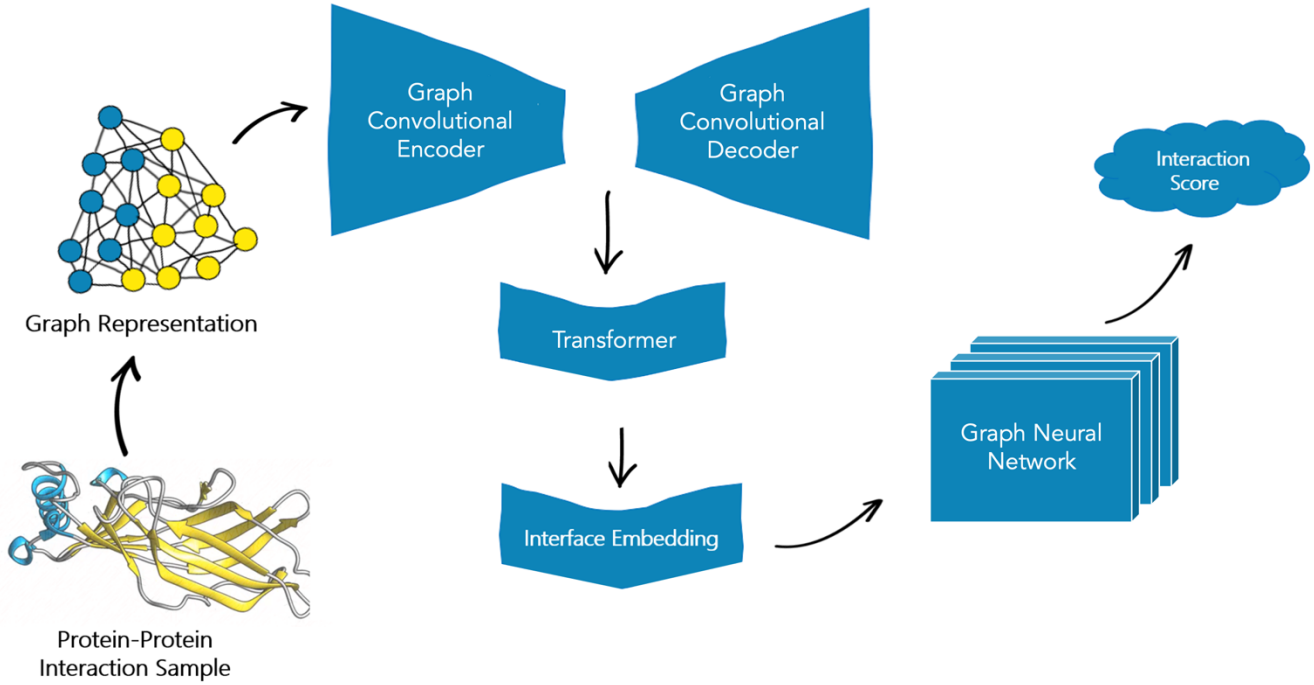
Predicting protein-protein interactions using GCN-based encoder-decoder combined with a transformer

Proteins are large and complex molecules that play countless roles throughout the biological world. They interact through their interfaces to fulfill essential functions in the cell. They bind to their partners in a particular manner and form complexes that highly affect understanding of the biological pathways in which they are involved. Any abnormal interactions may cause interruptions of their functions and may have detrimental effects on the organism. Furthermore, investigating residue interactions on protein-protein interfaces is essential for drug discovery in pharmaceutical research. As experimental data accumulates, artificial intelligence comes to the stage, and recent groundbreaking applications of AI profoundly impact the structural biology field. This talk will present our work on developing a graph-based framework for investigating protein-protein interfaces. Our framework converts protein-protein interfaces from PDB 3D coordinates files into graphs with nodes representing interface residues and edges corresponding to the residue interactions. Each residue is represented as a vector with the features: residue type, polarity, residue charge, relative accessible surface area (ASA), pair potentials, and backbone dihedral angles. Three edges are constructed: sequential edges, radius edges, and K-nearest neighbor edges. The graphs are further provided to a graph convolutional network-based encoder-decoder architecture combined with a transformer to generate embeddings. To train and test the encoder-decoder part, we use a large set that contains more than 500.000 PPIs deposited from PDB. Then, these embeddings are fed to a Graph Interaction Network, which is composed of a succession of graph convolution layers (GCL), non-linear activation (ReLU), and pooling layers. For the downstream task, which is the scoring docking model, we use 271.830 interfaces which are split into training, validation, and test sets. Positive samples come from PDB and PIFACE, and the negative data comes from PPI4DOCK and DOCKGROUND. Our method achieved 0.91 accuracy for the test set and outperformed existing CNN models, which achieved around 0.75-0.81 accuracy for the test set and GNN architectures, such as DeepRank-GNN and GNN-DOVE, whose accuracies are 0.89 and 0.88 respectively.


Graph Convolutional Encoder
Graph Convolutional Decoder
Interaction Score
Graph Representation
Transformer
Graph Neural Network
Interface Embedding
Protein-Protein Interaction Sample