Integrating Multi-Omics Data to Enhance Protein-Protein Interaction Predictions Using Variational 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 fundamental for understanding cellular processes, with implications in drug discovery, functional genomics, and disease research. However, existing computational models for PPI prediction face limitations, including reliance on static interaction data, exclusion of multi-omics information, and significant computational costs when applied to large-scale datasets.

To address these challenges, we introduce PPI-OMEGA (Protein-Protein Interaction with Omics-Enhanced Graph Autoencoder), a Variational Graph Autoencoder (VGAE)-based framework that integrates multi-omics data to improve PPI prediction. Our model incorporates RNA expression profiles and immunohistochemistry (IHC) protein expression levels as node attributes in a graph-based learning framework. Unlike conventional models, PPI-OMEGA captures both network topology and molecular context, learning biologically meaningful latent representations through graph convolutional layers and probabilistic embeddings.

To assess model performance, we conducted an ablation study, demonstrating that integrating multi-omics features significantly enhances prediction accuracy. The combined RNA and protein expression model achieved the highest AUROC (0.9235) and AP (0.9318), outperforming both RNA-only and protein-only models. Additionally, our analysis of latent representations revealed biological context awareness, as the model successfully distinguished between housekeeping and context-dependent proteins and exhibited structured convergence based on network connectivity.

PPI-OMEGA further reduces computational overhead by generating reusable embeddings, facilitating downstream applications such as disease association studies and functional classification. These findings highlight the potential of graph-based multi-omics integration for advancing PPI modeling and precision medicine applications.

Website: https://xgui17.github.io/PPI-OMEGA-Website/
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)). Moreover, many existing methods require retraining the entire model for each new prediction task, limiting their reusability and generalizability. This poses a bottleneck for efficiently leveraging PPI networks in downstream applications such as disease association studies and functional annotation.

To overcome these limitations, we introduce a graph-based framework, PPI-OMEGA (Protein-Protein Interaction with Omics-Enhanced Graph Autoencoder), that learns biologically context-aware latent representations from PPI networks, enabling both the discovery of generalizable protein embeddings and the accurate prediction of unseen protein interactions. Our approach leverages a variational graph autoencoder (VGAE) to generate biologically meaningful latent representations by integrating network topology with multi-omics features. Unlike standard graph autoencoders, VGAE introduces probabilistic modeling into the latent space, allowing for uncertainty quantification in protein embeddings and improving the model's robustness to noisy biological data. By capturing a distribution over possible protein representations rather than a single deterministic embedding, our approach enhances generalizability and supports more accurate prediction of unseen protein interactions.

We incorporate RNA and protein expression as node attributes because they serve as either

upstream or direct indicators of protein presence and activity within the cell. These features provide functional context across different tissues and cell types, extending beyond network topology alone. RNA-protein interactions are fundamental to gene regulation and can influence protein function and localization, suggesting that RNA expression may indirectly impact protein-protein interactions (Groot et al. (2019)). Additionally, studies indicate that co-expressed genes are more likely to produce interacting proteins, further supporting the relevance of RNA expression in PPI prediction (Grigoriev (2001)). Similarly, IHC-based protein expression captures post-transcriptional regulation and protein abundance, both of which are critical for interaction dynamics. Highly expressed proteins tend to have more interactions, partly due to their increased cellular abundance, which enhances the probability of interaction events (Drummond et al. (2005)). By integrating RNA and protein expression, our model learns biologically meaningful patterns in protein interactions rather than relying solely on network connectivity.

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:

- GTEx 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 (PPI) score data involves thresholding to retain high-confidence interactions and normalizing interaction scores to ensure consistency across datasets. For node features, preprocessing includes cleaning feature matrices, handling missing data, encoding protein expression levels, and applying dimensionality reduction to enhance data quality and improve model performance. These steps ensure that the input data is both biologically meaningful and computationally efficient for downstream learning.

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.

We then applied normalization to scale interaction scores, originally ranging from 611 to 999, to a standardized range between 0 and 1. This transformation ensures that interaction scores are on a comparable scale, preventing biases toward higher-magnitude values, and stabilizes gradient updates during training for improved optimization efficiency.

