Applications of Convolutional Graph Neural Networks for Proteomic Analysis

Robert Heeter, Arielle Sanford • *Rice University, Dept. of Electrical & Computer Engineering* (*COMP/ELEC 576 Intro. to Deep Learning*) • 5 December 2023

Abstract

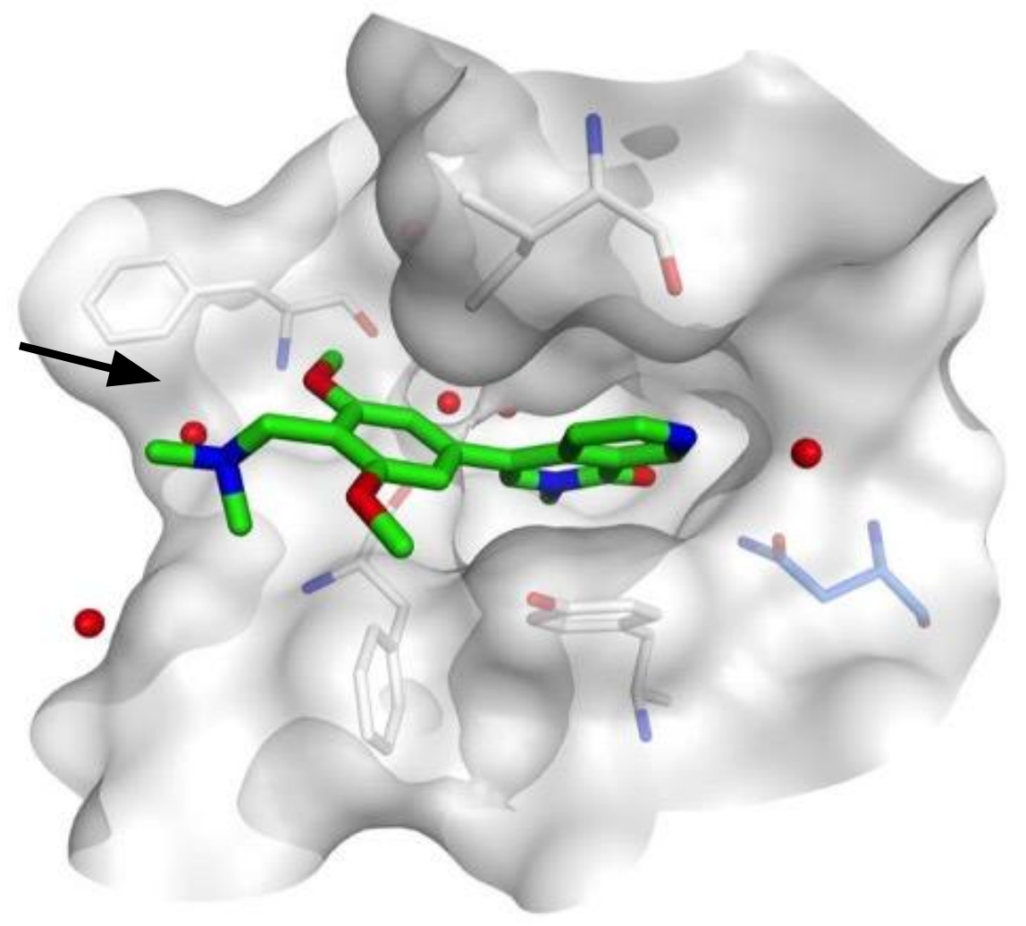
- Machine learning advances can accelerate drug development, but struggle with data imperfections and model interpretability
- Use convolutional graph neural networks (CNNs) for proteomic prediction tasks
  - Offer better interpretability by more naturally representing molecular data

1. **Explore use of graph CNNs by modeling the interaction between proteins using the popular Protein-Protein Interaction (PPI) dataset**
2. **Identify druggable cavities in novel proteins using graph representations of structural data in the Protein Data Bank (PDB) to train a graph CNN**

