Interpretable Prediction of Protein-Ligand Interaction by Convolutional Neural Network

Fan Hu[†]
Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, China
fan.hu@siat.ac.cn

Jiaxin Jiang[†]
School of Software Engineering, University of Science and Technology of China
Hefei, China
jx.jiang@siat.ac.cn

Peng Yin[*]
Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, China
peng.yin@siat.ac.cn

***Abstract*—Evaluation of protein-ligand interaction is a crucial step in the process of drug discovery. Recently, several methods based on deep learning have gained impressive binary classification performance on protein-ligand binding prediction. However, lack of three-dimensional complex data still limits the accuracy and robustness of evaluation of protein-ligand binding affinity, as well as the prediction of their binding sites. In this paper, we propose a novel convolutional neural network based method for estimating the binding affinity between protein and ligand using only 1D sequence data. Even with the same amount of sample size, this model outperforms other structure-dependent traditional and machine learning based methods in terms of both binary classification and regression task. Furthermore, we use this model to identify the key amino acid residues of protein that are vital for binding interaction, which provides biological interpretation.**

Keywords—drug discovery, representation learning, binding sites prediction, interpretability, occlusion

I. INTRODUCTION

Identifying the interaction of protein and ligand plays an import role in drug discovery. Computational methods for screening potential positive compounds to target protein at the initial phase of drug discovery actually improve the success rates[1]. However, the traditional methods like molecular docking have limitations such as expert knowledge dependence and high computational cost. That is, these structure-based methods first need to predict different binding poses of protein and compounds by “docking" them together before calculating their binding energies, which tends to be a bottleneck for computational speed and accuracy. In recent years, more attention has been given to the introduction of machine learning methods into drug discovery[2]. Models such as support vector machine, random forest are capable of capturing non-linear relationships in protein-ligand complex.

More recently, deep learning, which refers to neural network with many layers of non-linear transformations, has gained remarkable achievements in various fields such as computer vision, computer games.[3, 4] The main advantage of this algorithm is that it can extract useful features automatically from the raw data during the process of training. Inspired by these successes, several studies have introduced deep learning methods into drug discovery for protein-ligand interaction identification[5–8]. Using comprehensive 3D representation of a protein−ligand complex as input, Ragoza et al. described a convolutional neural network based scoring function. This model outperforms AutoDock Vina (a widely used docking software) on both pose prediction and virtual screening[6]. Similarly, Stepniewska-Dziubinska et al. proposed a model consisting of convolutional and dense layers with 3D grid represented structure as input[7]. Their model outperformed any other classical scoring functions Another study applied convolution operations along protein and drug sequences and gained better results than machine learning based model[8]. However, the applicability of structure based model may be limited by the lack of 3D data, whereas sequence based model always suffers from the lack of interpretability.

Here we present an interpretable convolutional neural network model to evaluate the protein-ligand interaction. The model outperforms traditional docking and machine learning methods on both binary classification (protein-ligand bind or not) and regression (protein-ligand binding affinity) task using only 1D sequence information, even with the same amount of sample size. In addition to such predictions, combined with designed occlusion, the model can trace the important sites of the input data, thus to predict key amino acid residues of protein that are crucial for binding.

II. METHODS

A. Data

The Directory of Useful Decoys Enhanced (DUD-E) set containing 102 targets, 22886 active compounds and 1.4M decoys (negatives) was used for classification task[9]. DUD-E is a benchmarking platform that makes it possible for comparing our model with previously proposed methods. The ratio of negatives to positives was set to 1.5:1 to avoid unbalanced data. The PDBbind v.2018 database contains 16151 protein-ligand complexes was used for regression[10].

†: These authors contributed equally to this work.
*: To whom correspondence should be addressed.

This database provides 3D crystal structures of protein-ligand and their experimentally measured binding affinity data expressed with pKa (-logKd or -logKi) values. To assess the model performance accurately, we used 5-fold cross validation in the training process. And the data were randomly split into train/valid/test set at ratio of 60/20/20.

