Integrating Multi-Omics Data to Enhance Protein-Protein Interaction Predictions Using Graph Autoencoders

Xiaoyu Gui
xgui@ucsd.edu

Siddharth Vyasabattu
svyasabattu@ucsd.edu

Utkrisht Rajkumar
utkrisht96@gmail.com

Thiago Mosqueiro
thiago.mosqueiro@gmail.com

Abstract

Protein-protein interaction (PPI) networks are essential for understanding molecular mechanisms governing biological systems, with applications in drug discovery, functional genomics, and disease pathway identification. However, existing computational approaches for predicting PPIs face limitations due to their reliance on static interaction data, the absence of multi-omics integration, and significant computational costs when applied to large-scale biological datasets. To address these challenges, we present PPI-OMEGA, a Variational Graph Autoencoder (VGAE)-based framework that integrates multi-omics data to enhance PPI prediction. Our model combines RNA expression profiles and immunohistochemistry (IHC) protein expression levels as node attributes in a graph-based learning framework. The encoder in the model architecture efficiently learns biologically meaningful protein embeddings by leveraging graph convolutional layers, while the decoder reconstructs the PPI network, capturing both network topology and molecular context.

To assess model performance, we benchmark our approach against traditional PPI prediction methods and evaluate it using Area Under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP). Results demonstrate that incorporating multi-omics data significantly improves predictive accuracy, with the protein expression-based model achieving the highest AUROC (0.7933) and AP (0.8102), outperforming RNA-only and combined models. Moreover, our framework reduces computational overhead by generating reusable embeddings that facilitate downstream tasks such as disease association studies and functional classification. These findings highlight the potential of graph-based multi-omics integration to improve disease-relevant PPI modeling and support precision medicine applications.

Code: https://github.com/EliteApex/PPI-OMEGA

1 Introduction

Protein-protein interaction (PPI) networks are fundamental to understanding the molecular mechanisms governing biological systems. These networks facilitate key cellular processes, including signal transduction, regulation of enzyme activity, and structural organization. Given their significance, PPIs play a crucial role in diverse domains such as drug discovery, functional genomics, and disease pathway identification. However, computational challenges persist in accurately modeling these interactions due to the large scale of high-confidence datasets (e.g. STRING) and the presence of experimental noise and missing data (Szklarczyk et al. (2023)). These limitations hinder progress in precision medicine, where accurate network-based representations of protein interactions are essential for developing targeted therapies for diseases like cancer.

Traditional computational methods for PPI prediction, such as ExactL3 and ProteinPrompt, have leveraged topological network characteristics and sequence-based descriptors (e.g., autocorrelation and amino acid properties) to infer protein interactions (Yuen and Jansson (2020),Duan et al. (2022)). While these approaches have demonstrated success in specific scenarios, they remain limited in scope due to their reliance on static interaction data without incorporating dynamic biological contexts and the absence of multi-omics integration. Specifically, RNA and protein expression profiles in different tissues and cell types - crucial for capturing condition-specific interactions — are often excluded from existing models, thus limiting their ability to account for biological variability and context-dependent changes in PPI networks.

One of the major challenges in PPI-based protein feature prediction is the extensive computational cost associated with training complex models, particularly when dealing with large-scale biological networks. Traditional machine learning methods often struggle with scalability, while deep learning approaches, such as graph-based models, require significant computational resources to process high-dimensional biological data (Greener et al. (2022)). To address this, our VGAE-based framework aims to generate generalizable protein embeddings that can be efficiently applied to future prediction and classification tasks with minimal retraining. By learning biologically meaningful representations that integrate both network topology and multi-omics features, our model facilitates downstream applications such as disease association studies and functional annotation while significantly reducing computational overhead.

To address these gaps, we introduce a graph-based framework that uses a graph autoencoder (GAE) to predict interactions in PPI networks. GAE is an unsupervised learning approach that learns low-dimensional protein embeddings by reconstructing the PPI network structure, effectively capturing latent relationships between proteins. Unlike previous methods that rely solely on network topology, our GAE-based model integrates multi-omics features, including RNA expression profiles, IHC protein expression levels, and sequence-based embeddings, as node attributes. The encoder employs graph convolutional layers (GCN) to encode structural and biological information into a latent space, while the decoder reconstructs the adjacency matrix, learning biologically meaningful representations of protein interactions.