Background & Motivation

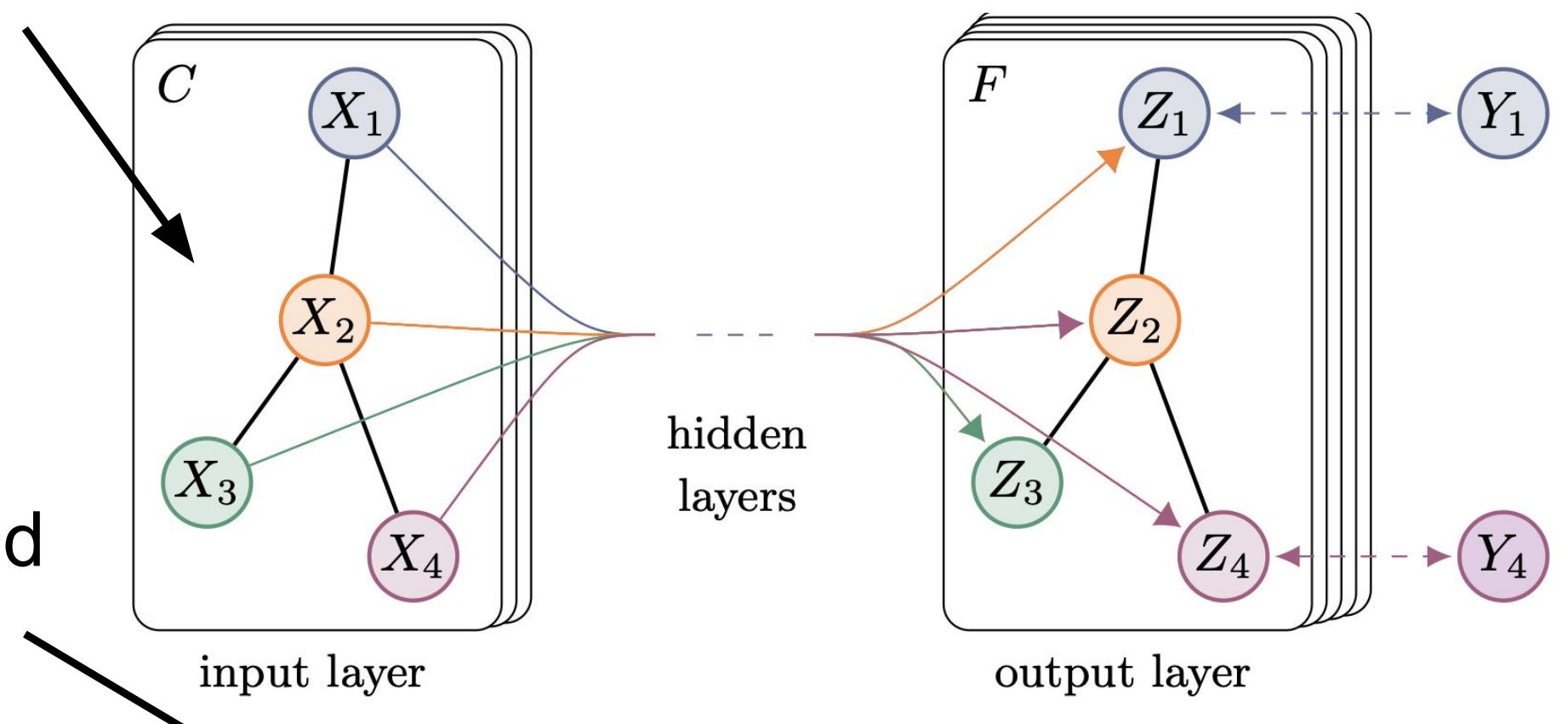
Protein Biology

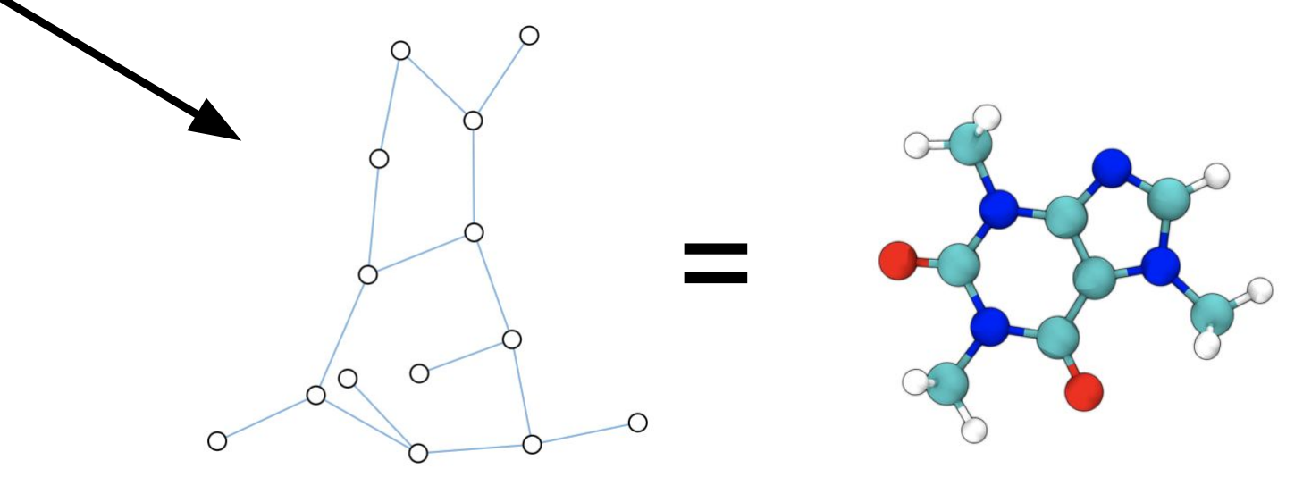
- Protein-protein & protein-ligand interactions
  - *Physical properties:* volume, "enclosure", surface protrusions or "roughness", opening size, depth
  - *Chemical properties:* hydrogen bonding, electrostatic interactions, hydrophobic and van der Waals forces



