**Caiya Zhang**

**11.23 Weekly Report**

**What you have done in the past week:**

1. I’ve successfully tried Pros\_gnn with S3421 (contains 3421 experimentally determined mutations from 150 proteins) and S2648 (includes 2648 single-point mutation in 131 different globular proteins) datasets, which are both given by the author. I think this can serve as a baseline.

|  |  |  |
| --- | --- | --- |
|  | **S3421** | **S2648** |
| **Direct mut** | ***r*** = 0.69 ， ***σ*** = 1.73 | ***r*** = 0.62 ，***σ*** = 1.11 |
| **Reverse mut** | ***r*** = 0.71，***σ*** = 1.69 | ***r*** = 0.60, ***σ*** = 1.12 |

1. I’m still learning about some popular graph transformers and trying to build one based on these projects and the recipe I mentioned last week. It may be a little bit difficult to build a transformer network from scratch according to the recipe, so I’m also referring to some established ones.
2. I’ve started writing the “First Check-in” progress report. I think there is not much to change in the Abstract part for now, except for the grammatical mistakes that I’ve already modified. I’m mainly focusing on the Technologies and Challenges parts mentioned in the assignment requirement.

**What are major challenges/issues you need to discuss in the meeting:**

I think there’s nothing having to be discussed in this week, and I’ll continue to try a transformer.

**What you want to accomplish in the next week/weeks:**

1. I’ll continue examining if a transformer will be suitable for my work and trying to build one;
2. I’ll complete a daft progress report this week and submit before next weekly meeting.

**11.16 Weekly Report**

**What you have done in the past week:**

1. Since we mentioned before that I can consider using a transformer, I followed an article this week that gives a recipe on how to build a general, powerful, scalable Graph Transformer. I'm browsing the source code and the paper to see if there’s any parts I can learn and apply.
2. Considering Yan's very pertinent suggestion, I decided to adopt a supervised learning strategy. This way, we don’t need very large amount of data can be accomplished and the training time is more limited, making it easier to adjust and modify.
3. I’m running S4169 and M1101 on GeoPPI, which are two of the datasets I plan to use as baseline data. Also, I’m still adjusting the configuration of Pros-GNN in my end, if possible, I still want to try the Pros-GNN, because this code structure is very simple and may have great reference significance.

**What are major challenges/issues you need to discuss in the meeting:**

I think there’s not many questions this week.

**What you want to accomplish in the next week/weeks:**

1. Yan just told me that her installation of Pros-GNN went smoothly, so it may be a problem with my local configuration. I'll try again later after reinstalling and then run the baseline within the week.
2. I’m going to draw a workflow diagram of my model out next week.

**11.9 Weekly Report**

**What you have done in the past week:**

1. Last week, I tried the GeoPPI code in its entirety. After referring to the code of this project, I also started to try to build a framework myself, using a simplified network structure. There are still some bugs in it, but overall, I think it's going relatively well. The core network of this code is still a simplest convolutional network, which is just temporary used to make the workflow run, and my final approach may consider other kinds like graph attention networks, autoencoder, and etc.
2. Because of the test and assignments due this week, I didn't get too much done this week, for which I’m so sorry about that.

**What are major challenges/issues you need to discuss in the meeting:**

I think not much need to be discussed in the meeting. The problems I encountered this week are mainly some trivial errors in my code and concerns about which specific kind of network to use and what kind of learning strategy to be implemented. These may be solved or decided after I tried more and discussed with Yan.

**What you want to accomplish in the next week/weeks:**

1. Discuss with Yan about specific learning strategies I might implement, such as whether I need to add pre-training and which specific types of GNN networks are worth trying.

**11.2 Weekly Report**

**What you have done in the past week:**

1. I briefly browsed the code for GeoPPI, and tried some given sample complexes, where the input .pdb files, mutation information and the two interaction partners are all given by the author. From this project, I’ve learned some about reading and processing .pdb files and usage of foldx, which I think maybe useful for me.
2. I’m still setting up the environment for Pros-GNN, because something went wrong with installing the Rosetta. I browsed the code on github, and I briefly searched some more about the embedder they used. Their method looks fitting our requirements for finding a baseline model, just as Yan said.