By incorporating multi-omics features alongside graph-based embeddings, this approach enables the identification of condition-specific protein interactions and functional protein clusters, particularly those associated with disease pathways and cellular regulation. This approach has the potential to significantly enhance our understanding of disease-associated PPIs, paving the way for targeted therapeutic strategies and precision medicine applications.

2 Methods

2.1 Data

Our project leverages comprehensive data sets from the STRING database and the Human Protein Atlas to construct a biologically meaningful Protein-Protein Interaction (PPI) network (Szklarczyk et al. (2023), Thul and Lindskog (2018)). These datasets collectively provide the necessary graph structure and biological annotations for developing and testing our framework.

2.1.1 Protein-Protein Interaction Network

The database obtained from STRING provides the core data consisting of a PPI network with approximately 19.6K nodes (representing proteins) and 13.7 million edges (representing interactions between proteins). Each edge is weighted, reflecting the confidence or likelihood of interaction between two proteins based on experimental data, computational predictions, and text mining. The network serves as the backbone of our graph, defining the structure and connectivity of proteins.

2.1.2 Node Features

To enrich the representation of each protein (node), we integrate biological features obtained from both STRING and the Human Protein Atlas such as:

- RNA Expression: Gene expression profiles across 35 human tissues, reflecting transcriptional activity and providing insights into the regulation of protein abundance and function. This data captures the relative mRNA levels of genes, helping to infer protein expression potential and functional roles across different tissues.
- IHC Protein Expression: Protein expression profiles across 45 human tissues based on immunohistochemistry (IHC) data. These measurements offer a tissue-specific perspective on protein presence and localization. This data helps link protein abundance with tissue-specific functions, enhancing the biological interpretability of graph representations.

These features enhance the biological relevance of each node, complementing the graph's topological information with molecular-level context.

2.2 Preprocessing

2.2.1 Data Preprocessing

The preprocessing phase of the protein-protein interaction score data involves thresholding to keep important interactions, while the preprocessing of node features includes cleaning feature matrices, addressing missing data, and performing dimensionality reduction to improve data quality and enhance model performance.

Protein-Protein Interaction (PPI) Data

The protein-protein interaction dataset consists of an edge list containing protein identifiers and their interaction scores. To reduce data size and focus on high-confidence interactions, we applied a thresholding strategy, retaining only the top 5% of scored interactions.

RNA Expression Data

For RNA expression data, we first utilized the PyEnsembl package (Yates et al. (2020)) to match protein IDs with gene IDs. To ensure data reliability, we removed rows with more than 15% missing values. Finally, we performed principal component analysis (PCA), reducing the dimensionality to 10 principal components (PCs), which were later used as node features.

IHC Protein Expression Data

To process immunohistochemistry (IHC) protein expression data, we first encoded categorical expression levels ('Not detected', 'Low', 'Medium', 'High') as numerical values (0, 1, 2, 3, respectively) to facilitate downstream computations. This encoding captures the relative expression magnitude while preserving interpretability. The data set was then reshaped to ensure that each protein had one measurement per tissue-cell type combination. Tissue-cell type combinations with over 15% missing values were removed. PCA was applied to reduce the feature space to 10 PCs, which were then concatenated with the RNA-seq feature matrix to create a comprehensive representation of protein expression patterns.

2.2.2 Graph Construction and Normalization

Protein-protein interaction data was transformed into a sparse adjacency matrix using PyTorch Geometric utilities. To normalize input data, node features were scaled, and the adjacency matrix was augmented with self-loops to ensure each node's features contributed to the aggregation process. Finally, the dataset was split into training and test subsets, ensuring appropriate separation for evaluation.