Graph Neural Networks

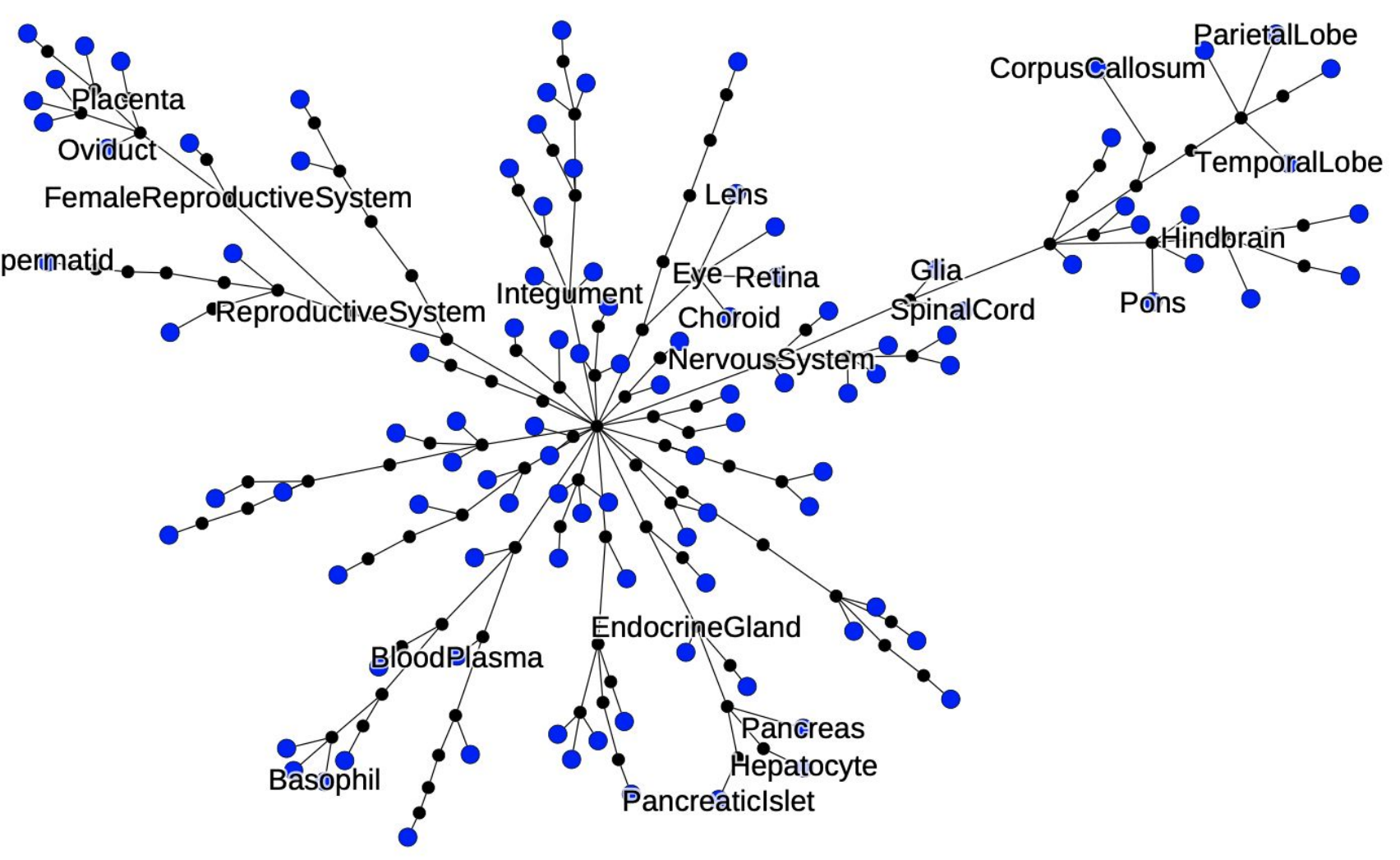
- Edges determine message passing between nodes in neural network
- Protein connectivity can be encoded more naturally in graph form
  - Atoms/residues = nodes
  - Bonds or interactions = edges
- More 'raw' chemical features assigned to nodes and edges





