STUDY ON THE PREDICTION OF PROTEIN ALLOSTERIC SITES USING GRAPH NEURAL NETWORKS

Hongfei Wu
College of Chemistry and Molecular Engineering
Peking University
Beijing
1900011764@pku.edu.cn

ABSTRACT

Allostery is a fundamental process in regulating protein activities, which is frequently reduced by changes in three-dimentional structure. A critical step in allosteric durg design is to identify the allosteric sites of target protein. Here, we study the application of graph neural network structure in residue level protein graph and its capacity on allosteric sites prediction task. On this basis, we propose to use DS-Net structure to realize the hidden representation of protein residues. The experimental results show that our model is more robust and accurate than simple GCN, and can help predict protein allosteric sites partly.

***K*eywords** Graph neural network · Residue level graph · Allosteric sites · Orthosteric sites · DS-Net

1 Introduction

Allosteric effect is an effect that involves multiple aspects of protein's property, resulting in conformational or dynamical changes of the protein, which frequently leads to the changes of binding activity of specific molecules[1, 2]. Activating or inhibiting target protein with less sides effect, the allosteric drugs could be more selective and less toxic than other types of drugs.

The prediction of allosteric sites is helpful for drug development. Generally, there are two ways of approaching the prediction of allosteric sites. (1) without considering the information especially the communication with orthosteric sites, identifying allosteric sites individually from protein structure information. (2) exploring the pathways between orthosteric sites and allosteric sites[3]. Considering the target of prediction, the two situations (with or without orthosteric sites information) are both foreseeable in real-world circumstances. Based on information without orthosteric sites, like topological and physicochemical characteristics of allosteric sites and non-allosteric sites, some methods have been proposed to approach the target[4, 5]. Normally, changes in three-dimensional structure always play an significant role in such allosteric regulation, which might be discovered by many theoretical analysis, in order to achieve the prediction task when allosteric site information is known. From the inspiration based on the structure, many theoretical methods have been developed to discover the allosteric sites from a given protein, such as normal mode analysis[6], anisotropic network model analysis[7, 8], molecular dynamics simulations[9]. These studies demonstrated the feasibility of allosteric sites prediction models which are related to pocket features and protein dynamics.

Another potential way to describe the three-dimensional structure information of proteins is to construct graph structure of proteins. Recently, atomic graph on protein has been used to discover the allosteric domain[10]. Additionally, residue graph is also an alternative scheme to construct protein graph structure, and has been invovled in some other prediction task of protein attributions[11].

Benefiting by the continuous enrichment and improvement of the allostery database[12, 13, 14] and the fast development of graph neural networks methods[15, 16] in recent years, deep learning methods based on data sets, especially graph neural networks, have been developed for this prediction task. For example, PASSer[17, 18] is a classification model to process accurate prediction of protein, which contains graph convolution network as a submodel.

Table 1: Mark description

Mark	Description	Size
i	The layer index (from interior to surface)	-
s	Number of layers	-
f_l	The feature dim in the l-th hidden dim (not for specific layer)	-
N	Number of residues (not for specific layer)	-
N_i	Number of residues in layer i, from 0 to d-1	-
o	Number of residues in orthosteric sites	-
a	Number of residues in allosteric sites	-
C	Input one-hot encoding vector	$\mathbb{R}^{N\times 21}$
H^l	The l-th hidden layer (not for specific layer)	$\mathbb{R}^{N\times f_l}$
W^l	Learning weights matrix of l-th hidden layer	$\mathbb{R}^{f_l\times f_{l+1}}$
$\tilde{A}$	Adjacency matrix with self-loop	$\mathbb{R}^{N\times N}$
$\tilde{A}'_i$	Adjacency matrix between layer i and layer i+1	$\mathbb{R}^{N_i\times N_{i+1}}$
$\tilde{D}$	Degree matrix with self-loop	$\mathbb{R}^{N\times N}$
d_{out}	Degree vector of layer i receiving from layer i-1	$\mathbb{R}^N$
d_{in}	Degree vector of layer i receiving from layer i+1	$\mathbb{R}^N$
e	Edge attribution vector	$\mathbb{R}^3$
σ	Nonlinear activation function	-
FF	Feed forward networks	-

In this study, we attempt to uncover the allosteric sites from residue-level graph structure of protein. We first calculate the residue depth to seperate the residues into several parts according to the msms algorithm[19]. Then, we construct a graph convolution network in each parts, and combine them with a aggregation operator, which named as DS-Net(deep surface networks) architecture. Further more, we apply an individual graph convolution network in surface part and predict the label of each residue to illustrate if the residue belongs to an allosteric site. To get more guides for the capacity of residue-level graph towards this task, we also apply some other graph neural network architectures. From our results, DS-Net seems to have more robust capacity than simple graph convolution network when catching the long-range connection between different residues and performs better on the prediction task.