To integrate multi-omics information into the graph, proteins were represented as nodes, while their interactions defined the edges. Biological features such as RNA expression and sequence data were incorporated as node attributes, providing a rich, multi-modal annotation for each protein. Edge weights from the PPI network were retained to capture interaction confidence, guiding tasks such as link prediction and node classification.

2.3 Model Architecture

The proposed model, PPI-OMEGA, is a Variational Graph Autoencoder (VGAE), selected for its ability to learn low-dimensional representations of proteins while capturing the structure of the Protein-Protein Interaction (PPI) network. The VGAE consists of an encoder-decoder architecture, where:

- The encoder learns node embeddings by leveraging Graph Convolutional Networks (GCNs).
- The decoder reconstructs the adjacency matrix, predicting the likelihood of protein-protein interactions.

2.3.1 Encoder: Graph Convolutional Layers

The encoder consists of two Graph Convolutional Network (GCN) layers:

- First GCN layer: Transforms input node features $X \in \mathbb{R}^{N \times F}$ into a hidden representation $H^{(1)} \in \mathbb{R}^{N \times 64}$, where N is the number of nodes (proteins), and F is the input feature dimension.
- Second GCN layer: Projects the hidden representation into a latent space of dimension 32.

Mathematically, the encoding process follows:

$$H^{(l)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l-1)}W^{(l)}\right)$$

where:

- $\tilde{A} = A + I$ is the adjacency matrix with self-loops
- $\tilde{D}$ is the degree matrix of $\tilde{A}$
- $W^{(l)}$ is the trainable weight matrix for layer l
- σ is the ReLU activation function

The latent space representation for each node (protein) is then computed as:

$$Z = GCN(X, A)$$

where $Z \in \mathbb{R}^{N \times 32}$ represents the latent embeddings.

2.3.2 Decoder: Inner Product for Edge Reconstruction

The decoder reconstructs the adjacency matrix $\hat{A}$ using an inner product between latent embeddings:

$$\hat{A}_{ij} = \sigma(Z_i^T Z_j)$$

where σ is the sigmoid activation function, ensuring interaction probabilities remain between 0 and 1.

2.4 Training Procedure

The VGAE is trained to maximize the likelihood of reconstructing the observed PPI network while enforcing a structured latent space via Variational Bayes.

2.4.1 Loss function

The training process minimizes a loss function composed of two components:

1. Reconstruction Loss (Binary Cross-Entropy Loss) ensures that predicted protein interactions match actual edges:

$$\mathscr{L}\text{recon} = -\sum(i,j) \in E \left[A_{ij} \log \hat{A}ij + (1 - Aij) \log(1 - \hat{A}_{ij})\right]$$

   where E is the set of edges in the training set.
2. Kullback-Leibler (KL) Divergence Loss enforces a Gaussian prior on latent embeddings:

$$\mathscr{L}\text{KL} = -\frac{1}{2}\sum i = 1^N \sum_{j=1}^{d} \left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)$$

   where μ and σ are the learned mean and variance of the latent space.

The final loss function combines both terms:

$$\mathscr{L} = \mathscr{L}\text{recon} + \beta\mathscr{L}\text{KL}$$

where β is a hyperparameter controlling the balance between reconstruction accuracy and latent space regularization.

2.4.2 Optimization

The model training process was optimized with the following strategies:

- **Optimizer:** Adam optimizer with a learning rate of 0.01 and weight decay of 5×10^{-4} to mitigate overfitting.
- **Training Configuration:** Conducted over 200 epochs using mini-batch processing and GPU acceleration for efficiency.
- **Early Stopping:** Monitored AUC on the validation set, stopping if no improvement (greater than a threshold δ) was observed for 60 consecutive epochs.
- **Model Checkpointing:** The best-performing model was saved and restored after training to ensure optimal final performance.