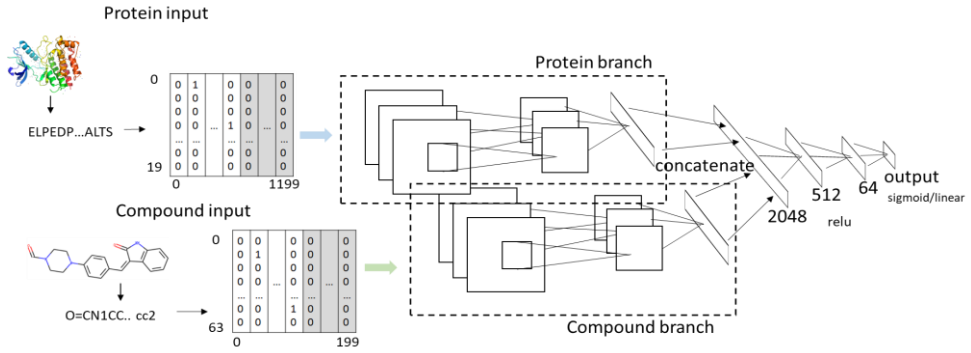

Fig.1. The architecture of proposed model.

B. *Model*

Basically, the proposed model consists of two parts: protein/ligand feature extraction by convolutional layers and interaction prediction by fully connection layers. The last dense layer is activated as output. The loss functions are defined according to different tasks: binary entropy for classification and mean squared error (MSE) for regression. To conserve space, some details would be added in the next version. Fig. 1 displays the basic architecture of the model, more details are described below:

- **Input:** The one hot encoding is used to represent input molecules. We traverse all protein amino acid sequences and ligand SMILES identifiers in databases to build dictionaries contained 20 and 64 characters for protein and ligand, respectively. Amino acid sequences are first one-hot encoded and then padded at the end to the length of fixed maximum lengths of 1200, which produced (20, 1200) dimensional matrices for proteins. Similarly, ligands SMILES identifiers are also one-hot encoded and padded to the length of 200, so the input for ligands are (64, 200) dimensional matrices.
- **Feature extraction:** The input 2D tensor is first processed by several stacked convolution layers. Specifically, two convolutional layers followed by one pooling layer are regarded as a block, totally we have three such blocks in feature extraction step. Similar to VGGNet, the techniques of "small convolution kernels" and "keeping input size" are applied in our model. In convolutional operation, the edge region is padded to keep feature map unchanged. The numbers of convolutional filters are set to 32, 32, 64, 64, 128 and 128 respectively. The size of convolutional core is set to 3*3 and the stride is set to 1. Then, the output of the last convolution layer of each branch is mapped to 1024 dimensional feature vector by a dense layer.
- **Concatenation:** Feature vectors from two branches are concatenated together and then fed into next dense layers with units of 512, 64 and 1, respectively. To avoid over-fitting, dropout layer are added after each dense layer, and the random inactivation probability is set to 0.5. Rectified linear units (ReLU) are chosen as the activation function in our model because it speeds up the training process and reduces the likelihood of vanishing gradient.
- **Output:** The activation functions of the output layer are sigmoid for classification and linear for regression. Also, the loss functions are defined according to different tasks: binary entropy for classification and MSE for regression.

C. *Training*

Xavier (Glorot) uniform weight initialization method is used to initialize weights in Convolution 2D filters and dense layers. This method corrects the variance of uniform distribution to ensure that the output variance and input variance of each layer are the same, without changing with the number of input neurons. The range of uniform distribution is from $-\sqrt{\frac{6}{n_{in}+n_{out}}}$ to $\sqrt{\frac{6}{n_{in}+n_{out}}}$, where n_{in} is the number of input units of layer, n_{out} is the number of output units of layer. The bias variables of all layers are initialized to 0. The Adam optimizer is used to train the model with 10^{-4} initial learning rate and 128 batch size. Other default parameters are set according to He *et al*.[11]. Meanwhile, the 10^{-3}, 10^{-5} learning rates and 256, 512 batch sizes were also tested but showed worse results of loss on validation set. In order to improve generalization of our model, we use RMSE with L2 regularization as the loss function, where the regularization parameter is set to 0.01. Early stopping is used to avoid over-fitting and the number of patience epochs is set to 10. Finally, the model with the minimum loss value on validation data is selected.