2 Marks

For simplicity, we agree to use the following notation below. Each mark and its corresponding meaning are shown in Table 1. For the expressed tensor, we also label its dimension characteristics.

3 Methods and Materials

3.1 Allosteric protein datasets

235 X-ray crystal structures of allosteric proteins were downloaded from the ASBench[20] database. Following the work of Nan Wu et al.[21], we use the selected pdb files to generate our dataset, where experimentally determined orthosteric and allosteric site residues for these proteins were attained from ASD Release 4.10.7. Note that we only randomly select 25 of these proteins in 118 to train our model. Firstly, we clear the binding ligands and water molecules in the pdb files. Then, the residues bond to the substrate molecules in the corresponding complex pdb file are selected as the residues of allosteric sites or orthosteric sites. After labeling each residue, all allosteric residues are selected and the training set and test set are divided in a 4:1 ratio. Each data set is divided into positive samples and negative samples in a 1:1 ratio. To make the dataset more reasonable, we randomly select the residues with the nearest atomic distance greater than 3Å on the generated surface as the negative sample.

3.2 Layer partition

To generate different hidden weights for different layers' residues and do prediction only on the surface layer, we divide the residues into several layers by different residue depth. The residue depth is calculated by MSMS algorithm using the accessible program attained from mgltools on the website `https://ccsb.scripps.edu/mgltools/#msms`[19],

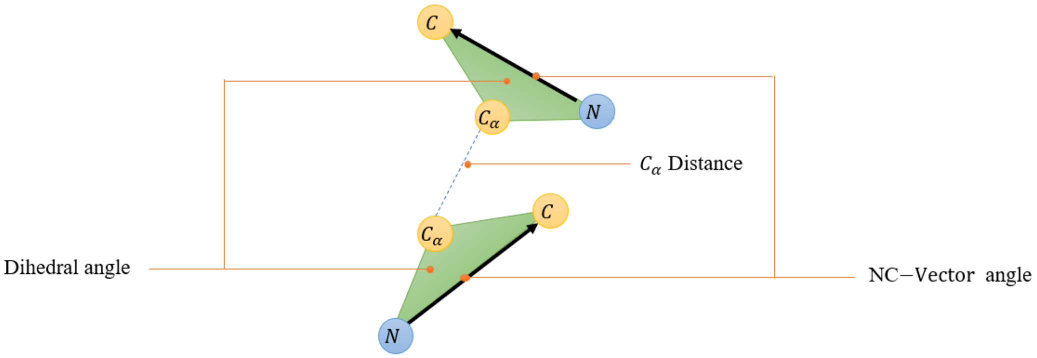


Figure 1: The edge attribution includes three numeric value: dihedral angle between two residues' backbones; NC-vector angle between two residues; C_α distance between two residues' α -C atoms.

and the api can be obtained in the python package Biopython[22]. Note that we have to promise our models to work correctly on different proteins, so it is necessary to set a fixed layer division method for general proteins.

After seperate the residues in s parts, we adopt the following criteria to determine whether the node is adjacent, where each residue is regarded as a graph node (1):

$$A_{uv} = \begin{cases} 1 & \text{if u, v has an atom pair within 3Å} \\ 0 & \text{else} \end{cases} \tag{1}$$

To get more detailed information of protein three-dimensional structure, we also compute some parameters to identify the edge attribution (2). The geometric relationship is shown in the figure 1.

$$A_{uv} = \begin{cases} e & \text{if u, v has an atom pair within 3Å} \\ None & \text{else} \end{cases} \tag{2}$$

In order to implement the message passing between adjacent layers, we also build up the layer adjacent matrix A' to record the connection between adjacent layers in the same rules as matrix A.

3.3 Residue descriptor and labels

One-hot encoding is used to generate the input feature for each residue. For example, the i-th layer input is C_i. The featrues will process in the after steps.