2.5 Evaluation

To assess model performance, we compute:

1. Area Under the Receiver Operating Characteristic Curve (AUROC): Measures how well the model distinguishes between interacting and non-interacting protein pairs.

$$\text{AUROC} = \frac{1}{m_+ m_-} \sum_{i=1}^{m_+} \sum_{j=1}^{m_-} \mathbb{1}(s_i > s_j)$$

   where m_+ and m_- are the number of positive and negative edges, and s_i and s_j are prediction scores for positive and negative edges.
2. Average Precision (AP): Measures ranking performance in imbalanced datasets.

$$\text{AP} = \sum_k (R_k - R_{k-1}) P_k$$

   where R_k is the recall at threshold k, and P_k is the precision at threshold k .

Both metrics quantify the quality of edge reconstruction and biological relevance of the learned embeddings.

2.6 Tools and Frameworks

The project utilized PyTorch Geometric for implementing the VGAE model and handling graph-based operations. Data preprocessing was performed using Pandas and sklearn, while visualization tasks were accomplished with Matplotlib. GPU resources were leveraged to efficiently train the model on large graph datasets.

3 Results

3.1 Ablation Study of Features

To evaluate the impact of different input feature combinations on model performance, we conducted an ablation study. Table 1 presents AUROC and AP scores for three configurations: using only RNA expression, using only protein expression, and combining both features.

From Table 1, we observe that the protein expression-based model achieves a very high AUROC (0.8073) and the highest AP (0.8102), outperforming both the RNA-only and combined models. This suggests that protein expression features provide stronger predictive signals for PPI modeling. Interestingly, the combined model does not significantly improve performance over protein expression alone, possibly due to feature redundancy or noise introduced by integrating RNA expression.

Table 1: Ablation study result on different input features for PPI prediction using PPI_OMEGA. The presence of RNA expression and IHC protein expression features is indicated with checkmarks (✓). AUROC and AP are reported for each configuration.

RNA Exp.	IHC Protein Exp.	AUROC	AP
✓		0.7835	0.7816
	✓	0.8073	**0.8102**
✓	✓	**0.8075**	0.8065

3.2 Training Dynamics and Model Performance

Aside from final AUC (Area Under the Curve) and AP (Average Precision) scores, we also evaluate model performance across training epochs (Fig. 1) and through the Precision-Recall curve (Fig. 2).

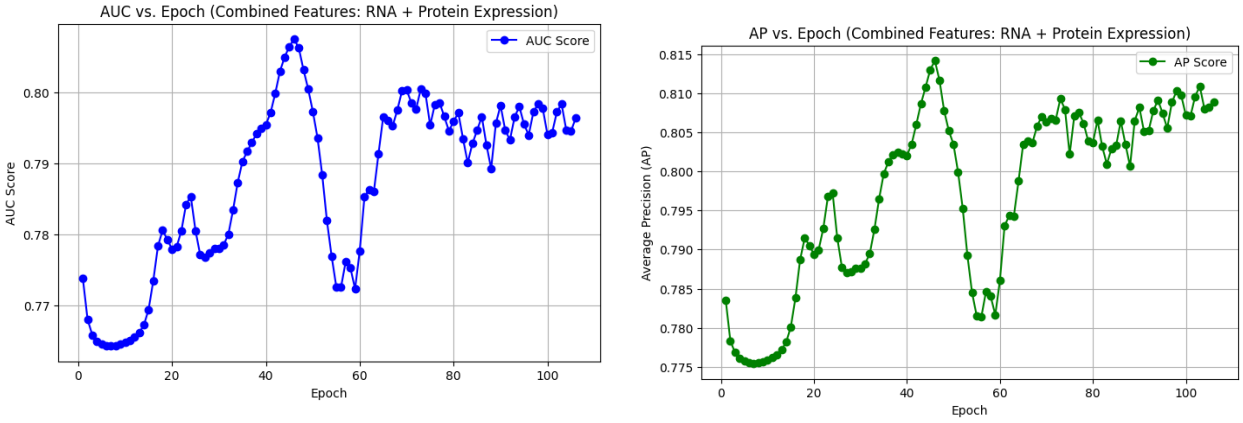

Figure 1: AUC and AP scores over training epochs for the PPI-OMEGA model trained on combined RNA and protein expression features with early stopping.