**What are major challenges/issues you need to discuss in the meeting:**

From my perspective, I think I have questions need to be discussed. I just met some problems about running the Pros-GNN, but I think I can handle it.

**What you want to accomplish in the next week/weeks:**

1. I may try the Pro-GNN this week, and if possible, I would like to compare the prediction results and speed of GeoPPI and Pros-GNN.
2. I’m going to learn more about graph transformer, which was mentioned last week.

**10.26 Weekly Report**

**What you have done in the past week:**

1. Prepare a summary report about the two articles. I have analyzed and summarized mainly from the aspects of Motivations, Methods, Data, Results, Evaluation, and Limitations. During the in-depth research and summary process, I did get some inspirations, including some new approaches in evaluation and the possibility of using a label-free learning strategy.
2. I browsed some unsupervised and self-supervised learning frameworks for graph neural networks processing 3D models and got some insight into the common techniques involved, which I may be able to use in my own model.

**What are major challenges/issues you need to discuss in the meeting:**

From my perspective I'm clear so far, but I have a little concern about the progress. I wonder if it's the time to start thinking about the code.

**What you want to accomplish in the next week/weeks:**

1. Accomplish the new tasks presented in today's meeting.
2. If necessary, I’m going to try to run some of the projects that provide the code, especially run on the benchmark datasets and get some baseline results.

**10.19 Weekly Report**

**What you have done in the past week:**

1. I finished the final draft of my proposal this week, and I plan to submit it after this week's meeting. Since the instructor has postponed the proposal submission deadline until Nov 7th and we can resubmit in unlimited times, I can still revise it if there’s any change of the main approach.
2. I'm preparing the report on the two key papers. This week I'm mostly combing through the proposed framework in “plos comb”, and I may need some more time on the other one.
3. I'm browsing some GNN frameworks that focus on interpretability, and I'm still studying about how to implement and evaluate a interpretable method.

**What are major challenges/issues you need to discuss in the meeting:**

Regarding the interpretability of the pre-training, I am looking at some data visualization methods, such as Maximum Mean Discrepancy (MMD), and I wonder if this is a viable idea for the database I will use. Of course, I am also learning about this, and I’m learning more about methods to achieve interpretability in the middle or later stages of training. I wonder if there’s any relevant suggestions.

**What you want to accomplish in the next week/weeks:**

1. Continue in-depth reading the two papers and focus more on the “nature method” to complete the report slides on the interpretation of the two articles.
2. I may need to schedule a meeting with Yan to make sure my understanding is accurate and to see if there is any other paper that may has significant instruction to my model.

**10.12 Weekly Report**

**What you have done in the past week:**

1. This week, I mainly focused on completing the draft proposal. During the process of writing proposal, I sorted out several goals that I need to achieve in order to complete the project, and I’m a little bit clearer about the overall time schedule.

2. I have quickly read over the two articles on Plos Computational Biology and Nature Methods. So far, I think GeoPPI can be an alternative to the baseline method or project for comparison, and the SKEMPI 2.0 dataset mentioned in several papers may also be an alternative to the baseline dataset.

**What are major challenges/issues you need to discuss in the meeting:**

I wrote my draft based on structure of the sample given by the instructor, combined with the proposal I received, and I found there is no module that requires me to describe the method and data sets in the given outline, which was what I might need to show from my perspective. So far, I have only mentioned them in the Abstract and Objectives, which are relatively brief and general, so I was wondering if I need to explain more about methods and datasets I may use in this project.

**What you want to accomplish in the next week/weeks:**

1. Continue to modify and improve my proposal.
2. Continue doing literature review and collect some ideas about choosing specific graph neural network structure in my work.
3. Carefully understand the feature extraction methods in the two key papers, as well as learn more about the use of AlphaFold database.

