PROTEIN-LIGAND BINDING AFFINITY PREDICTION USING DEEP LEARNING

ABENA ACHIAA ATWEREBOANNAH[1], WEI-PING WU[1,2*], LEI DING[2], SOPHYANI B. YUSSIF[1], EDWIN KWADWO TENAGYEI[1]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, P. R. China
[2]SipingSoft Co. Ltd., Tianfu Software Park, Chengdu, P. R. China
E-MAIL: atwereboannah@gmail.com, wei-ping.wu@uestc.edu.cn, xiandao.airs@gmail.com

Abstract:

Protein-ligand prediction plays a key role in drug discovery. Nevertheless, many algorithms are over reliant on 3D structure representations of proteins and ligands which are often rare. Techniques that can leverage the sequence-level representations of proteins and ligands are thus required to predict binding affinity and facilitate the drug discovery process. We have proposed a deep learning model with an attention mechanism to predict protein-ligand binding affinity. Our model is able to make comparable achievements with state-of-the-art deep learning models used for protein-ligand binding affinity prediction.

Keywords:

Deep learning; Protein–ligand binding affinity; Self-attention; Drug discovery

1. Introduction

. The rate of success during the early stages of drug discovery depends on the binding affinity a ligand assumes with a target protein. In silico techniques such as quantum mechanics and molecular dynamics have been deployed for the binding affinity prediction tasks. Nevertheless, they are unpopular because of high computational cost.

Computer-Aided Drug Discovery such as High-Throughput Screening (HTS) [1], has achieved immense success in drug discovery. Recently, Deep Learning (DL) techniques have excelled in solving problems in the Drug domain such as Drug Permeability studies [2] and Drug-Target Interaction Prediction and have superseded conventional screening methods.

Data featurization is believed to be important in ensuring the success of any model. In the drug discovery domain, such feature extraction techniques have over powered the utilization of traditional molecular descriptors and fingerprints, in that they are able to extract features that may have been illusive prior to the process of training.

In this work we propose a DL framework with an attention mechanism to focus on extracting relevant features from each of the three datasets we obtained from the PDBbind database version 2016 [3].

The subsequent sections are organized as follows: section 2 presents the related work of our study, section 3 discusses the data used and our proposed model, section 4 highlights the experiments design of our study and results obtained. We then discuss the results in session 5 and draw conclusions in session 6.

2. Related work

Researchers in the field of Bioinformatics have used DL models extensively for protein-ligand binding affinity prediction tasks. Structure-based and Ligand-based methods and are the two main categories. DL Models, such as TopologyNet [4] are structure-based algorithm highly dependent on the nature of the molecular input and only uses the provided complex protein-ligand structures. Structure-independent approaches include MONN [5] which gives a more interpretable prediction of binding affinities, DeepDTA [6] and WideDTA [7]**.** These models take advantage of SMILES (Simplified Molecular Input Line Entry System) ligands and sequences of Proteins. The ligand-based methods again represent the vectors of features with molecular fingerprints and the neural networks are established on top of them. Thus, some ligand-dependent ML models have achieved remarkable success in bioactivity [8] and toxicity prediction [9].