As shown in the AUC vs. Epoch and AP vs. Epoch plots (Fig. 1), the model's performance fluctuates significantly, particularly in the middle of training. This variability can be attributed to several factors, including sampling strategies, input normalization, and the early stopping criterion. The Precision-Recall curve further confirms that the model achieves a reasonable classification ability with a final PR AUC of 0.8075 (Fig. 2).

One possible cause of performance fluctuations is the input normalization process. Since the model incorporates features derived from different sources (e.g., RNA sequencing, protein expression), differences in scale may introduce instability. Ensuring that all input features are properly standardized (e.g., using z-score normalization) could potentially reduce noise and improve convergence stability.

The selection of the early stopping criterion also plays a crucial role in the observed performance pattern. In our training process, we employed AUC-based early stopping, but the presence of sharp fluctuations suggests that the stopping criterion may not be optimal. Adjusting the patience parameter or using a smoothed validation curve might help prevent

premature stopping or overfitting.

Last but not least, the VGAE architecture inherently introduces stochasticity through variational sampling during latent space encoding. This probabilistic nature causes variations in how embeddings are generated across epochs, leading to the observed performance oscillations. One potential improvement would be incorporating regularization techniques, such as KL divergence annealing, to reduce variance in latent representations.

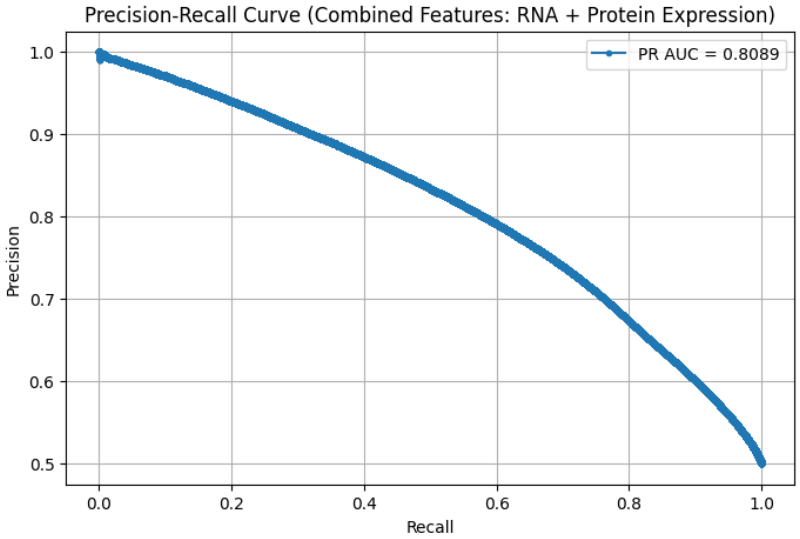

Figure 2: Precision-Recall Curve for PPI-OMEGA trained on combined RNA and protein expression features with early stopping

4 Discussion

5 Conclusion

6 Contribution Statement

Xiaoyu Gui worked on the Abstract and Introduction sections of the proposal as well as refinement of the data and preprocessing part in the Method section.

Siddharth Vyasabattu worked on model architecture, training and evaluation procedure, and tools and framework part in the Method section, as well as background research of related work and finalization onto the latex template.

Each person contributed 50% to this proposal.

References

Duan, Rong, Gong Cheng, Chenyang Wei, Haixuan Yang, and Zhi Wei. 2022. "ProteinPrompt: a webserver for predicting protein–protein interactions by integrating sequence features and graph neural networks." *Bioinformatics Advances* 2 (1), p. vbac059. [Link]

Greener, Joe G., Shaun M. Kandathil, Liam Moffat, and David T. Jones. 2022. "Machine learning solutions for predicting protein–protein interactions." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12 (2), p. e1618. [Link]

Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. 2023. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest." *Nucleic Acids Research* 51 (D1): D638–D646. [Link]

Thul, Peter J., and Cecilia Lindskog. 2018. "The Human Protein Atlas: A spatial map of the human proteome." *Protein Science* 27 (1): 233–244. [Link]

Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N. Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E. Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E. Loveland, Fergal J. Martin, Joannella Morales, Jonathan M. Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J. Trevanion, Fiona Cunningham, Kevin L. Howe, Daniel R. Zerbino, and Paul Flicek. 2020. "Ensembl 2020." *Nucleic Acids Research* 48 (D1): D682–D688. [Link]

Yuen, H. Y., and J. Jansson. 2020. "Better Link Prediction for Protein-Protein Interaction Networks." In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. [Link]

Appendices

A.1 Project Proposal

Q2 Project Proposal:

Integrating Multi-Omics Data to Enhance Protein-Protein Interaction Predictions Using Graph Convolutional Networks

Xiaoyu Gui
xgui@ucsd.edu

Siddharth Vyasabattu
svyasabattu@ucsd.edu

Utkrisht Rajkumar
utkrisht96@gmail.com

Thiago Mosqueiro
thiago.mosqueiro@gmail.com

1 Problem Statement

1.1 Broad Problem Statement

Protein-Protein Interaction (PPI) networks are crucial for understanding the molecular mechanisms of biological systems. They play a pivotal role in various domains, such as drug discovery, disease pathway identification, and functional genomics. While high-quality datasets such as STRING provide extensive PPI data, training models on these large datasets is computationally expensive, requiring significant time and resources (Szklarczyk et al. (2023)). Furthermore, gaps in experimental data and inherent noise in these datasets still hinder advancements in areas like precision medicine, where accurate PPI networks are essential for developing targeted therapies for diseases like cancer. Our project aims to address this limitation by combining graph-based methodologies and multi-omics data to accurately predict interactions in PPI networks and establish a versatile embedding framework for future predictive applications. Utilizing the state-of-the-art graph-based model GC-MERGE and integrating multi-omics features such as RNA expression and protein annotations, we aim to develop a comprehensive framework for PPI prediction that can be applied to a wide range of follow-up studies (Bigness et al. (2022)). This project has the potential to significantly enhance our understanding of PPI networks and their relevance to specific diseases, paving the way for targeted therapeutic strategies.

1.2 Narrow Problem Statement

Protein-Protein Interaction (PPI) networks are fundamental for understanding molecular mechanisms in biology and their roles in diseases. Existing computational methods, such as ExactL3 and ProteinPrompt, have sought to predict interactions within PPI networks using topological features or protein sequence data (e.g., autocorrelation and amino acid properties) (Yuen and Jansson (2020), Duan et al. (2022)). While these methods have demonstrated success in specific scenarios, they remain limited in scope. Notably, they fail to incorporate multi-omics data, such as RNA expression profiles and protein functional annotations, which are essential for capturing condition-specific interactions and dynamic changes within PPI networks. This gap restricts their applicability in domains like precision medicine, where detailed biological context is necessary for understanding disease mechanisms and informing therapeutic strategies.

Our project addresses these deficiencies by introducing a graph-based framework that integrates multi-omics data into PPI network analysis. Building on the GC-MERGE model, which was originally designed to predict gene expression by leveraging long- and short-range genomic interactions, we extend its application to PPI networks (Bigness et al. (2022)). By incorporating RNA expression and protein sequence data as node attributes, alongside graph-based features such as node degree and neighborhood embeddings, our approach provides a biologically enriched representation of PPI networks.

This investigation builds on foundational work in PPI prediction but introduces key advance-

ments by enabling the identification of condition-specific protein interactions and functional clusters, such as those associated with cancer. By addressing the lack of multi-omics integration in current methods, our work bridges the gap between network topology and biological context, offering a more comprehensive and adaptable tool for predicting PPIs and supporting precision medicine research.

2 Methods

2.1 Data

Our project leverages comprehensive datasets from the STRING database and the Human Protein Atlas to construct a biologically meaningful Protein-Protein Interaction (PPI) network (Szklarczyk et al. (2023), Thul and Lindskog (2018)). These datasets collectively provide the necessary graph structure and biological annotations for developing and testing our framework.

Protein-Protein Interaction Network

This database obtained from STRING provides the core data consisting of a PPI network with approximately 19.6 thousand nodes (representing proteins) and 13.7 million edges (representing interactions between proteins). Each edge is weighted, reflecting the confidence or likelihood of interaction between two proteins based on experimental data, computational predictions, and text mining. The network serves as the backbone of our graph, defining the structure and connectivity of proteins.