RNA Expression Data

The GTEx RNA expression dataset originally provides TPM values for genes across various tissues. To integrate this data with our model, we first used the *PyEnsembl* package (Yates et al. (2020)) to map gene IDs to their corresponding protein IDs. To ensure data reliability, we removed genes with more than 15% missing values, retaining a final set of 19,088 proteins. Next, we restructured the data so that each protein had a gene expression

measurement across 35 tissues. Finally, we applied principal component analysis (PCA) to reduce dimensionality, retaining the top 10 principal components (PCs) as node features for downstream modeling.

IHC Protein Expression Data

Immunohistochemistry (IHC) is a technique used to analyze protein expression in tissues by employing antibodies to label target proteins. This method inherently produces discrete categorical data, where protein expression is recorded in qualitative levels. To standardize this data for computational modeling, we first removed NaN values and descriptions labeled as 'Ascending', 'Descending', and 'Not representative', retaining only the four primary expression categories: 'Not detected', 'Low', 'Medium', and 'High'. We then encoded these categories as numerical values (0, 1, 2, and 3, respectively) to facilitate downstream computations while preserving the relative magnitude of expression levels.

The dataset originally contained 249 tissue-cell type combinations, but many had a high proportion of missing expression values for their corresponding proteins. To balance data completeness and biological coverage, we selected 49 tissue-cell type combinations that contained at least 13,000 protein expression measurements, resulting in a final dataset of 10,432 proteins. The data was then reshaped to ensure that each protein had one measurement per tissue-cell type combination.

Similar to the preprocessing of RNA expression data, to reduce dimensionality and extract the most informative features, we applied principal component analysis (PCA), retaining 10 principal components (PCs).

2.2.2 Graph Construction

The protein-protein interaction (PPI) data was transformed into a sparse adjacency matrix using PyTorch Geometric utilities. To ensure that each node's features contributed to the aggregation process, self-loops were added to the adjacency matrix. The dataset was then split into training, validation, and test sets with a proportion of 60:20:20, ensuring a well-balanced separation for model evaluation.

To incorporate multi-omics information, proteins were represented as nodes, while their interactions defined the edges. PCA-derived node features were either directly used or concatenated with additional features, depending on the specific task (further detailed in the Ablation Study section). This process created a unified multi-omics representation of protein expression patterns, resulting in a final dataset of 9,466 proteins, each represented by a set of principal components.

Additionally, normalized edge weights from the PPI network were retained to reflect interaction confidence, ensuring that the model leveraged biologically meaningful relationships for 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. Unlike standard Graph Autoencoders (GAE), VGAE introduces probabilistic latent representations, allowing the model to capture uncertainty in protein embeddings and improve robustness to noisy biological data. Specifically, the model learns a distribution over latent embeddings, rather than fixed point estimates, by parameterizing mean and variance vectors for each node in the latent space.

The VGAE follows an encoder-decoder architecture, where:

- The encoder learns node embeddings by leveraging Graph Convolutional Networks (GCNs) and generates a probabilistic latent space, where each latent dimension is parameterized by a mean (μ) and a variance (σ).
- The decoder reconstructs the adjacency matrix by estimating the probability of protein-protein interactions, using embeddings sampled from the learned latent distribution.

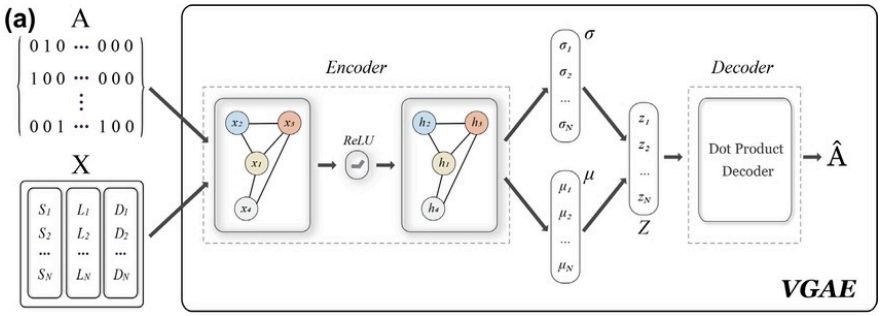

Figure 1: *The Variational Graph Autoencoder (VGAE) model structure. Reproduced from Zitnik et al. (2020).*

