Attention-enhanced Graph Cross-convolution for Protein-Ligand Binding Affinity Prediction

Xianbing Feng[1†], Jingwei Qu[1†], Tianle Wang[2], Bei Wang[3], Xiaoqing Lyu[1*], Zhi Tang[1]
[1]*Institute of Computer Science and Technology, Peking University, Beijing, China*
[2] *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China*
[3] *SenseTime Research, Shanghai, China*
fengxb@stu.pku.edu.cn, louiswong.cs@bupt.edu.cn, wangbei1@sensetime.com,
{qujingwei, lvxiaoqing, tangzhi}@pku.edu.cn

***Abstract*—The binding affinity between drugs and proteins is a substantial part of the drug discovery process. Graph neural networks (GNNs) have shown great promise in graph-related structure by learning the representations of graphs, which are suitable for tasks such as binding affinity prediction. However, most of the existing GNN architectures only pay attention to the information flow on a single graph, while the interaction between two graphs is unconcerned. In this paper, we propose an attention-enhanced graph cross-convolution network (GCAT) to explore binding affinity on pure 3D atomistic geometry. It consists of two components: cross-convolution and self-attention pooling. Specifically, cross-convolution performs an aggregate-update mechanism to simulate the interaction between the protein and the drug, then self-attention pooling is adopted to capture global interactions and get graph-level representations. Extensive experiments conducted on the PDBBind dataset demonstrate the effectiveness of our GCAT.**

***Index Terms*—protein-ligand affinity prediction, graph neural network, cross convolution, self-attention, interaction modeling**

I. INTRODUCTION

The research of protein-drug interaction has attracted attention for many years due to its potential in aiding drug discovery. Traditional methods of drug discovery are expensive and time consuming [1], which drives us to explore efficient models that can estimate the interaction strength of new drug–target pairs based on previous drug–target experiments.

Inspired by the success of deep learning in many areas such as computer vision and natural language processing, many researchers have also explored the combination of deep learning and biomedicine. There are three different representations of molecules, namely linear sequences (1D), chemical bond graphs (2D) and the 3D positions of the component atoms, which derived various exquisite models. Then for protein-drug affinity prediction, current deep neural network based methods can be summarized as a paradigm that uses well-designed models, e.g. RNN [2], CNN [3, 4], to learn feature representations and then predict affinity through a fully connected network. Most of these approaches are still emphasized on 1D or 2D representations rather than 3D atomistic geometry [5, 6], and 1D representation of protein can bring an over-fitting problem in training. As for how to make binary relationship predictions, these methods can be divided into two categories according to the number of branches, including one brunch way which combines protein and small molecular into complex with the heuristic algorithm for computation, e.g. [4]. More methods still use two branches, that is to update the representation of proteins and small molecules separately [3, 7]. These approaches can learn good representations for downstream tasks indeed, yet most of these approaches ignore the interactivity between the two entities, which is also important and reasonable.

A growing number of Graph Neural Networks (GNNs) have been proposed to update the feature on graphs with *aggregation* and *combination* mechanism [8, 9]. GNNs have shown superior performance by treating the molecule as a chemical bond graph and performing message passing scheme on it [10]. However, GNNs-based approach only focuses on local receptive field, and it is easy to encounter over-smooth or over-fitting problems [11]. A few works have been proposed to use Transformer-based techniques to extract global features on graphs [12, 13] which can be viewed as a variant of the GAT [9] that applied attention mechanism on a single graph. But there is little work exploring mutual attention mechanisms between two graphs for tasks like protein-ligand binding affinity prediction.

To address the aforementioned challenges, we propose an end-to-end framework, GCAT, which stands for **G**raph **C**ross convolution with **A**ttention enhanced affinity prediction ne**T**work, to simulate the interaction between the protein and the molecular by introducing attention mechanism. We summarized our contributions as follows:

- We propose an end-to-end GNN framework (**GCAT**) by interaction modeling for protein-ligand binding affinity prediction.
- We devise graph intro-convolution and attention enhanced graph cross-convolution to simulate the interactivity between protein and small molecular, which is learned on pure 3D atomistic geometry.
- We adopt a shared self-attentive pooling operation instead of previous graph pooling approach to aggregate information from the whole graph and obtain graph-level embedding.