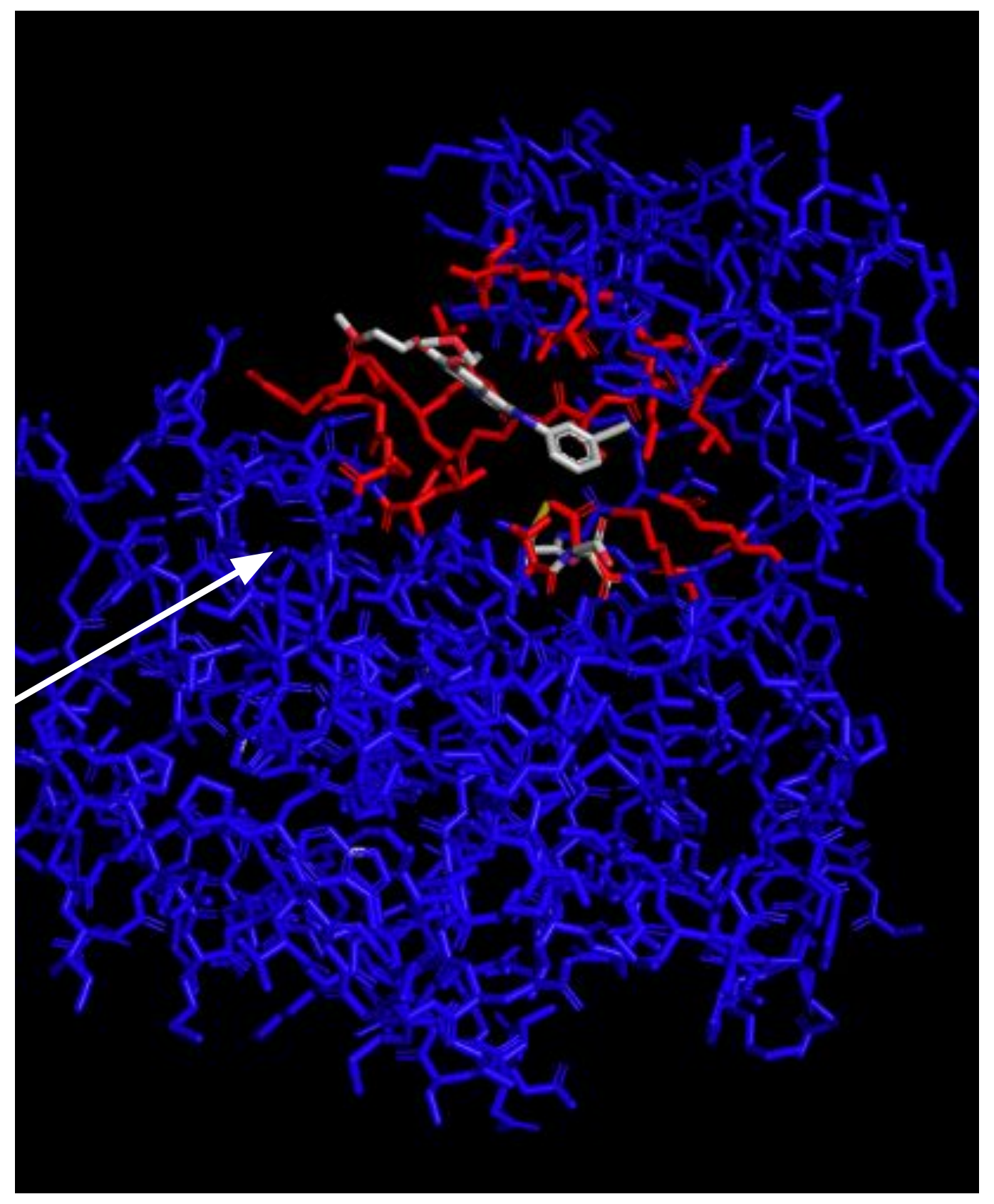
Protein-Protein Interaction Dataset

- *Graphs:* different human tissues
  - *Nodes:* proteins
  - *Edges:* interactions
- *Node features:*
  - Positional Gene Sets: shared physical locations on chromosome
  - Motif Gene Sets: commonly shared patterns on DNA/RNA
  - Immunological Signatures: gene sets involved in immune response
- *Labels:* gene ontology (vocabulary to describe roles of proteins)
- *Classes:* protein roles derived from the gene ontology labels
- 20 graphs, ~2.2k nodes, 61.3k edges, 50 node features