3. Materials and methods

3.1. Data

The three datasets obtained from the PDBbind database

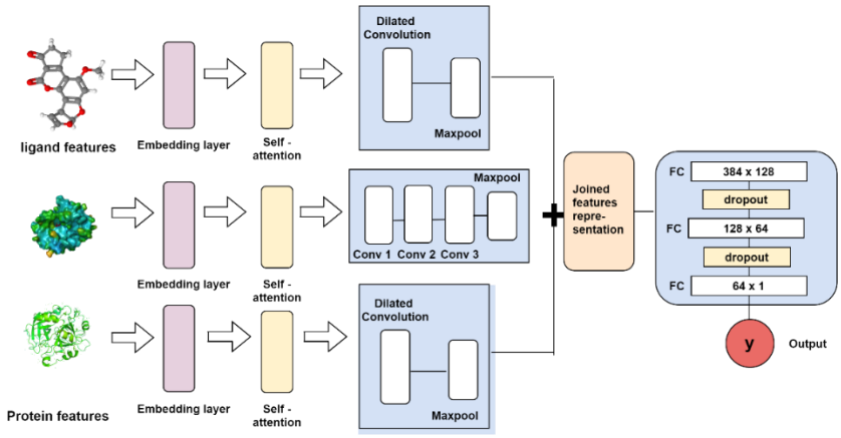

Figure 1. Our proposed deep learning architecture

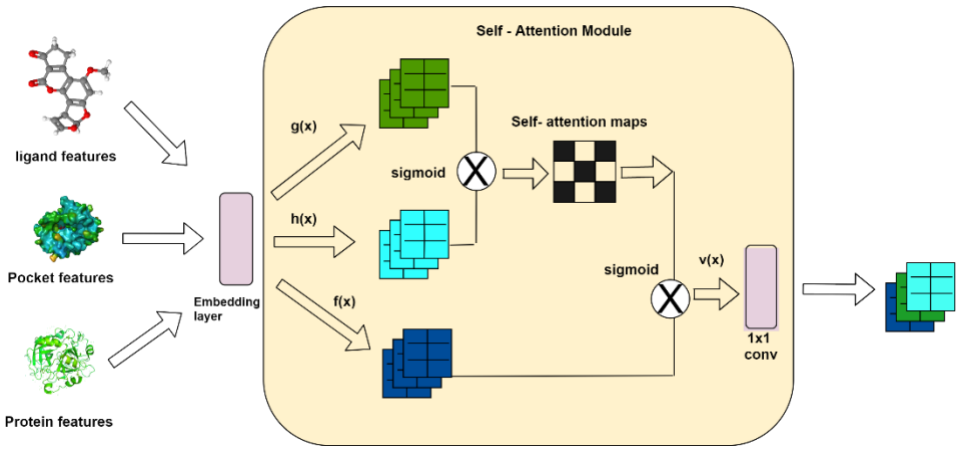

Figure 2. Detailed Self-attention architecture [11]

v2016 comprise 9221 protein-ligand components, for the general set, the refined set made up of 3685 binding affinities and the core 2016 set containing 290 complexes. In order to ensure a fair comparison, we followed the data pre-processing scheme of DeepDTA and tested on the core16 test set. Just like DeepDTAF [10], we fed our model with 1D sequence data. This was divided into 3 subsets consisting of the ligand, protein and pockets descriptions.

3.2. Model

We represented the three input datasets with embedded layers of vector dimension 128. In order to target the relevant features, we passed these features through Self attention layers before passing their one-dimension output through dilated convolution blocks with the exception of the output from the pocket module which we applied 3 normal convolution layers of one dimension with 3 corresponding filters of 32, 64 and 128 [8]. We then applied max pooling to all 3 modules and finally concatenated their outputs before passing them to a classification model made up of 3 fully connected layers and two dropout layers, both of rate 0.5 as shown in Figure 1.

3.3 Self-attention layers

The sublayers used for the implementation of Self-attention uses h heads of attention. The output from the sublayer is derived by applying a linear transformation to the concatenated outputs from each of attention heads [12]. Figure 2 shows the detailed Self- attention architecture.

Let $x = (x_1, ..., x_n)$ be the sequence of an input data each Self-attention head operates on. Number of elements is represented by n and $x_i \in R^{d_x}$. Again, a new sequence $z = (z_1, ..., z_n)$ with the same length $z_i \in R^{d_z}$ is computed by each attention head. Every one output z_i, is the sum of the weights of a linearly translated input elements given as:

$$x_i = \sum_{j=1}^{n} a_{ij}(x_j W^V) \quad (1)$$

Where a_{ij} is the coefficient of the weight which is calculated using a softmax function expressed as:

$$a_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \quad (2)$$

While e_{ij} is obtained with an equation which compares two inputs:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}} \quad (3)$$

Where W^Q , W^K , $W^V \in R^{d_x \times d_z}$ are matrix parameters different from layer to layer and attention heads. We adopted the Self-attention model of [10]**.**

3.4 Dilated convolution

Unlike traditional convolutional layers, the dilated convolution layers are able to retrieve contextual information in a multiscale.

Their major advantage over traditional convolution is that the dilated convolution initially records inherent sequence information by extending the field of convolution kernel without increasing the model's parameters. The operator of dilated convolution $*_l$ is expressed as:

$$(F *_l k)(P) = \sum_{s+lt=P} F(s)k(t) \quad (4)$$

where the discrete function is represented as $F : Z^2 \to R$. $k : \Omega_1 \to R$ is the distinct 3×3 filters, the dilation rate is l with s and t being the subscripts of element vector. We used the same dilation configuration as DeepDTAF for the protein and ligand modules.

3.5 Evaluation metrics

In order to make a fair comparison we used the same evaluation metrics as TopologyNet. The metrics used are described briefly.

3.5.1 Mean absolute error (MAE)

The Mean absolute error is the average of the absolute difference between the ground truth and predicted values in the dataset. This is expressed as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} | y_i - \hat{y} | \quad (5)$$

where, $\hat{y}$ is the predicted affinity of y and $\bar{y}$.