†: These authors contributed equally to this work.
*: Corresponding author

II. METHODOLOGY

A. Preliminaries

In our settings, for both protein graph and molecular graph, node is an atom and edges are defined between all atoms separated by less than 4.5Å as mentioned in atom3d. The nearest neighbors of node i are defined as $\mathcal{N}(i)$. Detailed symbols are shown in Table I.

TABLE I
SYMBOL NOTATIONS

G_p	protein graph	$\mathbf{H}_{1i}^l$	l-th layer embedding of node i in G_p
G_m	molecular graph	$\mathbf{H}_{2j}^l$	l-th layer embedding of node j in G_m
$\mathcal{N}(i)$	neighbors of node i	$\mathbf{h}_1^l$	l-th layer embedding of G_p
A	attention map	$\mathbf{h}_2^l$	l-th layer embedding of G_m
W	parameter matrix	$\mathbf{S}$	correlation adjacency matrix

B. Graph Intro-convolution

The massage passing scheme is inspired by GCN [8], where features are effectively aggregated from the adjacency nodes and the node itself.

$$\mathbf{H}_i^l = f_{update}(H_i^{l-1}, \frac{1}{||\mathcal{N}_{(i)}||}\sum_{j\in\mathcal{N}_{(i)}} f_{msg}(\mathbf{H}_j^{l-1})) \quad (1)$$

Where Eq. (1) is the message passing network which is denoted as graph intro-convolution. f_{msg} is the message passing function. f_{update} is the aggregating and updating function which fuses origin features and updated features in some ways, e.g. summing or concatenating. We perform graph intro-convolution in both G_p and G_m, which can extract local relations in G_p and G_m separately from the perspective of unary relations.

C. Attention Enhanced Graph Cross-convolution

For two graph instances, we calculate the distances among atoms of the two graphs by distance cutoff algorithm with 4.5Å threshold to identify the inter-graph correlation. That is, if the distance between two atoms is less than 4.5Å, there is a connection between them. We have $\mathbf{S} \in \mathbb{R}^{N_p \times N_m}$ which is the correlation adjacency matrix representation of the edges between G_p and G_m.

From another perspective, we could treat $\mathbf{S}$ as a predefined static attention of G_p to G_m which would limit expressivity of model compared to dynamic attention. Therefore, we could enhance $\mathbf{S}$ through a self-attention mechanism. For the dynamic attention between two graphs, we use Scaled Dot-Product Attention [14] calculated as:

$$\begin{aligned} Q = \mathbf{H}_1 W_Q, K = \mathbf{H}_2 W_K, \\ A = softmax(\frac{QK^T}{\sqrt{d_k}}) \end{aligned} \quad (2)$$

Where $\mathbf{H}_1$/$\mathbf{H}_2$ is projected into query Q/key K by parameter matrix W_Q/W_K. Attention map A captures the similarity between queries and keys by computing the dot products of the query with all keys, with the scaling factor $\frac{1}{\sqrt{d_k}}$ and softmax function.

Then an integrated correlation adjacent matrix $\mathbf{S}'$ can be obtained by:

$$\mathbf{S}' = (1-\lambda)\mathbf{S} + \lambda A \quad (3)$$

where λ is the balancing hyper-parameter to control the strength of attention mechanism, which is a trade-off between static and dynamic attention. Then we do cross-convolution between two graphs with $\mathbf{S}'$:

$$\begin{aligned} \mathbf{h}_1^l = W_{c1}(\mathbf{h}_1^{l-1} || \mathbf{S}' \mathbf{h}_2^{l-1}), \\ \mathbf{h}_2^l = W_{c2}(\mathbf{h}_2^{l-1} || \mathbf{S}'^T \mathbf{h}_1^{l-1}), \end{aligned} \quad (4)$$

where $W_{c1}, W_{c2} \in \mathbb{R}^{(2\times d_{l-1})\times d_l}$ are learned weight matrices and $||$ denotes vector concatenation.