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:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}}$$

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 α.
- **Training Configuration:** Conducted over 200 epochs using mini-batch processing and GPU acceleration for efficiency.
- **Regularization:** Applied dropout (rate p) and weight decay (λ) to mitigate overfitting.
- **Early Stopping:** Monitored AUROC on the validation set, stopping if no improvement (greater than a threshold $\delta = 0.01$) was observed for 60 consecutive epochs.
- **Model Checkpointing:** The best-performing model was saved and restored after training to ensure optimal final performance.

2.4.3 Hyperparameter Tuning

Prior to hyperparameter tuning, an ablation study was conducted to assess the impact of different feature combinations. The best-performing configuration, which integrated both RNA-seq and IHC features, was selected for hyperparameter tuning and final evaluation (see 3.1 for details).

To optimize the VGAE model, a systematic hyperparameter search was conducted using grid search across dropout rates, learning rates, and weight decay values. The grid search evaluated multiple combinations:

- Dropout rates p: {0.3, 0.4, 0.5}
- Learning rates α: {0.001, 0.005, 0.01}
- Weight decay values λ: {5e-4, 1e-3, 5e-3}

For each configuration, the model was trained on the graph dataset using a training-validation split to ensure generalization. The best configuration was selected based on the highest Area Under the Receiver Operating Characteristic curve (AUROC) score on the validation set.

To ensure robust evaluation, each model underwent training with early stopping (patience = 60 epochs). The optimal combination was found to be:

- Dropout rate p: 0.3
- Learning rate α: 0.01
- Weight decay λ: 5e-4

This configuration achieved the highest AUROC score among tested combinations and was subsequently used for final model training and evaluation.

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.

While both AUROC and AP evaluate model performance, they emphasize different aspects of classification. AUROC assesses the model's overall ability to distinguish between interacting and non-interacting protein pairs across all thresholds, whereas AP is particularly sensitive to ranking quality in imbalanced datasets by prioritizing high-precision predictions. In biological applications, minimizing false positives is crucial to reduce experimental validation costs and avoid misleading functional annotations. Therefore, if discrepancies arise between AUROC and AP trends across different models, we prioritize AP as it better reflects the reliability of predicted PPIs for downstream biological studies.

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 assess the impact of different input feature combinations on model performance, we conducted an ablation study. Table 1 reports the AUROC and AP scores for four configurations: (1) using only the PPI network without additional features, (2) incorporating only RNA expression, (3) incorporating only protein expression, and (4) combining both RNA and protein expression features.

RNA Exp.	IHC Protein Exp.	AUROC	AP
✗	✗	0.8194	0.8280
✓	✗	0.8888	0.9026
✗	✓	0.9215	0.9297
✓	✓	**0.9235**	**0.9318**

Table 1: *Ablation study results on different input features for PPI prediction using PPI_OMEGA. The presence of RNA expression and IHC protein expression features is indicated with checkmarks (✓), while absence is denoted with crossmarks (✗). AUROC and AP are reported for each configuration, with the best-performing configuration in* ***bold****.*

The results from Table 1 provide insights into the impact of different input features on PPI prediction performance:

1. **Integrating both RNA and protein expression features yields the best performance.** The configuration using both RNA expression and IHC protein expression achieves the highest AUROC (0.9235) and AP (0.9318), demonstrating that incorporating multi-omics features improves PPI prediction accuracy.
2. **Protein expression contributes more significantly than RNA expression alone.** The model trained with only IHC protein expression (✗, ✓) achieves an AUROC of 0.9215 and an AP of 0.9297, which is only slightly lower than the best-performing model. In contrast, using only RNA expression (✓, ✗) leads to a more pronounced performance drop (AUROC = 0.8888, AP = 0.9026). This suggests that RNA expression may play a relatively smaller role in improving PPI predictions compared to protein expression.
3. **Using direct expression-based features enhances PPI prediction even without sequence-based features.** The highest AUROC and AP scores indicate that leveraging RNA and protein expression data improves model performance, leading to a **12.54% increase in average precision (AP)** compared to the model without multi-omics features. This confirms that integrating multi-omics data provides a viable alternative or complement to traditional sequence-based PPI prediction methods.