3.5.2 Root mean square error (RMSE)

This is used to measure the average deviation in L2 norm between the ground truth value and the related predicted value. It is given as:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2} \quad (6)$$

3.5.3 Standard deviation (SD)

For regression tasks, SD is expressed as:

$$SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left[y_i - (a\hat{y} + b) \right]^2} \quad (7)$$

where the number of protein-ligand structures are represented by N . The gradient and intercept of the function line between the actual and predicted values are represented with a and b .

3.5.4 Concordance index (CI)

The ratio of the predicted and true affinity values for two randomly selected protein–ligand complexes is computed with CI in a particular order. This is defined as:

$$CI = \frac{1}{Z} \sum_{y_i > y_j} h(\hat{y}_i - \hat{y}_j) \quad (8)$$

where $\hat{y}_i$ is the predicted value for the greater binding affini ty value y_i and $\hat{y}_j$ is the predicted value for the smaller affin ity value y_j . Z is the sum of the protein-ligand complexes. A larger CI is indicative of a good prediction performance.

3.5.5 Correlation factor (R)

The Correlation factor *R* shows how closely related the predicted affinity values are to the ground truth. The closer the R score to 1, the better the correlation. This is expressed mathematically as:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

4. Training, results and evaluation

We trained our model for 30 epochs on 4 GeForce GTX 1080Ti server, Intel Xeon CPU E5-2687W with 128GB RAM. Our results are presented in Table 1 and compared in Table 2 with other state-of the art deep learning models as presented in [10] on the test105 set.

Table 1. Results obtained from Our Model

Metrics	Training	Validation	Test
RMSE	0.819	1.378	1.436
MAE	0.624	1.068	1.099
R	0.900	0.743	0.750
SD	0.812	1.373	1.438
CI	0.864	0.775	0.784

Table 2. Comparing with accuracies on test105 set

Metrics	TopologyNet	DeepDTA	Ours
RMSE	4.143	**1.425**	1.436
MAE	3.841	1.134	**1.099**
R	0.444	0.652	**0.750**
SD	1.530	**1.432**	1.438
CI	0.646	0.738	**0.784**

For the purpose of better evaluating performance of models, we mainly resort to MAE, R as well as CI. RMSE was not used as a major measurement because it casts a stronger restriction compared with MAE to force predicted value to conform with the ground value, but our ground value of affinity is naturally noisy. Meanwhile, SD also wasn't taken as a relatively accurate indicator of model's performance, since a and b vary in transformation space and such transformation is more likely to induce bias unavoidably.

As presented in Table 2, Our model outperformed the other models with an MAE of 1.099, a CR of 0.750 and a CI score of 0.784, thus corroborating the importance of combining attention with convolutional layers for better model performance.

5. Conclusion

In this work we have, proposed a deep learning algorithm with an attention mechanism to predict protein-ligand binding affinities. Our model is able to achieve comparative results with other leading deep learning models with the highest MAE, CR and CI scores.

References

[1] Agyemang B. Author, and WP. Wu, "Multi-view self-attention for interpretable drug–target interaction prediction". Journal of Biomedical Informatics. 2020 Oct 1;110:103547.

[2] Achiaa Atwereboannah A. Author, and WP. Wu, "Prediction of Drug Permeability to the Blood-Brain Barrier using Deep Learning" 4th International Conference on Biometric Engineering and Applications 2021 May 25 (pp. 104-109).

[3] Liu Z, Su M, Han L, Liu J, Yang Q, Li Y, Wang R. Forging the basis for developing protein–ligand interaction scoring functions. Accounts of chemical research. 2017 Feb 21;50(2):302-9.

[4] Cang Z. Author, and Wei GW. Topology net: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput Biol 2017;13:e1005690.

[5] Li S. Author, and F. Wan, MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems* 2020;**10**:308, e311–22.

[6] Öztürk H. Author, and A. Özgür, Deep drug-target binding affinity prediction. *Bioinformatics* 2018;**34**: i821–9.

[7] Öztürk H. Author, and E. Ozkirimli, Özgür A. Wide DTA: prediction of drug-target binding affinity. 2019arXiv preprint arXiv:1902. 04166.

[8] Lenselink, E. B. Author "Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set", Journal of Cheminformatics, 9, 45, (2017).

[9] Xu, Y. Author, Deep learning for drug-induced liver injury. Journal of Chemical Information and Modeling, 55(10), 2085–2093 (2015).

[10] Wang. K. Author, and R. Li, "DeepDTAF: a deep learning method to predict protein–ligand binding affinity". Briefings in Bioinformatics (2021).

[11] Zhang H. Author, and Han, "Self-Attention Generative Adversarial Networks" (2018).

[12] Peter S. Author, and J. Uszkoreit, "Self-attention with relative position representations",arXiv preprint arXiv:1803.02155, 2018.