particular, we colored the top-weighted residues of the example proteins and atoms of the ligand with red and compared them to the actual protein−ligand interaction sites retrieved from the PDB. We found that the highest-weighted amino acids and molecular atoms overlap substantially with the real interaction sites. For protein CA13 (UniProt ID: Q8N1Q1) in Figure 5a, the attention bar highlights residues His95, Thr200, and Trp210,

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

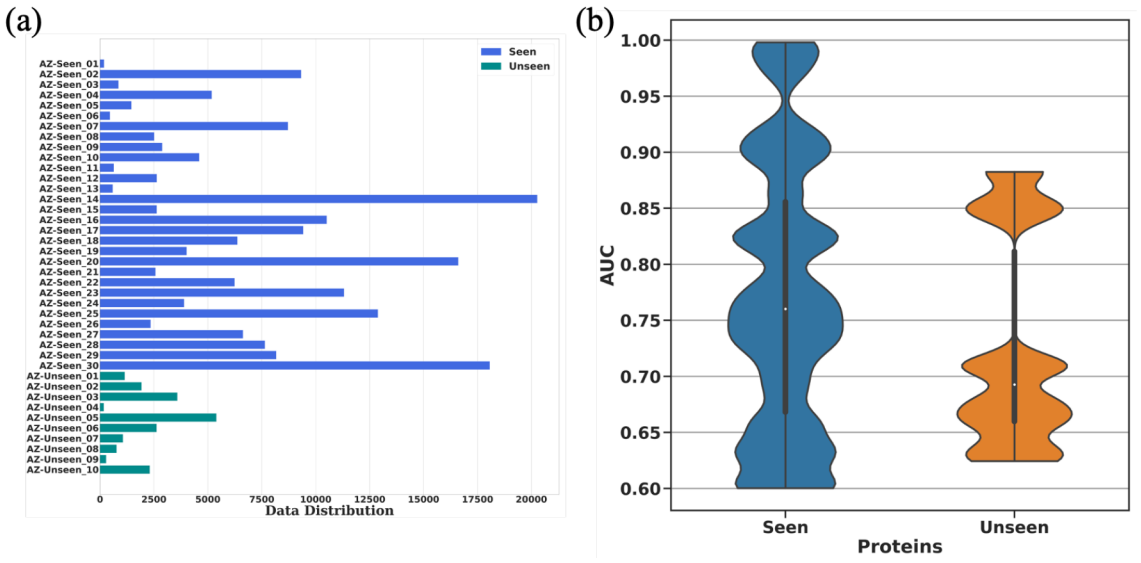


Figure 4. Information and evaluation results about the external test data set. (a) Data point distribution of individual target proteins. (b) AUC performance of our model on an external test data set. The violin plot represents the AUC distribution of individual target protein performance for seen and unseen proteins.

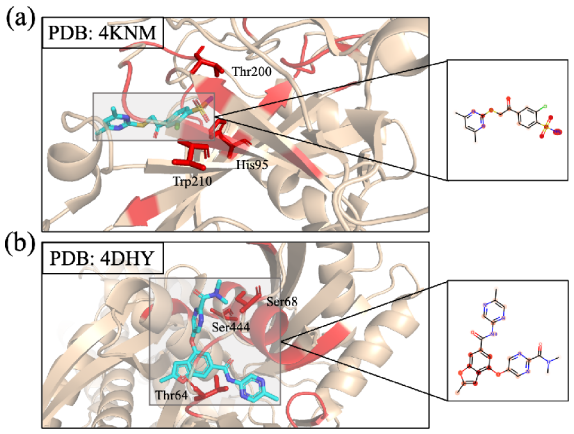


Figure 5. Attention weight visualization of pocket and ligand pairs. (a) Attention weight of interaction for CA13 and E1E (PDB: 4KNM). (b) Attention weight of interaction for GCK and S41 (PDB: 4DHY).

which highly overlap with the key pocket residues observed in the cocrystal complex (PDB: 4KNM). For protein GCK (UniProt ID: P35557) in Figure 5b, the highlighted residues (Thr64, Ser68, Ser444) and ligand functional groups in the importance maps show high similarity to observed interactions in the cocrystal complex (PDB: 4DHY). The results suggest that the model can be applied to analyze the interaction mechanism between molecules and target proteins and inspire researchers.

