SCAN ME

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

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

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.

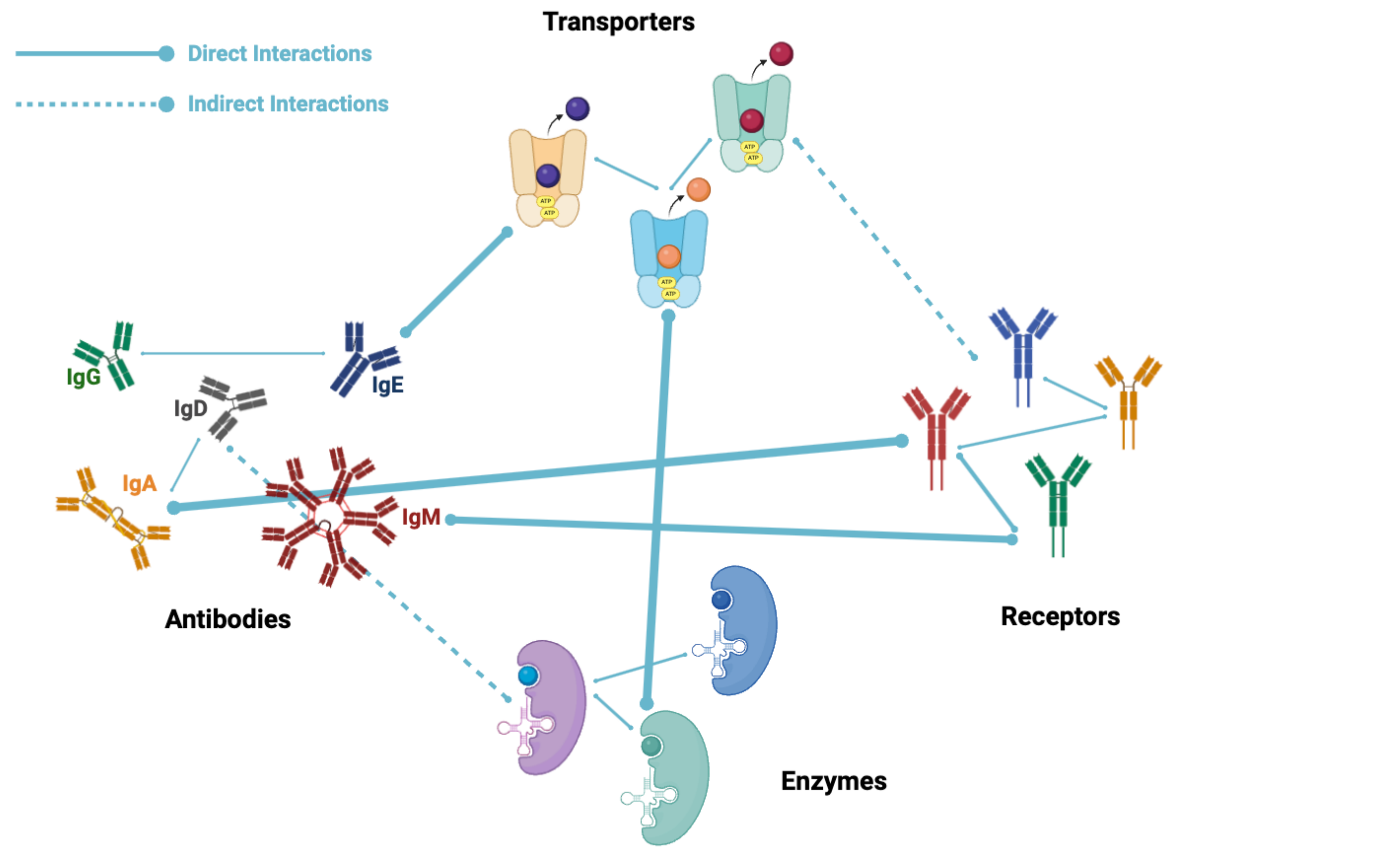

Figure 1: Example diagram illustrating interactions between different proteins. Note: This is a conceptual representation and not based on actual data from our study.