Node Features

To enrich the representation of each protein (node), we integrate biological features obtained from both STRING and the Human Protein Atlas such as:

- RNA Expression: Quantitative data reflecting gene activity at the transcriptional level, offering insights into protein abundance and function.
- Protein Sequence Information: Biological sequences that can be encoded as features for computational analysis, capturing structural and functional properties.

These features enhance the biological relevance of each node, complementing the graph's topological information with molecular-level context.

Integration into the Graph

Proteins are represented as nodes in the graph, while their interactions form the edges. Biological features such as RNA expression and sequence data are incorporated as node attributes, providing rich, multi-omics annotations for each protein. Edge weights from the PPI network capture the confidence of interactions and are used to guide tasks such as link prediction.

2.2 Model Pipeline

GC-MERGE Framework

The GC-MERGE framework is a graph-based model originally developed to integrate long- and short-range genomic interactions (Bigness et al. (2022)). In our adaptation for PPI networks, nodes represent proteins, and edges capture interactions, with weights reflecting interaction confidence scores. Key components of GC-MERGE include:

- Graph Construction: The graph is built using PPI data, where nodes represent proteins, and edges denote interactions. Node features are enriched with biological attributes such as RNA expression profiles and protein sequences, complementing the network's topological structure.
- Graph Convolutions: GC-MERGE employs graph convolutional layers to iteratively aggregate information from neighboring nodes, enabling embeddings to capture both local (short-range) and global (long-range) interaction patterns within the network.
- Enhancements: To adapt GC-MERGE for PPI tasks, we introduce reinforcement learning (RL)-based node sampling, which prioritizes influential neighbors during aggregation, and the ADOPT optimizer, which improves training stability and convergence compared to standard optimizers like Adam (Oh, Cho and Bruna (2019), Taniguchi et al. (2024)).

These components and enhancements ensure that the embeddings generated by GC-MERGE capture both the network's structural features and its underlying biological context, providing a robust foundation for downstream tasks such as link prediction and protein classification.

Feature Engineering

To incorporate tissue-level RNA expression into our PPI network, we will encode this information as node features in the graph. Tissue expression data, derived from the Human Protein Atlas, provides quantitative or categorical values representing RNA expression across various tissues (Thul and Lindskog (2018)). For each protein, we will create a feature vector where each dimension corresponds to the RNA expression level in a specific tissue. These values will be normalized to ensure comparability and handle potential missing data through imputation or default settings. The resulting feature matrix will integrate tissue-specific expression patterns into the graph, allowing the model to leverage functional context and distinguish condition-specific interactions during training.

To engineer protein sequence features, we will represent each sequence using one-hot encoding (OHE) of the 20 standard amino acids. Since protein sequences vary in length, this results in feature matrices of inconsistent dimensions. To standardize the feature shape, we will employ dimensionality reduction techniques such as Principal Component Analysis (PCA) or autoencoders, compressing the sequence data into fixed-size feature vectors while retaining key structural and functional information. These vectors will then be used as additional node attributes, enabling the model to capture biologically relevant sequence-based properties and improve its predictive power.

2.3 Tasks and Evaluation

2.3.1 Main task

The primary task involves predicting missing or low-confidence interactions within the PPI network. This task evaluates the ability of the embeddings to capture meaningful relationships between proteins.

The model's performance on this task will be evaluated by AUROC (Area Under the Receiver Operating Characteristic Curve), which measure the ability of the model to distinguish between interacting and non-interacting protein pairs.

We will benchmark our model's performance against established link prediction methods, including:

- ExactL3: A topology-based method that uses third-order connectivity patterns.
- ProteinPrompt: A sequence-based method integrating Random Forest and Graph Neural Networks.

