**Report on model of protein-protein interaction**

Author: Yuanqing Zhou

# Assumptions

The interaction between proteins is closely related to their structures. Graph is used to model structure of protein. The residues are regarded as nodes, establishing edges with neighborhoods.

## Task definition

Provided a set of proteinsand a set of PPI , where  is a binary PPI indicator function that is 1 if the PPI betweenandhas been confirmed, and 0 otherwise, the types of PPI can be represented by the label space with different categories of interactions, and the labels for a confirmed PPI can be represented as. The goal of multi-type PPI learning is to learn a function from the training set such that for any PPI,is the set of predicted labels for.

## Network architecture

First, protein graph was construct a protein graph with residues as nodes and contacts as edges. The physical and chemical properties of amino acids were treat as features of nodes. Then, GCN blocks are employed to characterize the protein graph, and all the output features of nodes are aggregated to a feature vector as a feature of the protein. Finally, products of two protein features were used for protein-protein interaction prediction.

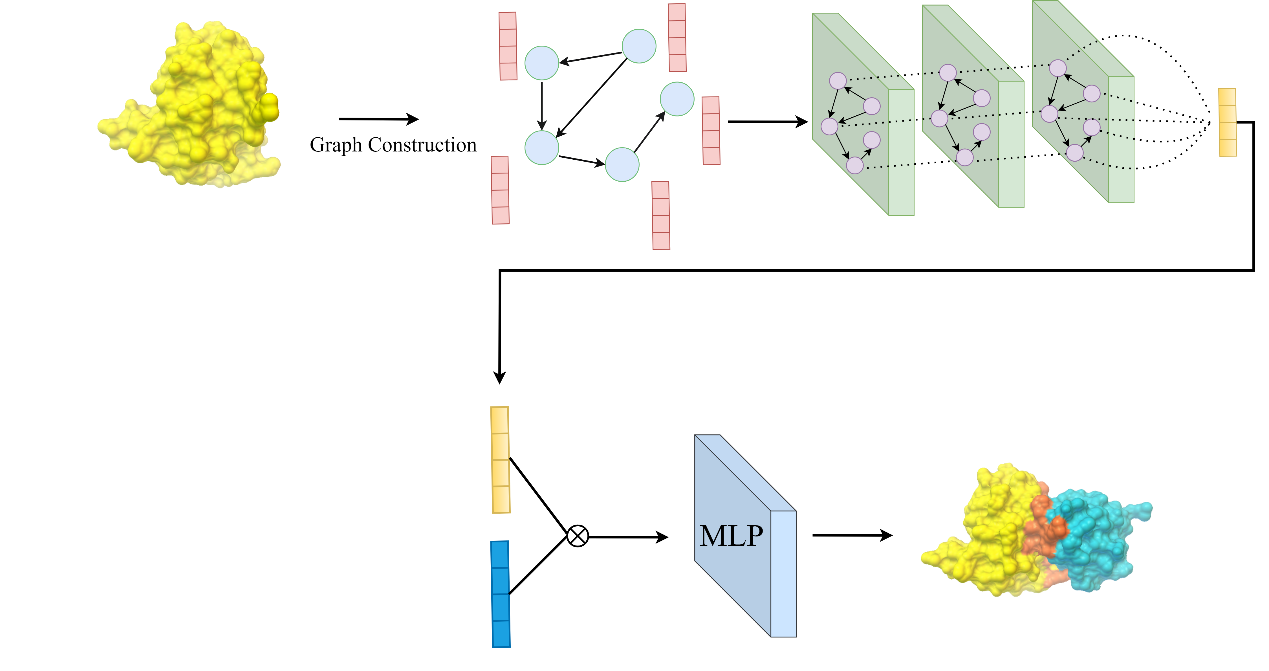


Figure 1 The overall framework

# Questions

## 2.1 What’s your prediction outcome variable?

In this model, prediction outcome variable is a probability vector with the size of 13, representing probability of each category of interactions.

## 2.2 How’s the training cohort constructed?

Due to limited time and power of devices, 10% of PPI data were used in training and validation. The ratio of the training set to the validation set is 8:2.

## 2.3 What are the features used and how are they constructed?

A set of amino acid residues in a protein are represented as . To characterize each residue, we utilize a diverse array of physicochemical properties. For the internal, bottom-up view of a protein, a protein graph  is constructed to capture the relationships between residues in Prot. This graph comprises a set of nodes denoted by, an adjacency matrixof size  that depicts the connectivity between the nodes, and a feature matrixthat encapsulates the properties of all the residues. The contact map is exactly equivalent to the adjacency matrix  in the protein graph. Contact maps are derived from the three-dimensional atomic coordinates of proteins. Firstly, we retrieve the predictive structures from the AlphaFold Protein Structure Database. We then designate the location of each residue by utilizing the three-dimensional coordinates of its alpha-carbon (Cα) atom. The existence or non-existence of contact between a pair of residues is ascertained by their physical distance. In this work, we choose the cutoff distance of 12 Å. In the feature matrix , each row represents a collection of properties of a specific amino acid residue. In this context, we take into account seven residue-specific properties: isoelectric point, polarity, acidity and alkalinity, capacity as a hydrogen bond acceptor, capacity as a hydrogen bond donor, octanol-water partition coefficient, and topological polar surface area.

## 2.4 How do you evaluate the model?

Sum up the TP, FP, FN, and TN for each category, aggregate them to form new TP, FP, FN, and TN values, and then calculate Micro-Precision and Micro-Recall to obtain Micro-F1. Specifically, generate confusion matrices for each category, add these confusion matrices together to obtain a multi-class confusion matrix, and then calculate the F1 score.







## 2.5 What would be the next steps?

1. Use 3D equivariant graph neural networks to extract structural information of proteins
2. Add protein surface information