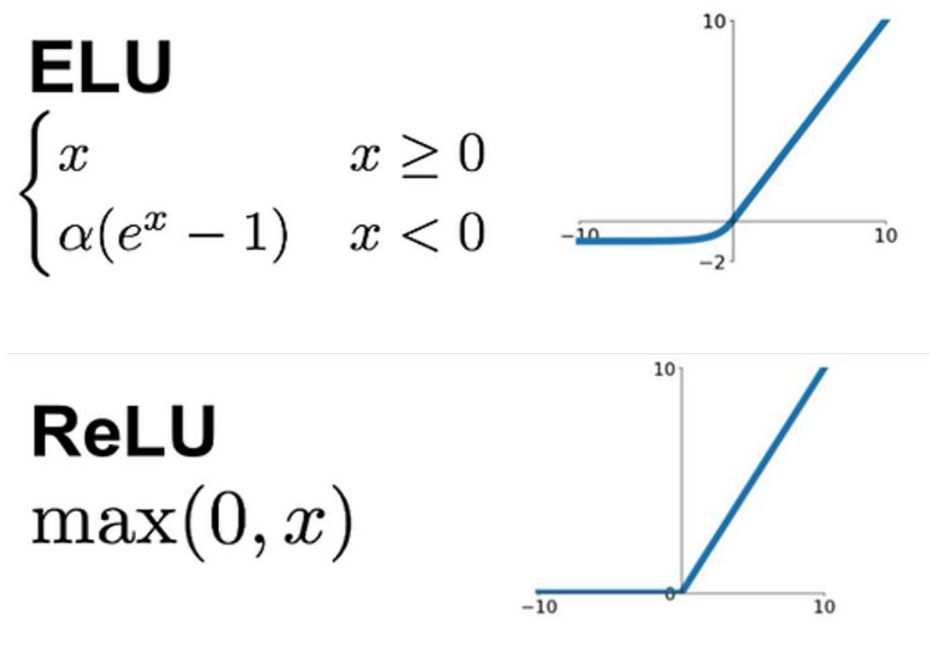
Protein-Ligand Binding Dataset

- *Graphs:* representative drug-binding proteins
  - *Nodes:* atoms (without hydrogens)
  - *Edges:* atomic interactions
- *Node features:* 39 physiochemical descriptors (element, atom degree, charge, radical electrons, hybridization, etc.)
- *Edge features:* binary covalent/non-covalent
- *Labels:* given node contributes to drug binding or not
- Built from scratch using Protein Data Bank
- 3,600 graphs, >1m nodes, >1b edges, 39 node features, 2 edge features