Eq. (4) is a single cross-graph update scheme from the perspective of graphs. We generalize from graph intro-convolution to graph cross-convolution which features are aggregated from nodes with similar features in the other graph in Eq. (5) from the perspective of nodes.

$$\mathbf{H}_{1i}^l = f_{update}(\mathbf{H}_{1i}^{l-1}, \frac{1}{||\mathcal{N}_{(i)}||}\sum_{j\in\mathcal{N}_{(i)}} \mathbf{S}'_{i,j} f_{msg}(\mathbf{H}_{2j}^{l-1})) \quad (5)$$

Here we apply dynamic self-attention on $\mathbf{S}$ instead of origin heuristic $\mathbf{S}$ for cross-convolution in Eq. (3). This self-attention enhanced $\mathbf{S}'$ to obtain a better performance proved in later ablation experiment.

D. Self-attentive READOUT Function

To encode the importance of different nodes into a unified embedding, attention mechanism could be used to dynamically adjust the node participation within a graph. Inspired by SEAL [15], we further utilize a shared self-attentive READOUT function to generate the graph-level embedding:

$$\mathbf{g} = softmax(W_2 tanh(W_1 \mathbf{H}^T))\mathbf{H}, \quad (6)$$

where W_1 and W_2 are two weight matrices and $\mathbf{H}$ is node embedding. $tanh$ activation function would introduce nonlinearity to this function. As final graph embedding, $\mathbf{g}$ is size invariant because it does not depend on the number of nodes, and is permutation invariant because the importance of each node is learned regardless of the node sequence. Then after a multilayer perceptron MLP, we can get a prediction of affinity $\hat{y}$.

E. Loss Function

For GCAT, we introduce Huber loss $\mathcal{L}_r$ for the regression task, which is more robust to outliers than Mean Square Error loss. Specially, when $\lambda = 1$, $\mathbf{S}$ would not be utilized as training information that is an unfair comparison. Thus, we use additional Cross Entropy Loss $\mathcal{L}_a$ on GCAT to minimize the gap between $\mathbf{S}$ and $\mathbf{S}'$. Then we get the final loss function:

$$\mathcal{L} = \mathcal{L}_r + \alpha\mathcal{L}_a, \quad (7)$$

where α serves as a weight coefficient to balance the $\mathcal{L}_r$ with $\mathcal{L}_a$.

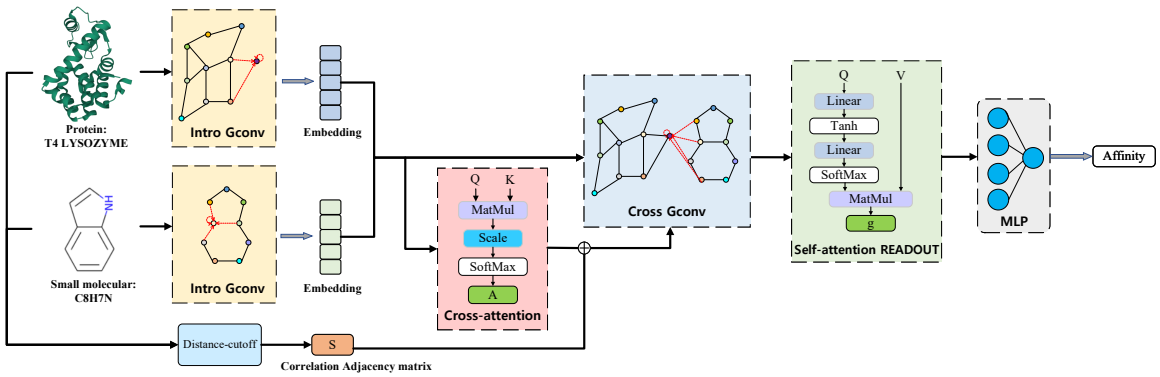


Fig. 1. Network architecture of the proposed graph intro-convolution and attention enhanced cross-convolution approaches for affinity prediction to learn interaction information flow between protein and small molecular, then learn global attention for protein-ligand binding affinity prediction. The graph intro-convolution model, graph cross-convolution and cross-attention model are all learnable in an end-to-end fashion.