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 to analyze training stability and convergence behavior. Figure 2 presents AP scores over training epochs for different feature sets, illustrating how RNA and protein expression features influence training dynamics.

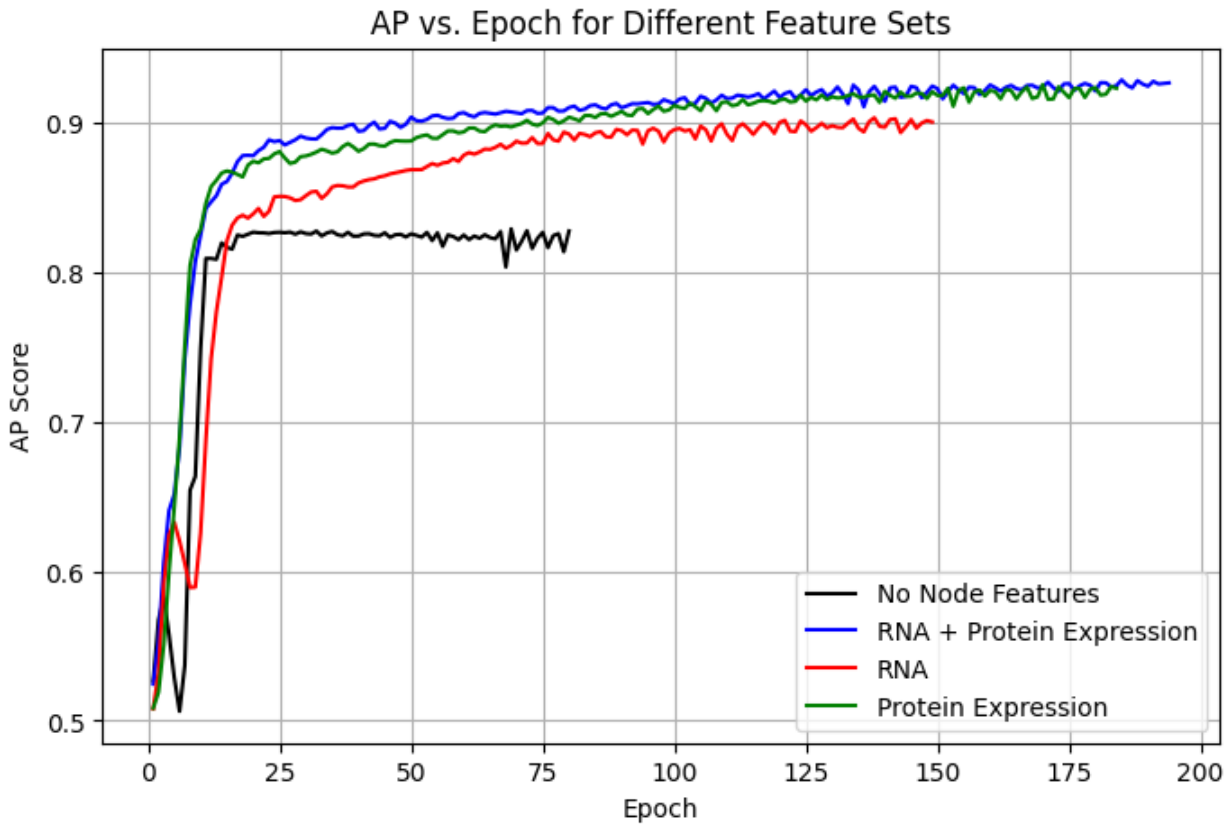

Figure 2: AP scores over training epochs for the PPI-OMEGA model trained on different feature sets tested in the ablation study with early stopping.