To ensure the reliability of our model, we will perform k-fold cross-validation to optimize hyperparameters, including embedding dimensions and learning rate, and select the best-performing configuration. The evaluation will focus on the consistency of the performance metric, AUROC, across folds. Additionally, we will analyze how variations in node features, such as excluding RNA expression or sequence data, impact link prediction accuracy, ensuring the model's robustness in capturing biologically meaningful interactions.

2.3.2 Subtasks

If time allows, we plan to explore additional subtasks that naturally extend the scope of our primary task and provide deeper biological insights.

Subtask 1: Multi-Class Classification for Functional Categories

This task involves classifying proteins into predefined functional categories based on their embeddings. Functional categories may include molecular functions, biological processes, or cellular components from the STRING database.

The task will be evaluated via macro-averaged precision, recall, and F1-score, which assess performance across all classes, treating them equally regardless of class size. We also plan to create confusion matrix that visualizes the distribution of correct and incorrect classifications across categories.

We will compare performance of embeddings derived from our method with a standard Random Forest Classifier trained using topological features only (e.g., node degree, edge weights).

Subtask 2: Oncogenic Protein Clustering

In this task, we investigate whether embeddings can group oncogenic proteins into meaningful clusters using k-means clustering. Oncogenic labels can be obtained from the STRING

database.

The cluster quality will be evaluated by Davies-Bouldin index, which measures the compactness and separation of clusters. We will also assess cluster purity by comparing clusters against labeled oncogenic and non-oncogenic proteins to calculate the proportion of correctly assigned proteins. Additionally, we will visualize the clusters formations with UMAP for dimensionality reduction, highlighting oncogenic proteins.

We also plan to evaluate the cluster quality quantitatively by these ways:

- Cluster Agreement with External Labels: If oncogenic proteins are labeled, evaluate clustering performance using adjusted Rand Index (ARI) or Normalized Mutual Information (NMI).
- Average Within-Cluster Distance: Measure the compactness of clusters in the embedding space, focusing on oncogenic proteins.

3 Expected Outputs

The expected outputs of this project include:

- A scientific report detailing the methodology, results, and significance of the project.
- A fully trained GC-MERGE model and enriched embeddings of proteins, published on GitHub, along with all datasets and scripts for reproducibility.
- A website showcasing the main parts of the study, especially visualizations and analysis results for link prediction and classification tasks.
- A poster presentation summarizing the key findings, methodologies, and significance of the study. The poster will include Introduction and Objectives, Methodology, Results, Visualizations, Conclusion and Future Work, and Acknowledgment and Reference.

4 Contribution Statement

Xiaoyu Gui worked on the writing of the proposal and finalization onto the latex template.

Siddharth Vyasabattu worked on background research of related work and data source validation.

Each person contributed 50% to this proposal.

References

Bigness, Jeremy, Xavier Loinaz, Shalin Patel, Erica Larschan, and Ritambhara Singh. 2022. “Integrating Long-Range Regulatory Interactions to Predict Gene Expression

Using Graph Convolutional Networks." *Journal of Computational Biology* 29 (5): 409–424. [Link]

Duan, Rong, Gong Cheng, Chenyang Wei, Haixuan Yang, and Zhi Wei. 2022. "ProteinPrompt: a webserver for predicting protein–protein interactions by integrating sequence features and graph neural networks." *Bioinformatics Advances* 2 (1), p. vbac059. [Link]

Oh, Jihun, Kyunghyun Cho, and Joan Bruna. 2019. "Advancing GraphSAGE with a Data-Driven Node Sampling." In *Proceedings of the Workshop at International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA. [Link]

Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. 2023. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest." *Nucleic Acids Research* 51 (D1): D638–D646. [Link]

Taniguchi, Shohei, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. 2024. "ADOPT: Modified Adam Can Converge with Any β_2 with the Optimal Rate." *arXiv preprint arXiv:2411.02853*. [Link]

Thul, Peter J., and Cecilia Lindskog. 2018. "The Human Protein Atlas: A spatial map of the human proteome." *Protein Science* 27 (1): 233–244. [Link]

Yuen, H. Y., and J. Jansson. 2020. "Better Link Prediction for Protein-Protein Interaction Networks." In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. [Link]