D. *Occlusion*

Here we present a non-parametric method “occlusion” to explore which parts of the input sequences are critical to the task. Then, compared to the positive binding pockets in proteins, we are able to see whether such sites are key parts for binding. Briefly, s_i from test samples (i = 0, 1, 2, ..., n-1, here n is sample size of test set) is expressed as tuple (protein input_i, compound input_i), where protein input_i and compound input_i are 2D tensors of shape (20, 1200) and (64, 200) respectively. While maintaining compound input_i unchanged, we systematically mask the protein input_i in s_i to track the changes of the output. Then the importance of each sub-sequence in the sequence to the prediction can be calculated. More details are described below:

- First, the upper left corner element of the two-dimensional matrix is regarded as the coordinate origin, the direction of one-hot amino acid types and sequence length are seemed as x-axis, y-axis respectively. Cartesian coordinate system is established in this step.
- Second, "mask" is carried out sequentially along y-axis direction to generate the occlusion result s_{ij}'. Values from x=0 to x=19, $y = j - \left\lfloor \frac{size-1}{2} \right\rfloor$ to $y = j + \left\lfloor \frac{size-1}{2} \right\rfloor$ are replaced by 0 to produce masked result protein input_ij'. The masked result protein input_ij' and compound input_ij' form sij' $\left(j = \left\lfloor \frac{size-1}{2} \right\rfloor, \left\lfloor \frac{size-1}{2} \right\rfloor + 1, \left\lfloor \frac{size-1}{2} \right\rfloor + 2, \ldots, 1200 - \left\lfloor \frac{size-1}{2} \right\rfloor - 1\right)$, here j denotes the y coordinate of blocked window central point and size denotes the length of blocked window. The hyper-parameter "size" can be optimized through experiments and it is set to 15 by default. The stride of blocked window is set to 1 to avoid any omission.
- Third, according to the occlusion result s_{ij}, the changes of output caused by occlusion are tracked. We define an evaluation measure K to quantify the changes as:

$$K_{ij} = \frac{|p_{ij} - v_i|}{|p_i - v_i| + \varepsilon} \quad (1)$$

where v_i denotes the actual binding value of sample s_i, p_i and p_{ij} denote the predicted value of s_i and the occlusion result s_{ij} respectively. ε is a small positive real number which is used to keep denominator not equal to 0. Then we can visualize the critical parts for binding in protein sequences through heat-map.

III. Results

A. Classification

Table I shows the comparison of methods including traditional docking, machine learning and other deep learning models for identifying actives and decoys on DUD-E dataset. AutoDock Vina and Smina are open source molecular docking programs that are widely used in traditional virtual screening. AtomNet[5] and 3D-CNN[6] are deep learning models for predicting interaction between protein-ligand based on 3D structure. Obviously, our model outperforms these 3D structure methods on classification even with 1D sequence input. It indicates that deep model achieve much better result than shallow model in this task.

TABLE I. Comparison Of Methods On DUD-E Dataset.

	Methods					
	Smina	***AutoDock Vina***	***SVM***	***AtomNet***	***3D-CNN***	***Our model***
AUC	0.696	0.716	0.811	0.895	0.868	0.997

The AUC scores of Smina, Vina, AtomNet, 3D-CNN are derived from[5, 6]

B. Regression

The measures root mean square error (RMSE), Pearson's correlation coefficient (R) and standard deviation (SD) are selected to evaluate our model's performance on regression. Among these, RMSE is used to calculate the differences between predicted and real values, Pearson's coefficient R is a measure of the linear correlation between predicted and real values and SD is used to quantify the amount of variation of values. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

where N is the sample size of corresponding set, y_i is the real binding value experimentally measured whereas $\hat{y}_i$ is the predicted value by our model.