We label all residues which discovered consist in known allosteric sites with 1 and the positive with 0. Such forms of tags enable the prediction task to be transformed into a node classification task on a graph network.

3.4 Graph information extraction

We apply two kinds of graph neural network to carry on the graph information extraction: GCN(Graph convolution network) and GAT(Graph attention network). These models are used to generate the embedding layer representation. Our model implementation is based on python library torch and torch geometric[23].

3.4.1 Graph convolution network[24]

Graph convolution network is a widely used model to proceed message aggregation. It follows the equation (**??**). The meanings of each mark is shown in section 2.

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{l+1}W^{l}) \tag{3}$$

This approximate normalization method is the approximation of spectral convolution under Chebsev polynomial.

3.4.2 Graph attention network[25]

When considering the edge attribution, GCN is not thoroughly a proper way to do graph information extraction because of the requirement for single channel adjacency matrix. To enable the edge attribution, we apply the GAT structure to approach this target. GAT learns aggregation weight for each neighbor of the host node, which depends on the features from both host and neighbor (4). Edge attribution also could be treated as the edge feature and becomes the learning weights source vector as shown in (5).

$$e_{ij} = Linear([Wh_i \| Wh_j]), j \in \text{neighbors} \tag{4}$$

$$e_{ij} = Linear(e), j \in \text{neighbors} \tag{5}$$

$$\alpha_{ij} = \frac{exp(leakyReLU(e_{ij}))}{\sum_{k \in \text{neighbors}} exp(leakyReLU(e_{ij}))} \tag{6}$$

Here $\|$ represents concatenate the two vectors in the first dim.

3.4.3 DS-Net structure

Due to the unreliable capacity in transfer learning on different protein graphs of simple graph convolution networks or graph attetion networks, we propose a new method to establish the graph neural network learning on different proteins. The core idea is that the residues in the layer have more similar solvent depth, so they tend to have similar implicit distribution. The residues of different layers should have a unique logic when the message is transmitted. Intuitively speaking, allosteric sites formed by surface residue interaction have near locality, while allosteric sites formed by internal residue interaction have long correlation, which is consistent with our model. Our layer-to-layer message passing could be represented as (7, **??**), which includes outward aggregation (7) and inward aggregation (8).

$$H_i = d_{out}^{-1} A_i^{'T} H_{i-1} W_{out,i} \tag{7}$$

$$H_i = d_{in}^{-1} A_i' H_{i+1} W_{in,i} \tag{8}$$

Here d^{-1} represent calculate the reciprocal for each value in the first dim. The complete network architecture shows in figure 2.

3.5 Loss function

As we have made the task into a binary classification task, we use the cross entropy loss to train our model (9).

$$Loss = -(y \cdot log(\hat{y}) + (1-y) \cdot log(1-\hat{y})) \tag{9}$$

3.6 Performance metrics

For binary classification, the results can be evaluated using a confusion matrix (Table 2).

Table 2: Binary classification results in a confusion matrix

	Real positive	Real negative
Predicted positive	TP(True positive)	FP(False positive)
Predicted negative	FN(False negative)	TN(True negative)