PPI Model Experimentation & Model Design

- 3 graph convolution layer operators:
  - *Linear layers:* combine the node features into higher-level representations
  - *GATConv layers:* graph attention network operator allowing for "nodes to attend over their neighborhoods' features"
    - Improves upon recurrent neural networks that favor more recently "seen" data
  - *Activation functions:* ELU vs. ReLU



- BCEWithLogitsLoss works well for multi-label classification datasets like PPI
- Training batch size = 1

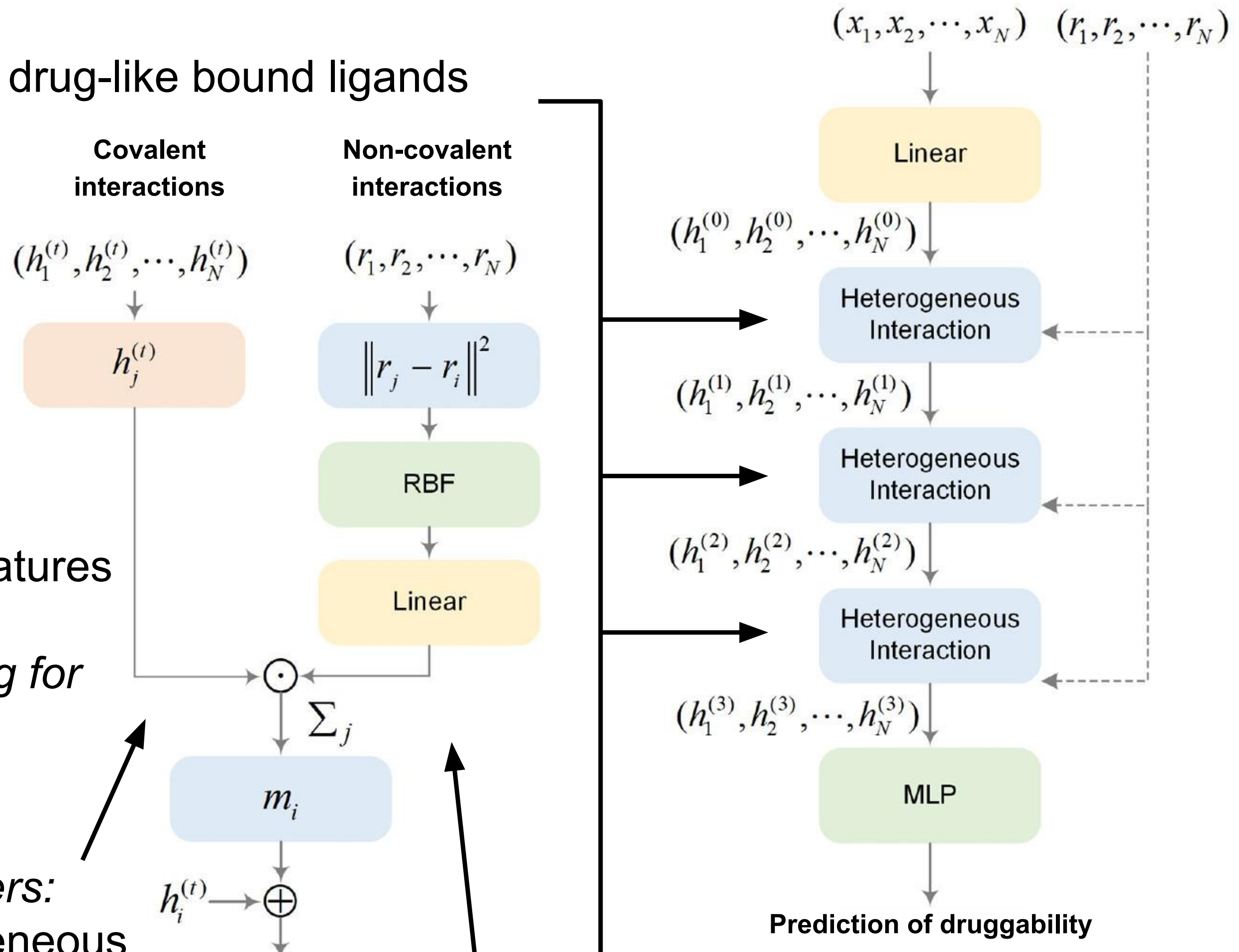
Druggability Classifier Experimentation & Model Design

