**Predicting Unseen Interactions in KEGG Pathways Using Unsupervised Learning**

Investigator: Yibo Chen

Real-World Task:

This project proposes to identify potential unseen protein-protein interactions within the KEGG pathway maps. The ability to predict such interactions can enhance our understanding of biochemical networks and may lead to the discovery of novel targets for drug development.

Dataset and Source:

The data for this project will be sourced from the KEGG PATHWAY database (https://www.kegg.jp/kegg/pathway.html), which comprises a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks. Given the condition where the prediction result is undesirable, other PPI interaction datasets like STRING (https://string-db.org/) might also be used to provide additional information on PPI interaction. Training, validation, and testing data will be acquired by randomly split KEGG Pathway database based on the pathway topic. An entire pathway will be only used once in training, validation, or testing.

Approach and Novelty:

The project will employ unsupervised learning techniques to learn patterns from known interactions within the KEGG pathways and apply this knowledge to predict unseen interactions in different pathway maps of KEGG.

1. Clustering Algorithms: Apply algorithms like spectral clustering to identify clusters of molecules within pathways that share similar interaction patterns, which may suggest potential new interactions.
2. Dimensionality Reduction: Use PCA and t-SNE to project the high-dimensional space of molecular interactions into a lower-dimensional space for pattern recognition and potential interaction prediction.

Procedure and Timeline:

Week 1 (11/13 – 11/19): Data preparation, including pathway selection, graph construction, and initial exploratory analysis using dimensionality reduction techniques.

Week 2 (11/20 – 11/26): Application of clustering algorithms to discern patterns in the interaction data, followed by the formulation of interaction predictions for unseen pathway maps.

Week 3 (11/27 – 12/3): Validation of predicted interactions against separate KEGG pathway maps not included in the training set and finalization of the project report.

Expected Outcomes:

The project aims to produce a list of predicted protein-protein interactions within KEGG pathways that have not been documented in the database. The success of the predictions will be validated against the KEGG database, providing a direct measure of the project's effectiveness.