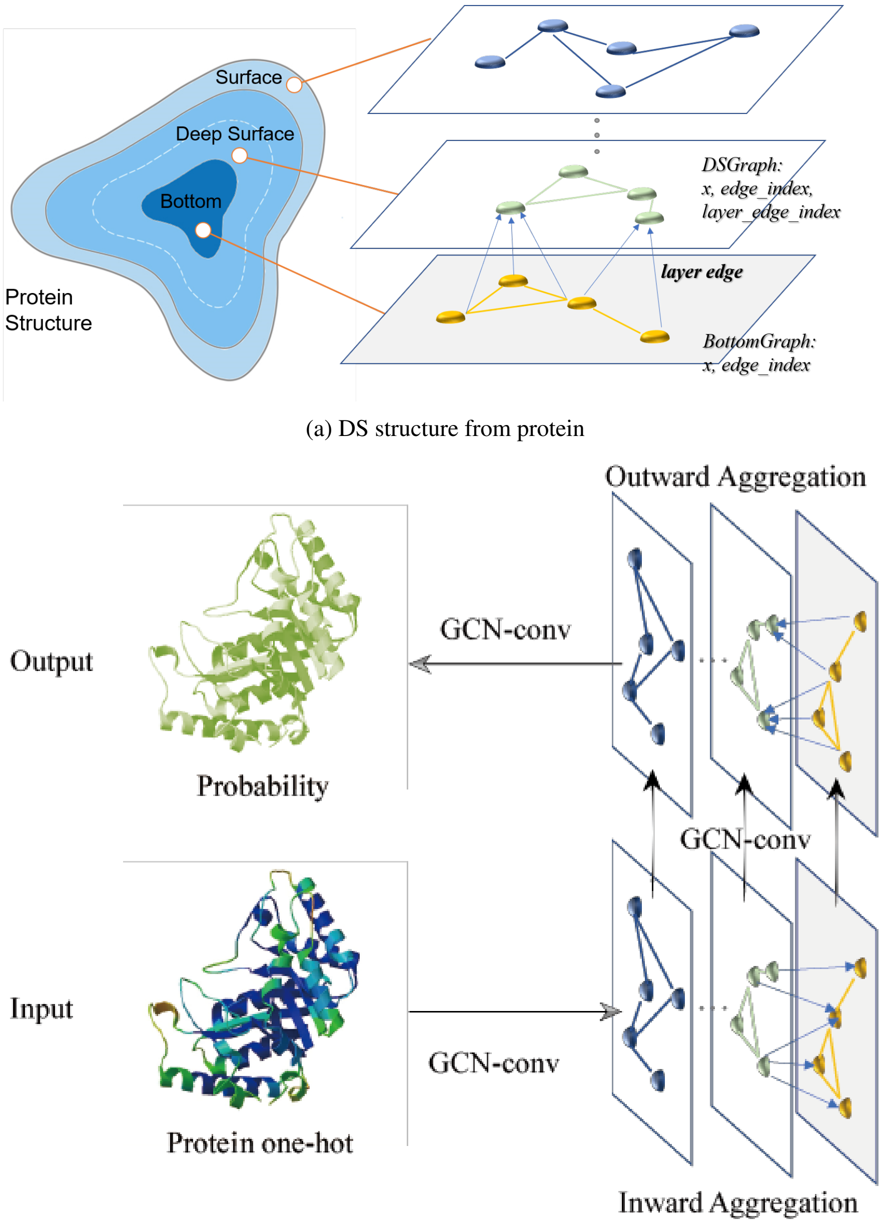
Various indicators could be constructed based on the confusion matrix to quantify the model performance, here we mainly calculate the recall(11)(measures how well the model can predict real positive labels) and precision(10)(measures the ability to classify true-positive and true-negative) and the F1-score(13) based on them and then we draw the ROC (receiver operating characteristic curve) , where the horizontal axis is FPR(12) and the vertical axis is TPR(11), to evaluate our model. Note that we are proceeding the residue-level prediction task, it is more difficult than pocket-level prediction, but soon we will show the advantages that our method might give more information about the allosteric sites. In the other hand, it is also feasible to count the predicted positive residue as a score of an allosteric sites.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{11}$$

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{13}$$



(a) DS structure from protein

(b) DS-Net

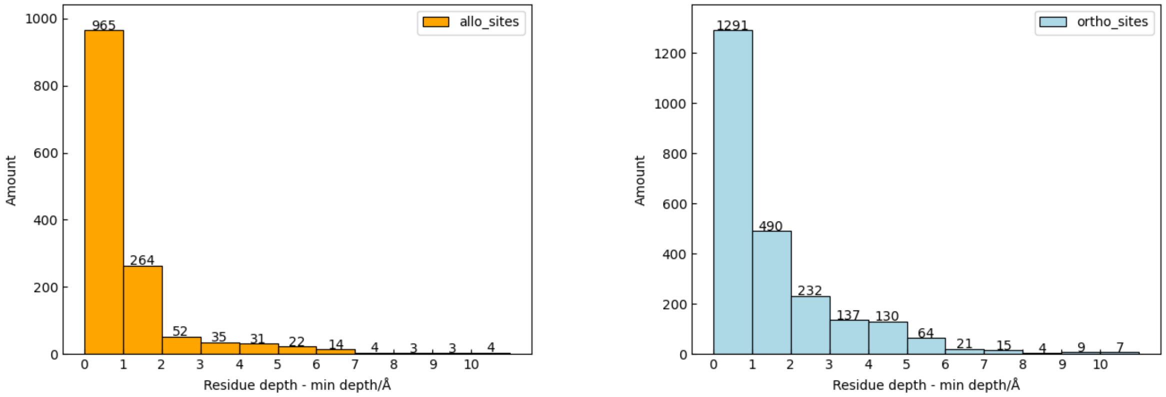
Figure 2: The construction of DS-Net and its workflow: 2a the protein residues are divided by pre-calculated residue depth into several layers. Each layer is treated as a subgraph and connected with a neighbor layer; 2b the DS-Net receives one-hot encoding tensor $C_0, C_1, ..., C_i$ as input. Then, first GCN generate the embedding representation of each node. DS-Net works following with inward aggregation, graph convolution and outward aggregation. Finally the last GCN is applied to predict the final probability.