CONCLUSION

In this study, we selected the Davis and GalaxyDB data set as the internal validation data set for our model, meanwhile, we further verified the generalization ability of our model on the external test set collected from AstraZeneca. Experimental evaluations show that our model consistently has the best performance on these three data sets. The ablation experiments have shown that the protein graph constructed by the contact map and amino acid node can provide richer and more accurate structure information for DPI prediction, and we can obtain better performance for DPI prediction when we combine high-level pretrained information from the proteins and molecules. Overall, we believe that our study provides a new SOTA model for DPI prediction research. Additionally, the benchmark data set that we constructed can be used for the community to develop and evaluate future structure-based virtual screening models. Combining the structure and pretrained information for both protein and ligand provides advantages in making DPI prediction and could be a new area to explore in the future.

DATA AND SOFTWARE AVAILABILITY

The data sets and source code are available on https://github.com/biomed-AI/STAMP-DPI.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00060.

Tables S1−S4, Figures S1−S5, and associated refs (PDF)

AUTHOR INFORMATION

Corresponding Authors

Dahong Qian − *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*; Email: dahong.qian@sjtu.edu.cn

Hongming Chen − *Guangzhou Laboratory, Guangzhou 510000, China*; orcid.org/0000-0002-8065-8333; Email: chen_hongming@grmh-gdl.cn

Yuedong Yang − *School of Data and Computer Science, Sun Yat-Sen Universit, Guangzhou 510275, China*; orcid.org/0000-0002-6782-2813; Email: yangyd25@mail.sysu.edu.cn

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317

Authors

Penglei Wang − *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0002-1966-3491

Shuangjia Zheng − *School of Data and Computer Science, Sun Yat-Sen Universit, Guangzhou 510275, China;* orcid.org/0000-0001-9747-4285

Yize Jiang − *Galixir, Beijing 100080, China*

Chengtao Li − *Galixir, Beijing 100080, China*

Junhong Liu − *Galixir, Beijing 100080, China*

Chang Wen − *Guangzhou Laboratory, Guangzhou 510000, China*

Atanas Patronov − *MolecularAI, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Gothenburg 405 30, Sweden*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00060

Author Contributions

○P.W. and S.Z. contributed equally to this work.

Funding

This study has been supported by the National Key R&D Program of China (2020YFB0204803), National Natural Science Foundation of China (61772566), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), and Guangzhou S&T Research Plan (202007030010).

Notes

The authors declare the following competing financial interest(s): This work is done when P.W. worked as an intern at Galixir; S.Z., Y.J., C.L., and J.L. were employees of Galixir; and A.P. was an employee of AstraZeneca.

■ ACKNOWLEDGMENTS

We thank the Galixir team for its support and discussion, with special thanks to Jixian Zhang, Zixuan Liu, and Da Wei for the experimental design discussion and technical support.

■ REFERENCES

(1) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(2) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **2016**, *44*, 1045−1053.

(3) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.

(4) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting drug−protein interaction using quasi-visual question answering system. *Nature Machine Intelligence* **2020**, *2*, 134−140.

(5) Gao, K. Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. Interpretable Drug Target Prediction Using Deep Neural Representation. In *IJCAI*; 2018; pp 3371−3377.

(6) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.

(7) Tsubaki, M.; Tomii, K.; Sese, J. Compound−protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **2019**, *35*, 309−318.

(8) Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. TransformerCPI: improving compound−protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **2020**, *36*, 4406−4414.

(9) Huang, K.; Xiao, C.; Glass, L. M.; Sun, J. MolTrans: Molecular Interaction Transformer for drug−target interaction prediction. *Bioinformatics* **2021**, *37*, 830−836.

(10) Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; Venkatesh, S. GraphDTA: Predicting drug−target binding affinity with graph neural networks. *Bioinformatics* **2021**, *37*, 1140−1147.

(11) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug−target binding affinity prediction. *Bioinformatics* **2018**, *34*, 821−829.

