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

Siddharth Vyasabattu svyasabattu@ucsd.edu
Xiaoyu Gui xgui@ucsd.edu
Mentors: Utkrisht Rajkumar utkrisht96@gmail.com
Thiago Mosqueiro thiago.mosqueiro@gmail.com
University of California San Diego, Halıcıoğlu Data Science Institute, La Jolla, CA

SCAN ME

INTRODUCTION

Protein-protein interaction (PPI) networks are essential for understanding molecular mechanisms like signal transduction, gene regulation, and metabolic processes. Accurate PPI predictions are crucial for drug discovery, disease pathway analysis, and personalized medicine.

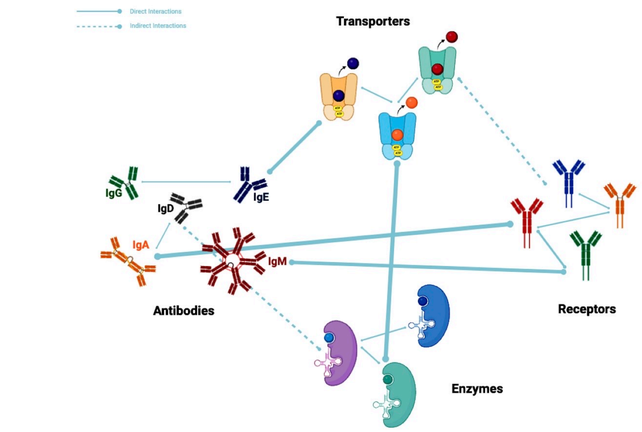

Figure 1: Interactions between different proteins

Our project, Protein-Protein Interaction with Omics-Enhanced Graph Autoencoder, a.k.a. **PPI-OMEGA**, integrates multi-omics data—specifically RNA and protein expression profiles—into a Variational Graph Autoencoder (**VGAE**) framework. This integration allows for biologically meaningful protein embeddings, improving prediction accuracy while maintaining computational efficiency.

DATA

- Graph Structure (Sourced from STRING Database)
  - Protein-Protein interaction: ~19.6K proteins and ~13.7M interactions with weighted edges determined by the strength of the link
- Node Features (Sourced from the Human Protein Atlas)
  - Bulk RNA expression: across 35 human tissues
  - IHC (immunohistochemistry) Protein expression: across 45 human tissues

METHODOLOGY

Preprocessing

- Applied thresholding to retain top 5% of high-confidence PPIs.
- Normalized node features and interaction scores.
- Encoded discrete protein expression levels (IHC) into numerical values (0-3).
- Performed Principal Component Analysis on RNA and protein expression data to reduce dimensionality to 10 PCs each.
- Constructed a graph with proteins as nodes, PPIs as edges, and multi-omics data as node attributes.

Model Architecture

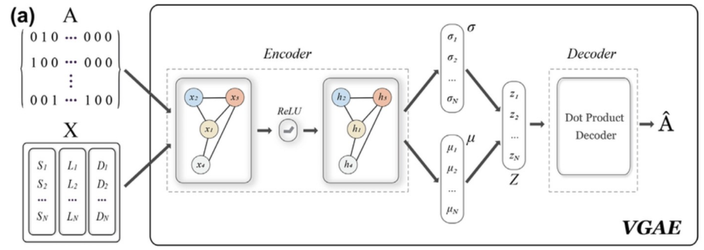

Figure 2: Architecture of the Variational Graph Autoencoder (Fan et al., 2020).

Training & Evaluation

- Loss Function:

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

- Early stopping & regularization
- Feature Importance Analysis via Ablation Study
  - No feature; RNA Exp only; Protein Exp only; Combined
- Evaluation Metrics: AUROC & AP
- Hyperparameter tuning: Dropout rate p = 0.3; Weight decay λ = 0.0005; Learning rate α = 0.01

RESULTS

RNA Exp.	IHC Protein Exp.	AUROC	AP
❌	❌	TBD	TBD
✅	❌	0.8837	0.8972
❌	✅	0.9057	0.9135
✅	✅	**0.9115**	**0.9198**

Table 1: Comparison of AUROC and AP from Ablation Study with Different Feature Sets

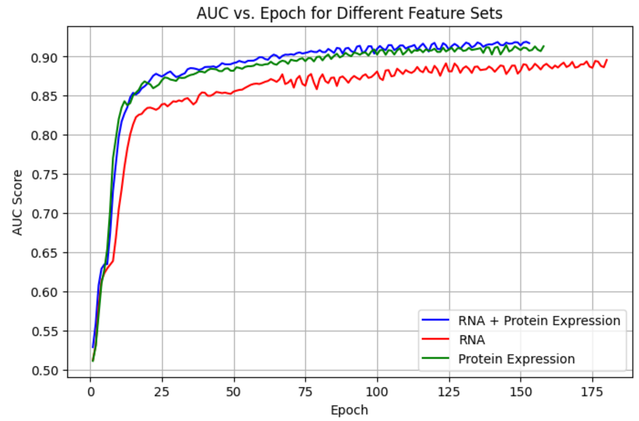

Figure 3: Comparison of Model Performance among different feature sets

DISCUSSION

- Integrating RNA and protein expression improves PPI prediction accuracy, enabling better interaction modeling and guiding drug discovery and functional genomics research.
- Limitations: Encoding of protein expression introduces variability; protein sequence data was omitted due to time/memory constraints.
- Future Work: Incorporate an attention mechanism to highlight critical interactions; integrate protein sequence data to benchmark against traditional models; expand the model to include multi-omics data beyond RNA and protein expression; benchmark against existing models with protein sequence features.

ACKNOWLEDGEMENTS

We sincerely thank the Halıcıoğlu Data Science Institute (HDSI) for supporting this capstone project. We are especially grateful to our mentors, Dr. Utkrisht Rajkumar and Dr. Thiago Mosqueiro , for their invaluable guidance. We also appreciate the assistance of our TA, Aritra Das, for their insightful feedback throughout the project. Additionally, we acknowledge the contributions of our teammates and the resources provided by UC San Diego.