4 Results and Discussions

4.1 Statistics of residue depth

We study the distribution of residue depth in all proteins in the dataset of 118 proteins(firgure 3). For the residue depth, we take the calculated minimum depth as the zero point (generally, the calculated minimum depth is exactly 0). The statistical results show that the vast majority of allosteric or orthostructural site residues are in the part with deep depth, so reasonable selection and division can retain enough residues on the surface layer of the protein as the sites to be identified.

In addition, it is noted that the residue depth of normal sites tends to be deeper, which may be ascribed that the data and proteins are obtained by deleting ligands and water molecules from allosteric complexes, and some normal sites may be bound only under the induced fit.



(a) Allosteric sites depth distribution

(b) Orthosteric sites depth distribution

Figure 3: The distribution of residue depth on sites, allosteric and orthosteric sites follow a same distribution approximately. 3a distribution of allosteric sites; 3b distribution of allosteric sites.

4.2 Surface depth

Note that our model only predicts allosteric label on surface layer, so it is greatly significant to choose a proper parameter to divide the surface residues.

Based on this, we calculate the distribution of the maximum depth of each site in all residues in the data set, and the proportion of the depth of each site in all residues. We can see from the figure 4a, figure 4b that the residues whose depth is less than the maximum depth of site residues reach more than 90%, and the proportion of site residue depth in all residues is almost evenly distributed. Therefore, using simple depth division can not effectively separate the potential site residues and other residues on the surface.

In order to find a suitable partition index, we further calculate the regularization depth as 14.

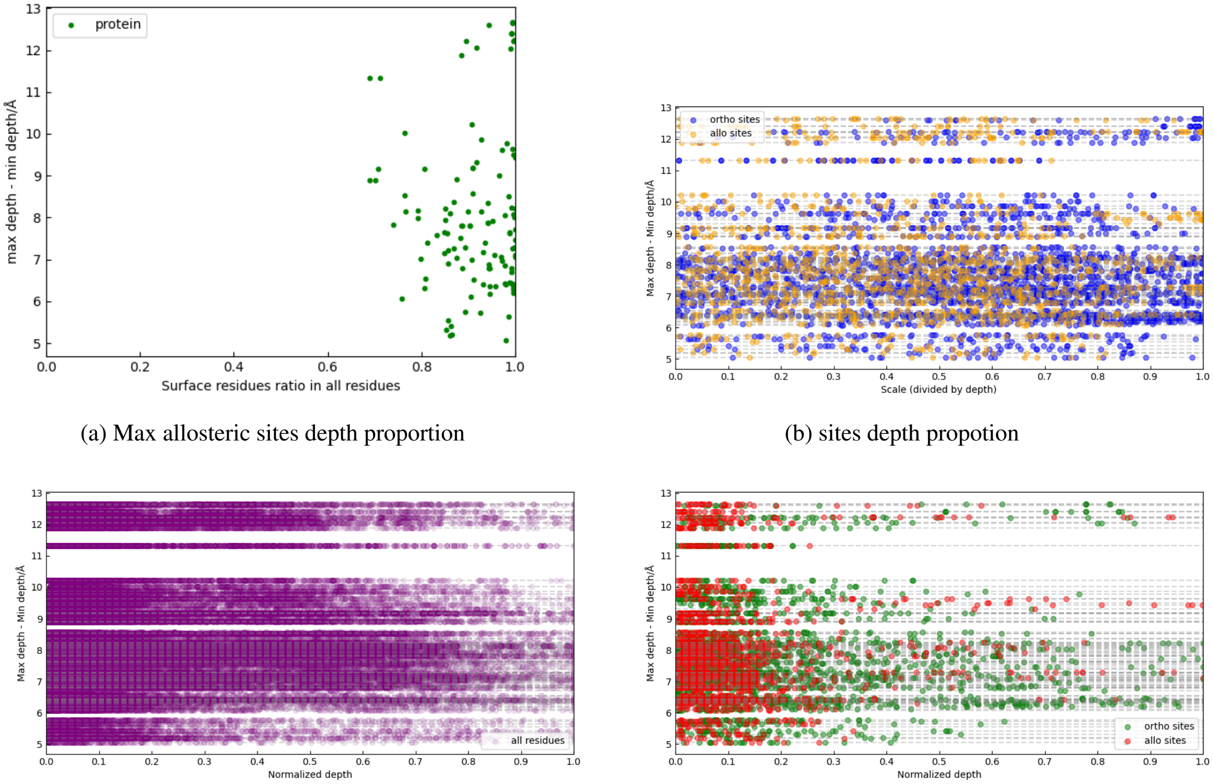
$$\text{Normalized depth} = \frac{d - (mindepth)}{(maxdepth) - (mindepth)} \tag{14}$$