(12) Öztürk, H.; Ozkirimli, E.; Özgür, A. WideDTA: prediction of drug-target binding affinity. *arXiv.org*; 2019; 1902.04166.

(13) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* **2008**, *36*, 901−906.

(14) Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E. G.; Gewiess, A.; Jensen, L. J.; Schneider, R.; Skoblo, R.; Russell, R. B.; Bourne, P. E.; Bork, P.; Preissner, R. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **2007**, *36*, 919−922.

(15) Liu, H.; Sun, J.; Guan, J.; Zheng, J.; Zhou, S. Improving compound−protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **2015**, *31*, 221−229.

(16) Sieg, J.; Flachsenberg, F.; Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.

(17) Litfin, T.; Zhou, Y.; Yang, Y. SPOT-ligand 2: improving structure-based virtual screening by binding-homology search on an expanded structural template library. *Bioinformatics* **2017**, *33*, 1238−1240.

(18) Wee, J.; Xia, K. Forman persistent Ricci curvature (FPRC)-based machine learning models for protein−ligand binding affinity prediction. *Brief. Bioinf.* **2021**, *22*, 1.

(19) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology* **2018**, *14*, No. e1005929.

(20) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein−ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666−3674.

(21) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv.org*; 2015; 1510.02855.

(22) Jiang, M.; Li, Z.; Zhang, S.; Wang, S.; Wang, X.; Yuan, Q.; Wei, Z. Drug−target affinity prediction using graph neural network and contact maps. *RSC Adv.* **2020**, *10*, 20701−20712.

(23) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry* **2002**, *45*, 4350−4358.

(24) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(25) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997**, *25*, 3389−3402.

(26) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminf.* **2017**, *9*, 1−9.

(27) Sturm, N.; Mayr, A.; Le Van, T.; Chupakhin, V.; Ceulemans, H.; Wegner, J.; Golib-Dzib, J.-F.; Jeliazkova, N.; Vandriessche, Y.; Böhm, S.; Cima, V.; Martinovic, J.; Greene, N.; Vander Aa, T.; Ashby, T. J.; Hochreiter, S.; Engkvist, O.; Klambauer, G.; Chen, H. Industry-scale application and evaluation of deep learning for drug target prediction. *J. Cheminf.* **2020**, *12*, 1−13.

(28) Chen, J.; Zheng, S.; Zhao, H.; Yang, Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J. Cheminf.* **2021**, *13*, 1−10.
(29) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **2017**, *33*, 2842−2849.
(30) Habibi, N.; Hashim, S. Z. M.; Norouzi, A.; Samian, M. R. A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC bioinformatics* **2014**, *15*, 1−16.
(31) Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **2017**, *45*, 170−176.
(32) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* **2018**, *34*, 4039−4045.
(33) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems 32*, 2019; Vol. *32*, p 9689.
(34) Kenton, J. D. M.-W. C.; Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019; pp 4171−4186.
(35) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.
(36) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017; pp 5998−6008.
(37) G, L. *RDKit: Open-source cheminformatics*; http://www.rdkit.org (accessed 8 Aug, 2017).
(38) Zhang, M.; Lucas, J.; Ba, J.; Hinton, G. E. Lookahead Optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, 2019.
(39) Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the variance of the adaptive learning rate and beyond. *arXiv.org*; 2019; 1908.03265.
(40) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, 2019.
(41) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry* **2004**, *47*, 2977−2980.
(42) Ye, J.; McGinnis, S.; Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic acids Res,* **2006**, *34*, 6−9.
(43) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Machine Learn. Res.* **2008**, *9*, 2579−2605.
(44) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Kalro, T.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Periskova, I.; Prlic, A.; Randle, C.; Rose, A.; Rose, P.; Sala, R.; Sekharan, M.; Shao, C.; Tan, L.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Woo, J.; Yang, H.; Young, J.; Zhuravleva, M.; Zardecki, C. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **2019**, *47*, 464−474.

https://doi.org/10.1021/acs.jcim.2c00060
J. Chem. Inf. Model. 2022, 62, 1308−1317