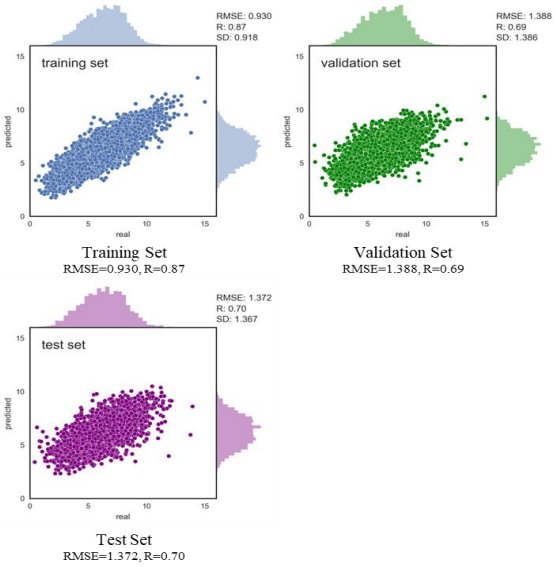

Fig.2. Evaluation of binding affinity on the PDBbind v2018 dateset.

The loss changes on training set and validation set were monitored during training. After 28 epochs, the validation loss began to rise while the training loss continued declining, that means the model started to over fit. So the model on 28th epoch was saved and selected as the final model.

Fig.2 displays the predicted against real binding values of protein-ligand complex on PDBbind v2018 dataset. As shown, our model achieves the lowest RMSE on training set which is used to learn "common rules" for evaluation. The model also performs well on validation and test set. That means, our model learned some important features of interactions between proteins and ligands and can use this "knowledge" to predict interactions even it has never seen before. Actually, the results predicted by our model are slightly better than those achieved by 3D structure based deep model (pafnucy)[7], which had RMSE=1.44, SD=1.43 and RMSE=1.42, SD=1.37 for validation and test set, respectively. To the best of our knowledge, pafnucy provides the state-of-the-art protein-ligand prediction results on PDBbind dataset. Although the values of R exhibited in their study are higher than ours (0.78 for test), the accuracy and generalization of 1D input model would be easier to improve as it can extract more binding features from large amounts of data without known structural information.

C. Biological interpretation

As a representation learning method, deep learning can automatically learn features from raw data. While fast and accurate, deep learning model is a "black box" that is difficult to know why it works well on specific tasks. Hence we develop a method to explore how the model processes the biological data. Similar to the occlusion algorithm which used to explore whether CNN model can locate the key target

in input image[12], this method is actually masking some subsequences of the input protein, thus can be used to define where is the binding sites.

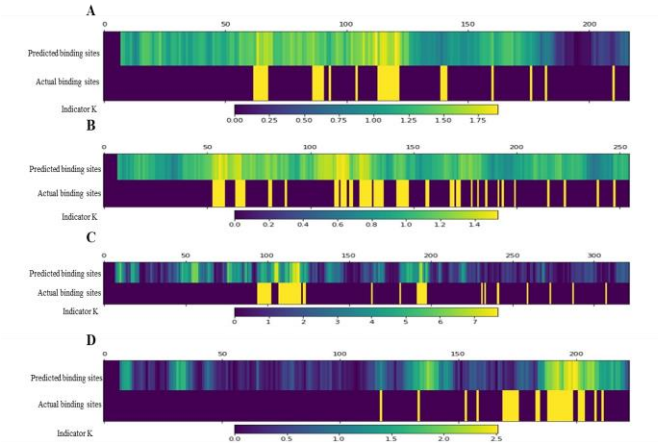

Fig.3. The alignments of predicted and actual binding sites of protein sequences. The corresponding PDB ID: (A) 5k09; (B) 4i4f; (C) 4a51; (D) 3ati. (The abscissa axis is the length of protein sequences.)