According to the distribution map of normalized depth, we find that site residues are concentrated in the part with less than 0.3 normalized depth, while for all protein residues, a considerable part remains in the part with more than 0.3 normalized depth. Therefore, we use normalized depth value 0.3 to divide the protein surface layer.

4.3 Prediction performance

We trained our models on the generated dataset. As shown in 5, after training 600 epochs, the simple GCN model is still difficult to converge to a better result. Our model converges to about 90% accuracy on the training set stably at about 150 epochs, and achieves about 70% accuracy on the test set. At the same time, our model has higher AUC than simple GCN.

Under the assumption of GCN homogeneity, the residue tags are distributed more intensively and can correspond to certain pockets. At the same time, as shown in 5e, residue level labels are helpful to discover hidden allosteric sites. Taking PDB 3H6O as an example, the prediction of residue level is basically concentrated near the pocket, and error tags of some residues can be simply eliminated by the pocket recognition algorithm.



(a) Max allosteric sites depth proportion

(b) sites depth propotion

(c) Normalized depth of all residues

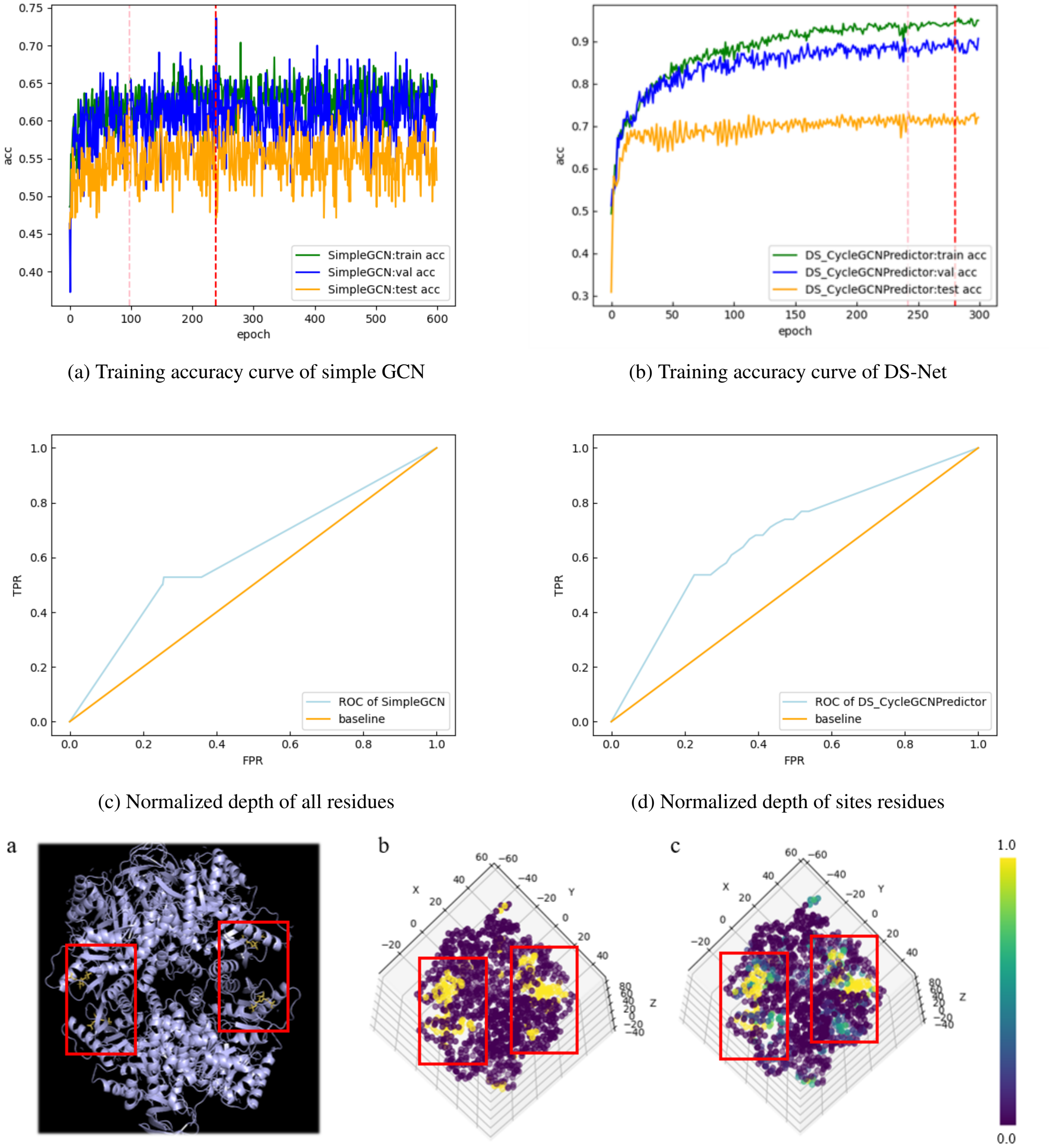
(d) Normalized depth of sites residues

Figure 4: The statistics of sites residue depth proportion. 4a the proportion of maximum residue depth of sites; 4b all proportion of residue depth of sites; 4c normalized depth of all residues in different proteins; 4d normalized depth of sites residues in different proteins.

5 Conclusion

Several machine learning-based methods have been developed for allosteric site prediction over the past few years. Considering the latent graph data structure of proteins, our study starts from residue-level graph structure, and applies some graph neural networks to learn the relationship between protein structure and allosteric sites. Our model has an effective accuracy on prediction task, and residue-level prediction allows our model to get better performance when combining with some pocket recognition algorithms.