III. EXPERIMENTS

A. Dataset

We use PDBBind database refined set $v.2019$ [16], following the split way of train/validation/test sets in atom3d. It sets a 30% sequence identity threshold to limit homologous proteins appearing in both train and test sets rather than 90% identical to several proteins in the training set and test set in CASF [17] set, which prevents overfitting to specific protein families in previous methods.

B. Evaluation Metrics and Compared Methods

We use Root Mean Square Error(RMSE), Pearson's correlation coefficient(R_P), and Spearman correlation(R_S) as metrics. To demonstrate the effectiveness of our approach, we choose DeepDTA [3] (CNN based), DeepAffinity [2] (RNN-CNN based), Pafnucy [4] (3D-CNN based) to compare with our approach with the new split ways in atom3d.

C. Implementation Details

For proteins and small molecules, the input feature dimensions are 18 and 43. For self-attention on cross-convolution, we set attention hidden size as 128, head as 4 and the dropout rate as 0.1. For the self-attention READOUT function, we set hidden size as 256 and attention output size as 64. We use Adam algorithm [18] as an optimizer, the learning rate is set as 0.001 initially.

D. Results and Analysis

We first compare our proposed GCAT with baseline approaches. As shown in Table II, we can observe that GCAT achieves the best performance, with 3.5% and 12.2% improvement of RMSE over 3DCNN and ENN baseline respectively.

Among all methods, DeepAffinity [2] show relatively poor performance due to the failure of considering the spatial structure of proteins and small molecules. It indicates that simply modeling the molecules into SMILES with protein sequence information is not capable of predicting structure-based protein-ligand binding affinity. By contrast, Pafnucy achieves better results for taking atomic position coordinates as input. However, Pafnucy ignores the interaction between the protein and the molecular and can not take advantage of long-range interaction features. Our proposed **GCAT** can not only capture spatial structural information, but also handle interactions in the complex through attention mechanism.

TABLE II
RESULTS OF PROTEIN-LIGAND AFFINITY PREDICTION ON PDBBIND REFINED SET

Methods	30% identity		
	RMSE ↓	R_P ↑	R_S ↑
3DCNN [6]	1.416	0.550	0.553
ENN [6]	1.568	0.389	0.408
DeepAffinity [2]	1.893	0.415	0.426
DeepDTA [3]	1.565	0.573	0.574
Pafnucy [4]	1.489	0.539	0.537
Ours	**1.382**	**0.585**	**0.592**

a) Ablation Studies: To verify the effectiveness of factors that influence the final performance, we compare GCAT with its variants and the results are exhibited in Table III.

First of all, we compare the results of the proposed method with the model with simple graph cross-convolution, which have better performance in both GCN and GAT intro-convolution backbone. Secondly, in the case of simple graph global pooling, the performance of the model has decreased, implying that global attention enhanced pooling method could capture better graph-level representations. Then, we compare the effects of GAT and GCN in the **GCAT**, i.e. **GCAT-GAT** and **GCAT-S**, and find that using the attention mechanism regardless of the tasks does not yield good results. Thirdly, we use pure attention enhanced model **GCAT-A**, namely use the learned attention with GAT backbone and self-attention READOUT, which has an unsatisfactory performance because attention mechanism lacks adequate supervised information. For a fair comparison, we set L as cross-entropy loss function and $\mathbf{S}$ as ground truth correlation matrix to minimize the gap between attention map A and $\mathbf{S}$. With this setting, **GCAT-L** have got 2.89% improvement compared with **GCAT-A**.

TABLE III
ABLATION STUDIES OF OUR PROPOSAL.

method	local feature		global feature	valid RMSE↓
	intro	cross		
GCAT-NP	GCN	-	global add pooling	2.021
GCAT-SP	GCN	$\mathbf{S}$	global add pooling	1.570
GCAT-GAT	GAT	$\mathbf{S}$	self-attention	1.475
GCAT-S	GCN	$\mathbf{S}$	self-attention	1.453
GCAT-A	GCN	A	self-attention	1.462
GCAT-AL	GCN	A+$\mathcal{L}_a$	self-attention	1.411
GCAT	GCN	$\mathbf{S}'$	self-attention	**1.382**