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. By generating protein embeddings within a multi-omics biological context, PPI-OMEGA enhances prediction accuracy while maintaining computational efficiency.

DATA

- Graph Structure (Sourced from STRING Database)
  - Protein-Protein interaction: ~19.6K proteins and ~13.7M interactions, where edges are weighted based on interaction strength.
- 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 interaction scores.
- Encoded discrete protein expression levels to numerical values (0-3) and retained tissue/cell types with sufficient data.
- 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

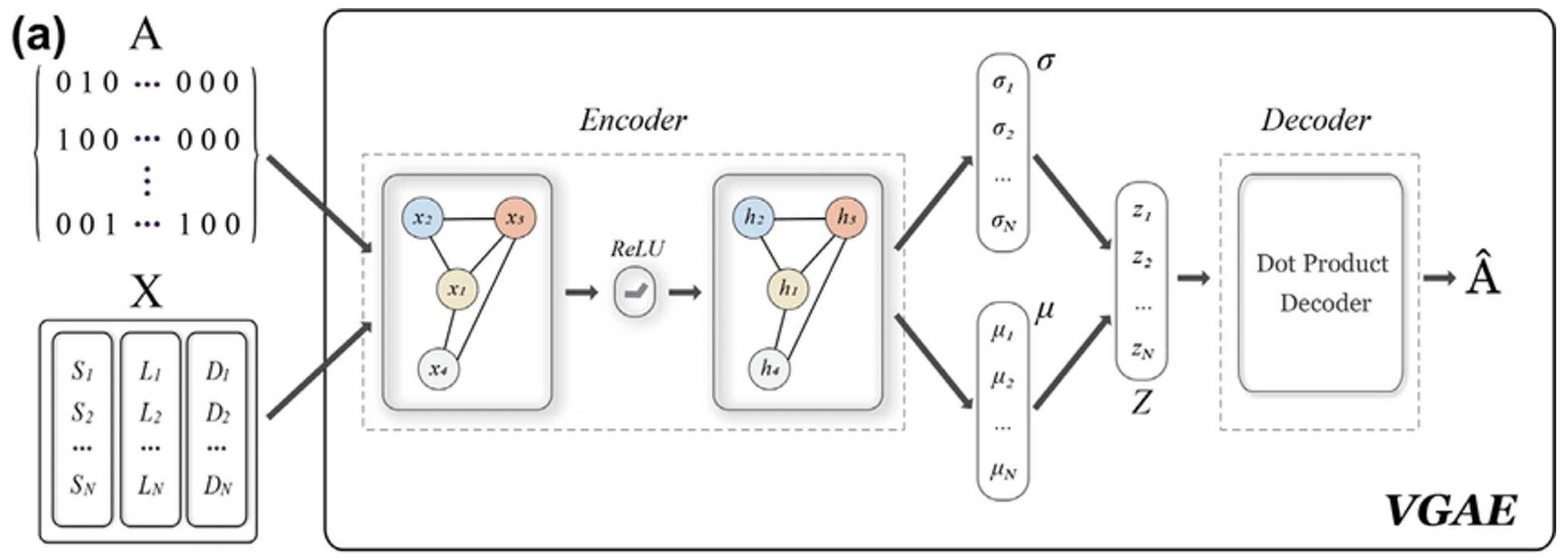

Figure 2: Architecture of the Variational Graph Autoencoder (Fan et al., 2020).

Training & Evaluation

- Loss Function:

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

- 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

Combined features yield the best predictive performance

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

Latent Representations Capture Biologically Meaningful Structure

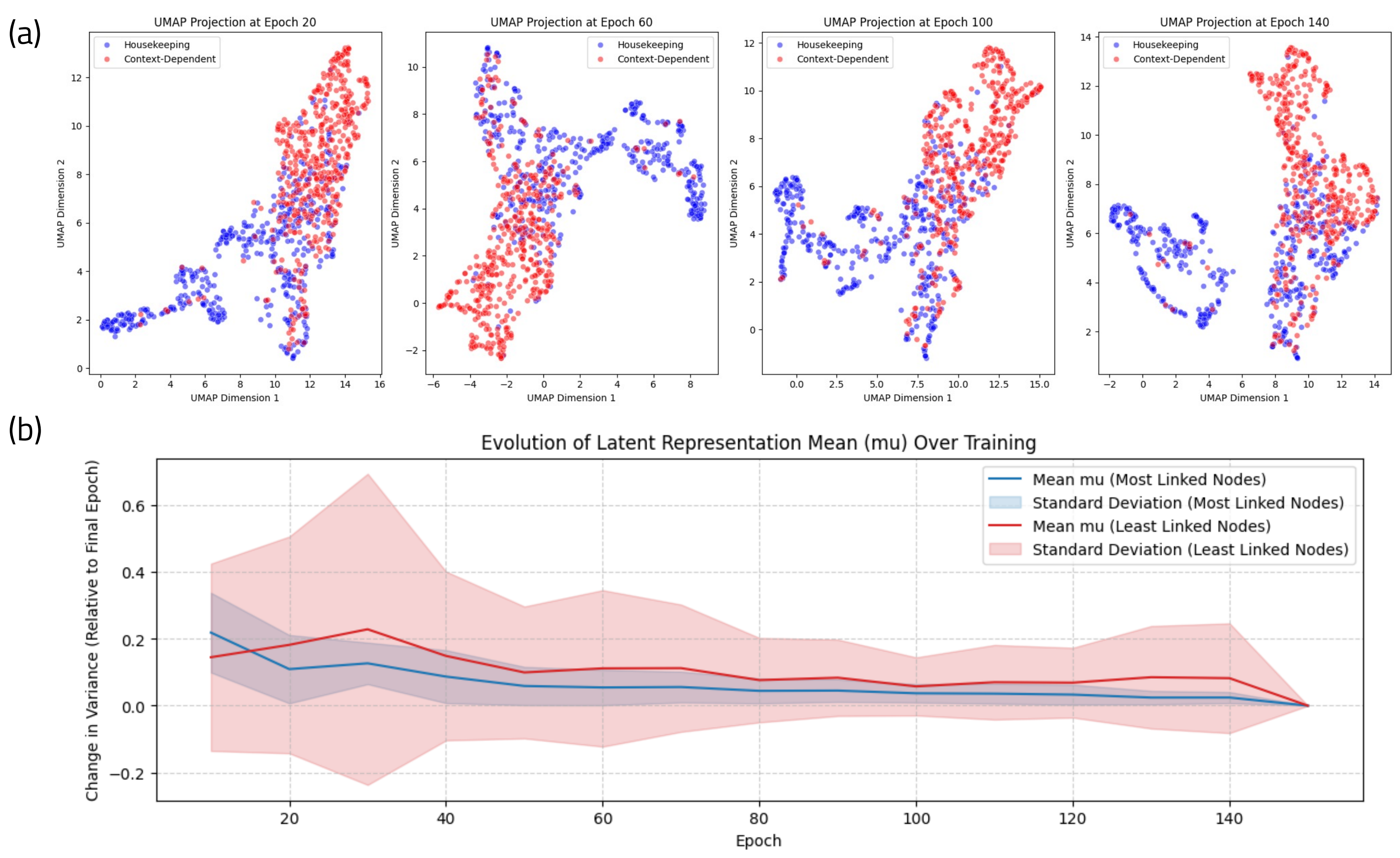

Figure 3: Evolution of Latent Representations Over Training Epochs. a. *UMAP projections of latent representations for housekeeping (blue) and context-dependent (red) proteins at different training epochs.* ***b.*** *Changes in variance of the latent representation mean (μ) over training. The y-axis represents the change in variance relative to the final epoch, measuring how quickly different groups converge.*

DISCUSSION

- Integrating multi-omics data 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; diminishing returns with RNA expression features.
- Future Work: Incorporate an attention mechanism; integrate protein sequence data; include multi-omics data beyond current ones; benchmark against existing models (e.g. Exact L3, ProteinPrompt).

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.