As a prospect, our model is expected to further add the prediction function based on normal site information. At present, the main idea of this related implementation is the combination of neural network based on attention structure and DS Net. The use of normal site information is expected to improve the ability of the model in specific prediction tasks.



(a) Training accuracy curve of simple GCN

(b) Training accuracy curve of DS-Net

(c) Normalized depth of all residues

(d) Normalized depth of sites residues

(e) The prediction of PDB_ID: 3H6O.(a)Allosteric complex structure; (b)binary result; (c)probability-hot result.

Figure 5: The prediction model performance on selected dataset. 5a, 5c shows the performance of simple graph convolution network. It is obvious that simple GCN has difficulty in converging a well performance. 5b, 5d shows better performance of our model than simple GCN; 5e prediction results of 3H6O as an example. The homogeneity assumption of GCN makes the distribution of residue labels concentrated.

Acknowledgments

The research is supported by Peking University President Fund.

Thank my tutor Liu Zhirong for guiding my research. Prof. Liu Zhirong's suggestions gave me adequate ability and motivation to complete this project.

References

[1] Kannan Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure*, 57, 2004.

[2] Bharath Srinivasan, Farhad Forouhar, Arpit Shukla, Chethana Sampangi, Sonia Kulkarni, Mariam Abashidze, Jayaraman Seetharaman, Scott Lew, Lei Mao, Thomas B. Acton, Rong Xiao, John K. Everett, Gaetano T. Montelione, Liang Tong, and Hemalatha Balaram. Allosteric regulation and substrate activation in cytosolic nucleotidase ii from legionella pneumophila. *The FEBS Journal*, 281, 2014.

[3] Xavier Daura. Advances in the computational identification of allosteric sites and pathways in proteins. *Advances in experimental medicine and biology*, 1163:141–169, 2019.

[4] Wenkang Huang, Shaoyong Lu, Zhimin Huang, Xinyi Liu, Linkai Mou, Yu Luo, Yanlong Zhao, Yaqin Liu, Zhongjie Chen, Tingjun Hou, and Jian Zhang. Allosite: a method for predicting allosteric sites. *Bioinformatics*, 29 18:2357–9, 2013.