As defined above, the range of K_{ij} is $(0, +\infty)$. When $K_{ij} > 1$, occlusion of the corresponding subsequences would increase the difference between real and predicted value, which means the masked subsequences play an important role in prediction. The larger K_{ij} value, the greater influence of the corresponding region on the prediction. If $K_{ij}=1$, the corresponding region has no effect on the prediction results. As for $K_{ij} < 1$, which means mask of corresponding sequences can reduce the error of the predictions. One possible explanation for that is the whole protein sequence may produce some noise information from the portions of the sequence that are not involved in the binding.

Heat-maps of some samples from test set are exhibited in Fig.3. The yellow regions in actual binding sites are the pockets (the binding site for ligand) in protein sequences. In the heat maps of predicted binding sites, as indicated by K value, the regions close to yellow are considered to be important for binding by the model. Obviously, the predicted binding sites are very close to the actual ones which suggests our deep model indeed processes the data in a proper way. Although there is a slight shift in alignments, this may be partially caused by local translation invariance of CNN introduced by pooling operation, which is especially useful when we only care about whether the desired characteristics exist in a certain area not a specific location. Additionally, the selected coordinates of masked area at the center point in the calculation of heat map more or less affect these shifts.

IV. Conclusion

In this paper, we introduce an interpretable convolutional neural network based model to predict the binding between protein and ligand. This model is shown to accurately predict both binding possibility and value by using only 1D information. It should be noted that our model performs even slightly better than the state-of-the-art 3D structure based deep learning model. Although 3D complex retains complete binding information, the amount of such data is too small and an end-to-end model cannot extract enough common features. The poor performance on unseen complex may be a common problem for 3D based predictive models. The generalization of model with 1D input would be easier to improve because of large amounts of data without structural information.

One of the most common problems for deep learning is that they are regarded as black boxes, lacking accountability, trustworthiness and effective ways for debugging. Although convenient and accurate, deep model may yield favorable results for the misguided reasons. Therefore, it is necessary to check if the model utilized input information correctly. Interestingly, we show that our model can be used to predict key amino acid residues which are important for binding in combination with feature extraction algorithm applied in computer vision. It should be emphasized again that our model pinpoints the binding sites using only 1D sequence information, which suggests the deep model processed data in a proper way. We are excited about the future application of such sequence based predictive model. Different datasets would be used to improve the generalization ability of our model. We plan to apply multi-task learning in the downstream of the model, thus it can be extended to problems involving drug toxicity and sensitivity prediction.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NO. 11801542), the Shenzhen Fundamental Research Projects (JCYJ20170818164014753, JCYJ20170818163445670 and JCYJ20180703145002040) and China Postdoctoral Science Foundation (2018M643242).

References

[1] D.-L. Ma, D. S.-H. Chan, and C.-H. Leung, “Drug repositioning by structure-based virtual screening,” *Chem. Soc. Rev.*, vol. 42, no. 5, p. 2130, 2013.

[2] A. Varnek and I. Baskin, “Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis ?,” *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1413–1437, Jun. 2012.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[4] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep Learning for Computer Vision: A Brief Review,” *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, 2018.

[5] I. Wallach, M. Dzamba, and A. Heifets, “AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery,” *Data Min. Knowl. Discov.*, vol. 22, no. 1–2, pp. 31–72, Oct. 2015.

[6] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, “Protein-Ligand Scoring with Convolutional Neural Networks,” *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 942–957, 2017.

[7] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, Nov. 2018.

[8] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: deep drug–target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[9] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, “Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking,” *J. Med. Chem.*, vol. 55, no. 14, pp. 6582–6594, 2012.

[10] R. Wang, X. Fang, Y. Lu, and S. Wang, “The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures,” *J. Med. Chem.*, vol. 47, no. 12, pp. 2977–2980, Jun. 2004.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[12] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *European conference on computer vision (ECCV)*, 2014, pp. 818–833.