This figure provides insights into the impact of different input feature sets on training performance. The model trained with both RNA and protein expression (blue) converges faster and achieves the highest final AP score, demonstrating the effectiveness of multi-omics integration. The model using only protein expression (green) achieves similar final performance to the multi-omics model, suggesting that protein expression plays a major role in improving PPI predictions. The RNA-only model (red) lags behind the protein-only and multi-omics models in both convergence speed and final AP score, indicating that RNA expression alone provides less predictive power for PPI. Finally, The model trained without node features (black) struggles with stability and achieves the lowest AP score, confirming that direct biological features (RNA/protein expression) are essential for accurate PPI predictions.

One important consideration is that the VGAE architecture inherently introduces stochas-

ticity 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.

These results reinforce the findings from the ablation study (Table 1), confirming that incorporating protein expression is more beneficial than RNA expression alone and that multi-omics integration provides the best predictive performance.

3.3 Biological Context Awareness in Latent Representations

To investigate whether the learned latent representations capture biologically meaningful structure, we examined their organization over training epochs using two analyses: UMAP projections of latent embeddings for biologically distinct protein groups, and variance evolution of latent means for highly and sparsely connected proteins.

3.3.1 Separation of Housekeeping and Context-Dependent Proteins

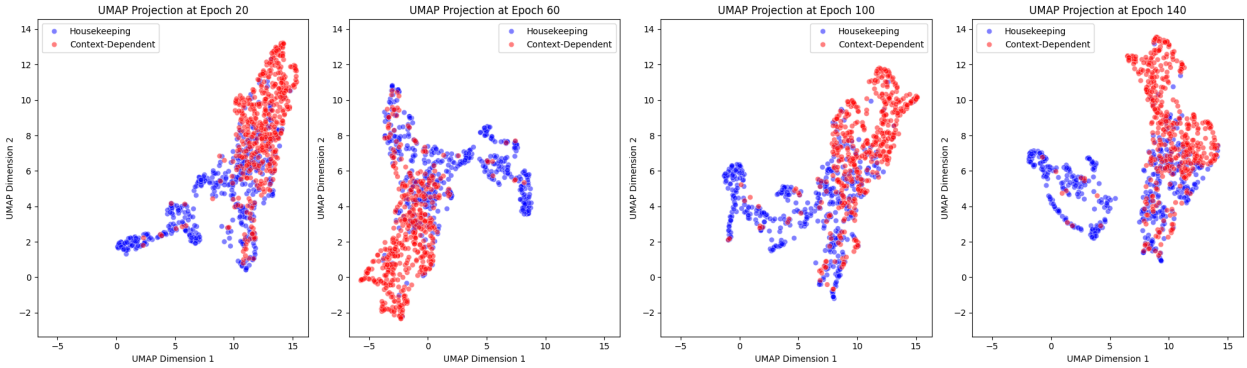

Figure 3: *UMAP projections of latent representations for housekeeping (blue) and context-dependent (red) proteins at different training epochs.*

Figure 3 presents UMAP projections of latent representations at different training epochs, highlighting proteins categorized as housekeeping (blue) or context-dependent (red). These categories were assigned using a keyword-based filtering approach, where ribosome-related proteins were labeled as housekeeping, while signaling proteins were designated as context-dependent (see Appendix A.1 for the full list of keywords used for classification).

Across training epochs, we observe an increasing separation between the two categories, indicating that the model progressively refines its embeddings to distinguish biologically distinct protein groups. However, some overlap persists even at later epochs, which may be biologically meaningful. These mixed proteins could represent cases where these defined housekeeping and context-dependent proteins interact, leading to embeddings that capture shared functional relationships rather than strict categorical separation.

3.3.2 Convergence of Highly Connected and Sparsely Connected Proteins

To further assess the latent space organization, we analyzed the convergence behavior of proteins based on their connectivity in the PPI network.