[5] Ava S.-Y. Chen, Nicholas James Westwood, Paul D Brear, Graeme W. Rogers, Lazaros Mavridis, and J. B. O. Mitchell. A random forest model for predicting allosteric and functional sites on proteins. *Molecular Informatics*, 35, 2016.

[6] Alejandro Panjkovich and Xavier Daura. Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics*, 13:273 – 273, 2012.

[7] Le Yan, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Architecture and coevolution of allosteric materials. *Proceedings of the National Academy of Sciences*, 114:2526 – 2531, 2017.

[8] Miao Yu, Yixin Chen, Zi-Le Wang, and Zhirong Liu. Fluctuation correlations as major determinants of structure- and dynamics-driven allosteric effects. *Physical chemistry chemical physics : PCCP*, 21 9:5200–5214, 2019.

[9] Élodie Laine, Christophe Gonçalves, Johanna C. Karst, Aurélien Lesnard, Sylvain Rault, Wei-Jen Tang, Thérèse E. Malliavin, Daniel Ladant, and Arnaud Blondel. Use of allostery to identify inhibitors of calmodulin-induced activation of bacillus anthracis edema factor. *Proceedings of the National Academy of Sciences*, 107:11277 – 11282, 2010.

[10] Nan Wu, Léonie Strömich, and Sophia N. Yaliraki. Prediction of allosteric sites and signaling: Insights from benchmarking datasets. *Patterns*, 3, 2021.

[11] Charles Abreu Santana, Sabrina de Azevedo Silveira, João P. A. Moraes, Sandro Carvalho Izidoro, Raquel de Melo-Minardi, António J. M. Ribeiro, Jonathan D. Tyzack, Neera Borkakoti, and Janet M. Thornton. Grasp: a graph-based residue neighborhood strategy to predict binding sites. *Bioinformatics*, 36 Supplement_2:i726–i734, 2020.

[12] Zhimin Huang, Liang Zhu, Yan Cao, Geng Wu, Xinyi Liu, Yingyi Chen, Qi Wang, Ting Shi, Yaxue Zhao, Yuefei Wang, Weihua Li, Yixue Li, Haifeng Chen, Guoqiang Chen, and Jian Zhang. Asd: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Research*, 39:D663 – D669, 2011.

[13] Wenkang Huang, Guanqiao Wang, Qiancheng Shen, Xinyi Liu, Shaoyong Lu, Lv Geng, Zhimin Huang, and Jian Zhang. Asbench: benchmarking sets for allosteric discovery. *Bioinformatics*, 31 15:2598–600, 2015.

[14] Kun Song, Jian Zhang, and Shaoyong Lu. Progress in allosteric database. *Advances in experimental medicine and biology*, 1163:65–87, 2019.

[15] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4–24, 2019.

[16] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *ArXiv*, abs/1812.08434, 2020.

[17] Hao Tian, Xi Jiang, and Peng Tao. Passer: prediction of allosteric sites server. *Machine Learning: Science and Technology*, 2, 2021.

[18] Sian Xiao, Hao Tian, and Peng Tao. Passer2.0: Accurate prediction of protein allosteric sites through automated machine learning. *Frontiers in Molecular Biosciences*, 9, 2022.

[19] Michel F. Sanner, Arthur J. Olson, and Jean-Claude Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38 3:305–20, 1996.

[20] Wenkang Huang, Guanqiao Wang, Qiancheng Shen, Xinyi Liu, Shaoyong Lu, Lv Geng, Zhimin Huang, and Jian Zhang. Databases and ontologies asbench : benchmarking sets for allosteric discovery. 2015.

[21] Xinyi Liu, Shaoyong Lu, Kun Song, Qiancheng Shen, D. Ni, Qian Li, Xin heng He, Hao Zhang, Qi Wang, Yingyi Chen, Xinyi Li, Jing Wu, Chunquan Sheng, Guoqiang Chen, Yaqin Liu, Xuefeng Lu, and Jian Zhang. Unraveling allosteric landscapes of allosterome with asd. *Nucleic Acids Research*, 48:D394 – D401, 2020.

[22] Peter J. A. Cock, Tiago R. Antão, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczyński, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422 – 1423, 2009.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[24] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2017.

[25] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2018.