**Reference**

Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J., Fernández-Recio, J., & Moal, I. H. (2018). SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, *35*(3), 462–469. https://doi.org/10.1093/bioinformatics/bty635

Liu, X., Luo, Y., Li, P., Song, S., & Peng, J. (2021). Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLOS Computational Biology*, *17*(8). <https://doi.org/10.1371/journal.pcbi.1009284>

Tubiana, J., Schneidman-Duhovny, D., & Wolfson, H. J. (2022). Scannet: An interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, *19*(6), 730–739. <https://doi.org/10.1038/s41592-022-01490-7>

**10.5 Weekly Report**

**What you have done in the past week:**

1. Based on last week's session, this week I first continued my understanding of the mCSM -PPI2 [1] dataset.
   1. I reviewed the modeling approach of the original mCSM [2] project and supplemental files of mCSM-PPI2, which are both predecessors of the target project, and specifically understood the idea of "Graph-based structural signatures" construction relied on in these projects.
   2. Then I focused back to the mCSM-PPI2 paper, to understand the six features presented in “Modelling effects of mutation” section. With previous understanding of mCSM-PPI2 data, I found the graph modeling ideas of above-mentioned projects are invariant, all of them represent atoms as nodes and their interactions as edges, and use pharmacological features to illustrate the effects of physicochemical changes caused by point mutations. Regarding the six newly proposed features in mCSM-PPI2, they are different effects on single point mutations, and are used to combine with their well-established graph-based signatures for regression and classification.
2. I discussed the investigated dataset with Yan in our meeting, and we found, although this project uses a graph-based structure, its most dominant predictive model is still a traditional machine learning algorithm, where there may not much guidance on our GNN-based framework. So we still need to spend some time doing some literature review in the next weeks.
3. Yan has sent me three papers she found on related work, of which I am following one to understand how to generate the graph [3].
4. I was also preparing a short report for presentation in this week's meeting.

**What are major challenges/issues you need to discuss in the meeting:**

1. There is not much concern so far, since I’m still in progress of accumulating knowledge of relevant works.

**What you want to accomplish in the next week/weeks:**

1. Continue studying the papers Yan sent me, while I can also find one or two similar projects that may give me some inspirations.
2. Based on the ideas that have been accumulated so far and the proposal I received at the very beginning, I’m going to write a draft for the formal proposal I need to submit two weeks later.

**Reference**

[1] Rodrigues, C. H. M., Myung, Y., Pires, D. E. V., & Ascher, D. B. (2019). mCSM-PPI2: Predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Research*, *47*(W1), W338–W344. <https://doi.org/10.1093/nar/gkz383>

[2] Pires, Ascher, D. B., & Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics, 30(3), 335–342. <https://doi.org/10.1093/bioinformatics/btt691>

[3] Liu, Luo, Y., Li, P., Song, S., & Peng, J. (2021). Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Computational Biology*, *17*(8), e1009284–. <https://doi.org/10.1371/journal.pcbi.1009284>

**9.28 Weekly Report**

**What you have done in the past week:**

1. This week I focused on the dataset in mCSM-PPI2 and went through the general structure of input data for that project;
2. To understand how to get and prepare similar data in real code, I referred to a github project called GeoPPI, which was recommended by Yan;
3. Tried to retrieve one of the csv files in mCSM dataset (S4169) on Google Colab, and split each row of data according to the method used in GeoPPI;
4. Prepare for the short presentation of describing and handling the data.

**What are major challenges/issues you need to discuss in the meeting?**

1. I wonder is there a specific type of graph neural network is recommended for PPI prediction, such as GCN or CGAT, etc.

**What you want to accomplish in the next week/weeks:**

1. In next week, my plan is to reproduce the data preparation part of mCSM on Colab, following the idea of GeoPPI to obtain data and build graphs, and check the results by printing the matrix;
2. Ready to start writing a formal thesis proposal, which will be due on Oct 20th.

**9.21 Weekly Report**

**What you have done in the past week:**

1. Quickly read over the four papers mentioned in “Literature Review”, and I’ve had a general understanding of the task mentioned in the proposal.
2. A github repository was created to upload weekly report, and this is the repo link: <https://github.com/Caiya-Zhang/gnn_pro_mut>

**What are major challenges/issues you need to discuss in the meeting?**

1. Whether I can develop my project basing on the code of the recommended project, or is there a recommended framework. If I need to start from scratch, I may first build the framework as I'm used to and may modify later.

**What you want to accomplish in the next week/weeks:**

1. For the binding affinity changes equation given in the proposal, I am not particularly sure how this equation is actually applied in code. I may need Frederic's help.
2. Focus on reading and understanding the DATA and METHOD parts of mCSM-PPI2 and MutaBind2. Hopefully, through these, I can **have an initial idea of the form of data I need to use for training and how to apply the objective function**. I hope I can get some inspiration from this.