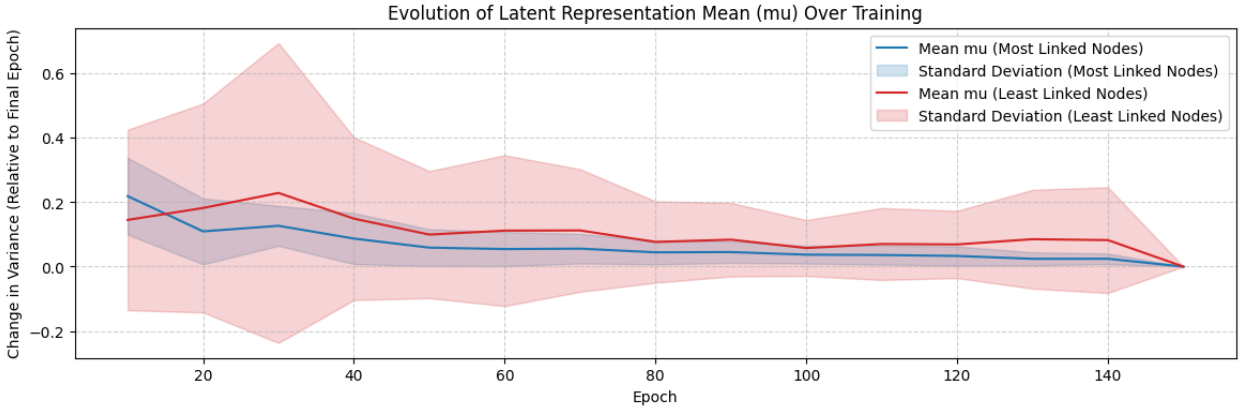

Figure 4: *Changes in variance of the latent representation mean (μ) over training. The y-axis represents the change in variance relative to the final epoch, measuring how quickly different groups converge.*

Figure 4 tracks the variance of the latent representation mean (μ) over training epochs for highly linked proteins (blue) and sparsely linked proteins (red). The top and bottom 50 proteins by node degree were selected to represent these two groups.

The results indicate that highly linked proteins converge more quickly, as their variance stabilizes earlier in training. This suggests that the model rapidly learns robust embeddings for well-connected nodes, which likely benefit from stronger structural constraints in the graph. In contrast, sparsely connected proteins exhibit greater variance fluctuations before stabilizing, implying that their representations require more training epochs to reach a stable state. This observation aligns with expectations, as less connected proteins have fewer interaction-based constraints guiding their embeddings.

Together, these analyses demonstrate that the PPI-OMEGA model captures biologically meaningful latent structures, separating functionally distinct protein groups while encoding network connectivity constraints into the learned representations.

4 Discussion

Our study demonstrates that integrating multi-omics features, specifically RNA and protein expression, enhances protein-protein interaction (PPI) prediction beyond traditional sequence-based approaches. However, several aspects of our methodology warrant further refinement. One notable limitation is the treatment of dropped or unused data during preprocessing. While thresholding and PCA helped manage noise and reduce dimensionality, these steps may have inadvertently discarded informative low-confidence PPIs or subtle expression variations. Future work could explore more nuanced filtering techniques, such as

adaptive thresholding or data imputation, to retain valuable biological signals without compromising model performance. Additionally, refining data selection criteria could allow the inclusion of a broader range of proteins while maintaining robustness.

Another challenge lies in our current approach to encoding protein expression, particularly the reliance on immunohistochemistry (IHC) data. Although IHC provides tissue-specific protein localization, its imaging-based nature introduces variability, making quantification less reliable than transcriptomics or proteomics-based methods. Employing more precise protein quantification techniques, such as mass spectrometry-based proteomics or normalized RNA-protein expression values, could improve the consistency of our feature set. Furthermore, incorporating temporal data, such as time-series expression profiles, could enhance the model's ability to capture dynamic PPI interactions, reflecting the transient nature of certain protein functions.