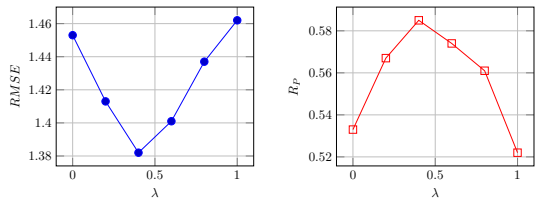
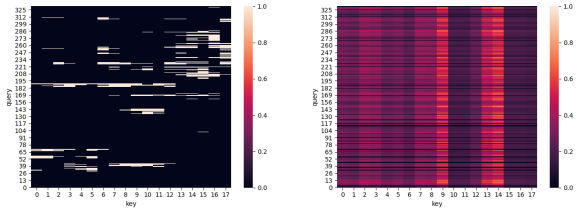


Fig. 2. Parameter analysis of λ. With the increase of λ, more attention information between G_p and G_m is available to our model and beneficial for learning complex representation better, which leads to more performance improvements when $\lambda \leq 0.4$. After that, too much attention information will introduce additional noise and degrade the performance.

b) Parametric Analysis.: As depicted in Fig. 2, we investigate the performance variation for GCAT of necessary hyperparameter by varying each parameter while keeping others fixed as default settings.

E. Visualization

To verify the effectiveness of the attention mechanism, we visualize the learned attention map A. Fig. 3 shows a heatmap of GCAT learned attention scores between G_p and G_m.



(a) correlation matrix $\mathbf{S}$ (b) attention map A

Fig. 3. For entity of PDB ID 1c86, origin correlation matrix $\mathbf{S}$ (Fig. 3(a)) can be considered as a static attention map: the ranking of attention coefficients is global for all nodes between G_p and G_m. In contrast, GCAT(Fig. 3(b)) can actually compute dynamic attention map A, where every query has a different ranking of attention coefficients of the keys.

IV. CONCLUSION

In this paper, we proposed an attention enhanced graph convolution based model, **GCAT**, to dynamically simulate the interaction between the protein and the ligand by leveraging the 3D atomistic geometry. In the future work, researchers can explore more structural information than just as a basis for constructing graphs with simple heuristic algorithms, which could be better with dynamic learned threshold.

ACKNOWLEDGMENT

This work was supported by National Key Research and Development Program of China (No. 2019YFB1406303), National Natural Science Foundation of China (No. 61876003), and Beijing Natural Science Foundation (No. L192024). It is also a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

REFERENCES

[1] Philip Cohen. Protein kinases—the major drug targets of the twenty-first century? *Nature reviews Drug discovery*, 1(4):309–315, 2002. I

[2] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019. I, III-B, III-D, II

[3] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. I, III-B, II

[4] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018. I, III-B, II

[5] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018. I

[6] Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. Atom3d: Tasks on molecules in three dimensions, 2021. I, II

[7] Yihan Zhao, Kai Zheng, Baoyi Guan, Mengmeng Guo, Lei Song, Jie Gao, Hua Qu, Yuhui Wang, Ying Zhang, and Dazhuo Shi. Dldti: A learning-based framework for identification of drug-target interaction using neural networks and network representation. *bioRxiv*, 2020. I

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. I, II-B

[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. I

[10] Shuo Zhang, Yang Liu, and Lei Xie. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. *arXiv preprint arXiv:2011.07457*, 2020. I

[11] Sergi Abadal, Akshay Jain, Robert Guirado, Jorge López-Alonso, and Eduard Alarcón. Computing graph neural networks: A survey from algorithms to accelerators. *arXiv preprint arXiv:2010.00130*, 2020. I

[12] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation?, 2021. I

[13] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *arXiv preprint arXiv:2007.02835*, 2020. I

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. II-C

[15] Jia Li, Yu Rong, Hong Cheng, Helen Meng, Wenbing Huang, and Junzhou Huang. Semi-supervised graph classification: A hierarchical graph perspective. In *The World Wide Web Conference*, pages 972–982, 2019. II-D

[16] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005. III-A

[17] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018. III-A

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. III-C