While most existing PPI prediction models are heavily reliant on protein sequence data, our approach prioritized multi-omics integration due to time constraints. Sequence data remains a fundamental predictor of protein interactions, but our findings suggest that RNA and protein expression patterns contribute significantly to prediction accuracy. Given these results, future work should focus on integrating protein sequence features alongside multi-omics data to build a more comprehensive and accurate PPI prediction framework. This could help bridge the gap between sequence-based and expression-based models, leveraging the strengths of both approaches.

In addition to improving feature representation, refining our model architecture could further enhance predictive performance and interpretability. One promising avenue is the incorporation of an attention mechanism, which would allow the model to focus on the most biologically relevant nodes and interactions. This could be particularly beneficial for identifying disease-associated proteins, where certain key interactions may have stronger biological significance than others. Implementing attention layers could improve both interpretability and the robustness of predictions in complex biological networks.

Beyond methodological refinements, broader extensions of this work could provide a more holistic view of cellular processes. Expanding beyond RNA and protein expression to include epigenetic modifications, metabolic profiles, or spatial transcriptomics data could reveal additional layers of regulation influencing PPIs. Additionally, evaluating the biological context awareness of learned embeddings through domain-specific tests would provide deeper insights into their interpretability, though such validation would require guidance from experimental studies.

Finally, addressing practical challenges such as class imbalance is essential for real-world applications. In contexts such as drug discovery, minimizing false positive predictions is crucial, as incorrect interactions could lead to costly downstream experimental validation. Fine-tuning the model to reduce false discovery rates (FDR) while maintaining sensitivity remains an important goal. Furthermore, improving the generalizability of the model across different tissues and biological conditions could support its application in precision medicine, enabling personalized treatment strategies tailored to individual molecular profiles.

Taken together, these considerations highlight both the strengths and limitations of our current approach while outlining key directions for future improvements. By integrating additional biological data sources, refining model architecture, and improving validation strategies, we aim to develop a more comprehensive, biologically-informed framework for PPI prediction.

5 Contribution Statement

Xiaoyu Gui worked on the writing the Abstract, part of Introduction, data preprocessing in Methods, training dynamics in Results, and Discussion sections of this report.

Siddharth Vyasabattu worked on model architecture, training and evaluation procedure, and tools and framework part in the Method section, latent representation interpretation in Results as well as background research of related work and part of Introduction.

Each person contributed 50% to this report.

References

Drummond, D. Allan, Jesse D. Bloom, Christoph Adami, and Frances H. Arnold. 2005. "Why highly expressed proteins evolve slowly." *Proceedings of the National Academy of Sciences* 102 (40): 14338–14343. [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]

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]

Grigoriev, Andrei. 2001. "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae." *Nucleic Acids Research* 29 (17): 3513–3519. [Link]

de Groot, Natalia Sanchez, Alexandros Armaos, Ricardo Graña-Montes, Marion Alriquet, Giulia Calloni, R. Martin Vabulas, and Gian Gaetano Tartaglia. 2019. "RNA structure drives interaction with proteins." *Nature Communications* 10 (1), p. 3246. [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]

Zitnik, Marinka et al. 2020. "Graph2GO: a multi-modal attributed network embedding method for inferring protein functions." *Bioinformatics* 36 (Supplement_2): i418–i426. [Link]

Appendices

A.1 Keywords for Protein Classification

The following keywords were used to classify proteins as housekeeping or context-dependent based on their functional annotations in the protein description database from the Human Protein Atlas.

A.1.1 Housekeeping Proteins

Proteins were labeled as housekeeping if their descriptions contained any of the following terms after converting to lower cases:

- ribosome
- translation
- elongation factor
- rrna
- ribosomal protein
- rna polymerase

A.1.2 Context-Dependent Proteins

Proteins were labeled as context-dependent if their descriptions contained any of the following terms after converting to lower cases:

- cytokine
- chemokine
- interleukin
- tnf
- interferon
- complement
- granzyme
- perforin

These keywords were selected to distinguish proteins involved in core cellular processes (housekeeping) from those involved in signaling and regulatory pathways (context-dependent).

A.2 Project Proposal

For further details on the project scope and methodology, refer to the proposal document available here.