Dataset Generation

- Filter proteins with drug-like bound ligands
  - Completeness
  - Resolution
  - Uniqueness
- Convert 3D structure to PyTorch graph object
- Assign 39 physiochemical features
- *Use subset of 30 proteins for training for now*

Model

- *3 GATv2Conv layers:* based on heterogeneous interaction layers with 256 hidden channels
- *Activation functions:* ReLU

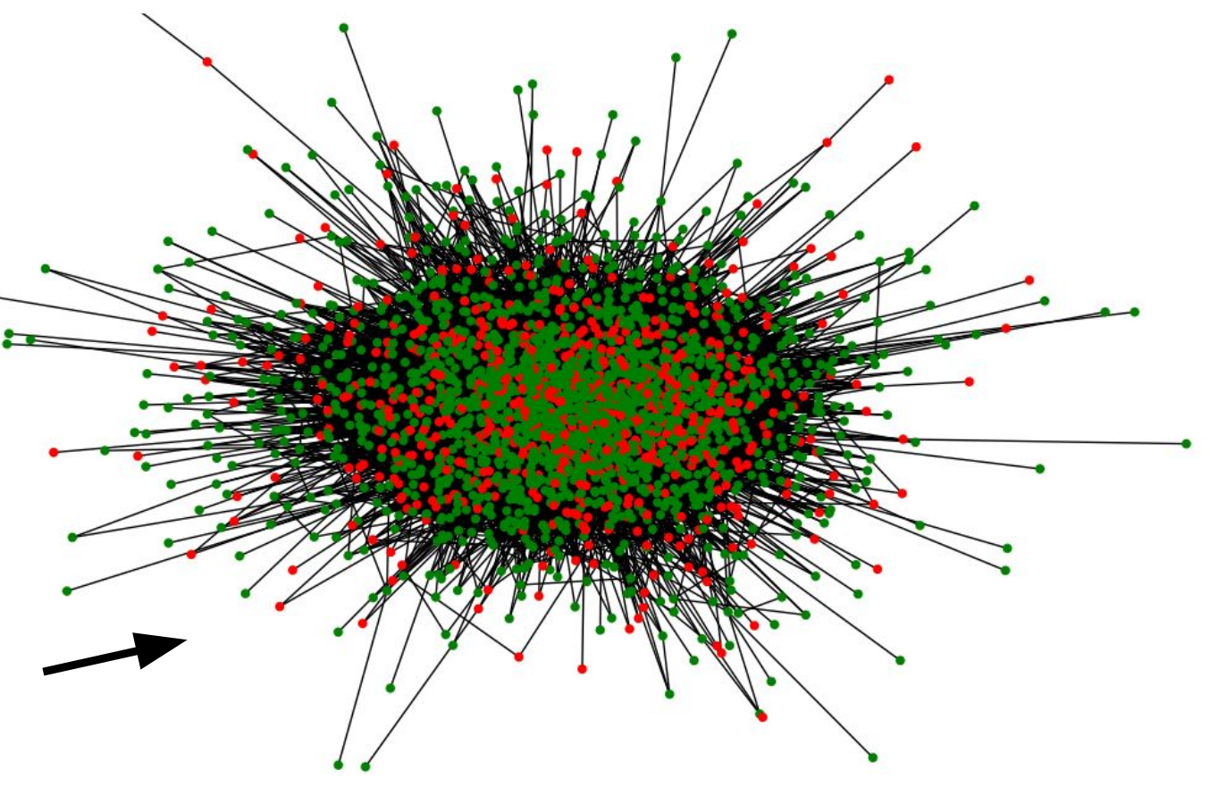


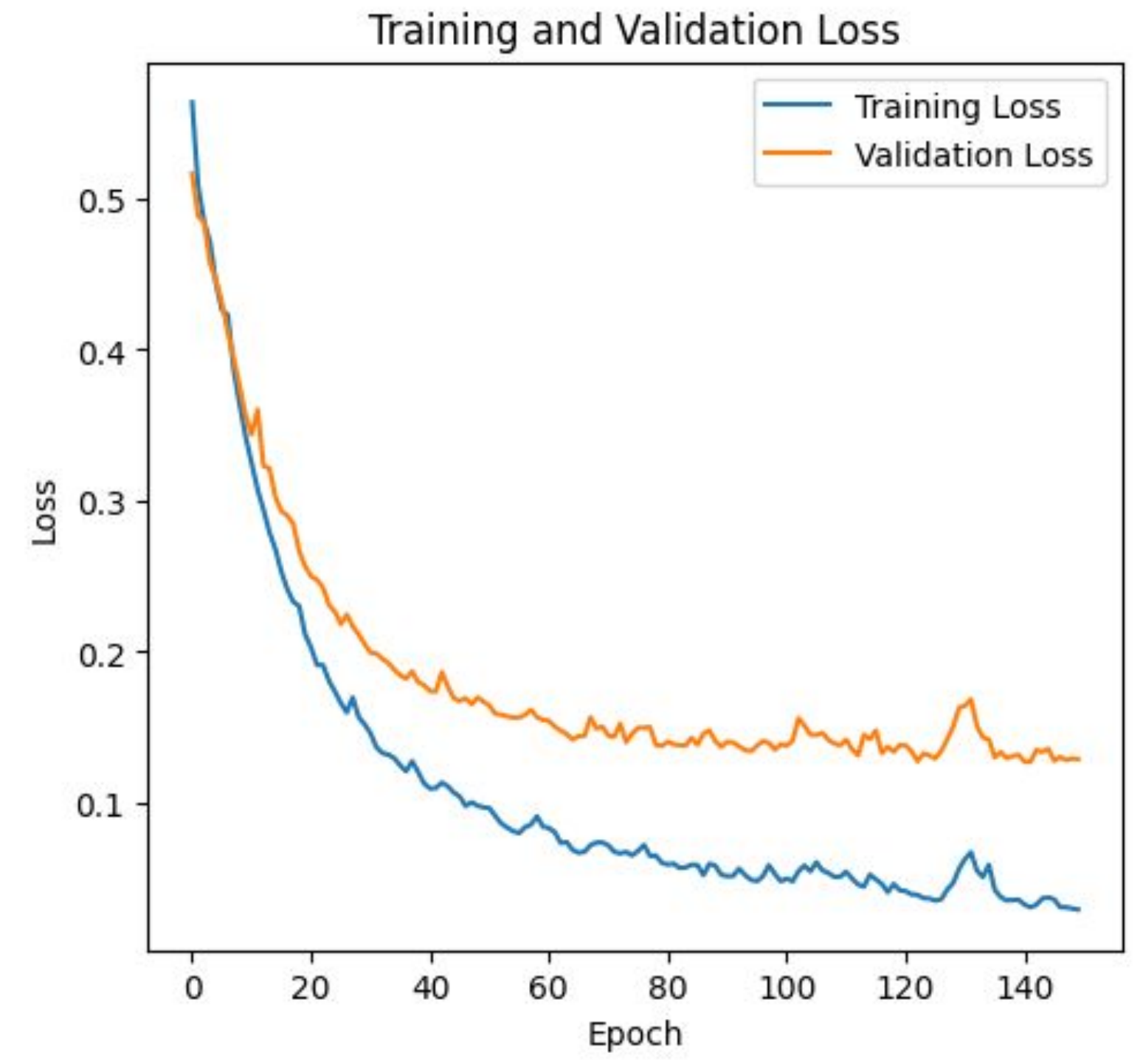
Covalent interactions
Non-covalent interactions
$(h_1^{(t)}, h_2^{(t)}, \cdots, h_N^{(t)})$
$(r_1, r_2, \cdots, r_N)$
$h_j^{(t)}$
$\|r_j - r_i\|^2$
RBF
Linear
$\odot$
Σ_j
m_i
$h_i^{(t)}$
$\oplus$
$(x_1, x_2, \cdots, x_N)$
$(r_1, r_2, \cdots, r_N)$
Linear
$(h_1^{(0)}, h_2^{(0)}, \cdots, h_N^{(0)})$
Heterogeneous Interaction
$(h_1^{(1)}, h_2^{(1)}, \cdots, h_N^{(1)})$
Heterogeneous Interaction
$(h_1^{(2)}, h_2^{(2)}, \cdots, h_N^{(2)})$
Heterogeneous Interaction
$(h_1^{(3)}, h_2^{(3)}, \cdots, h_N^{(3)})$
MLP
Prediction of druggability

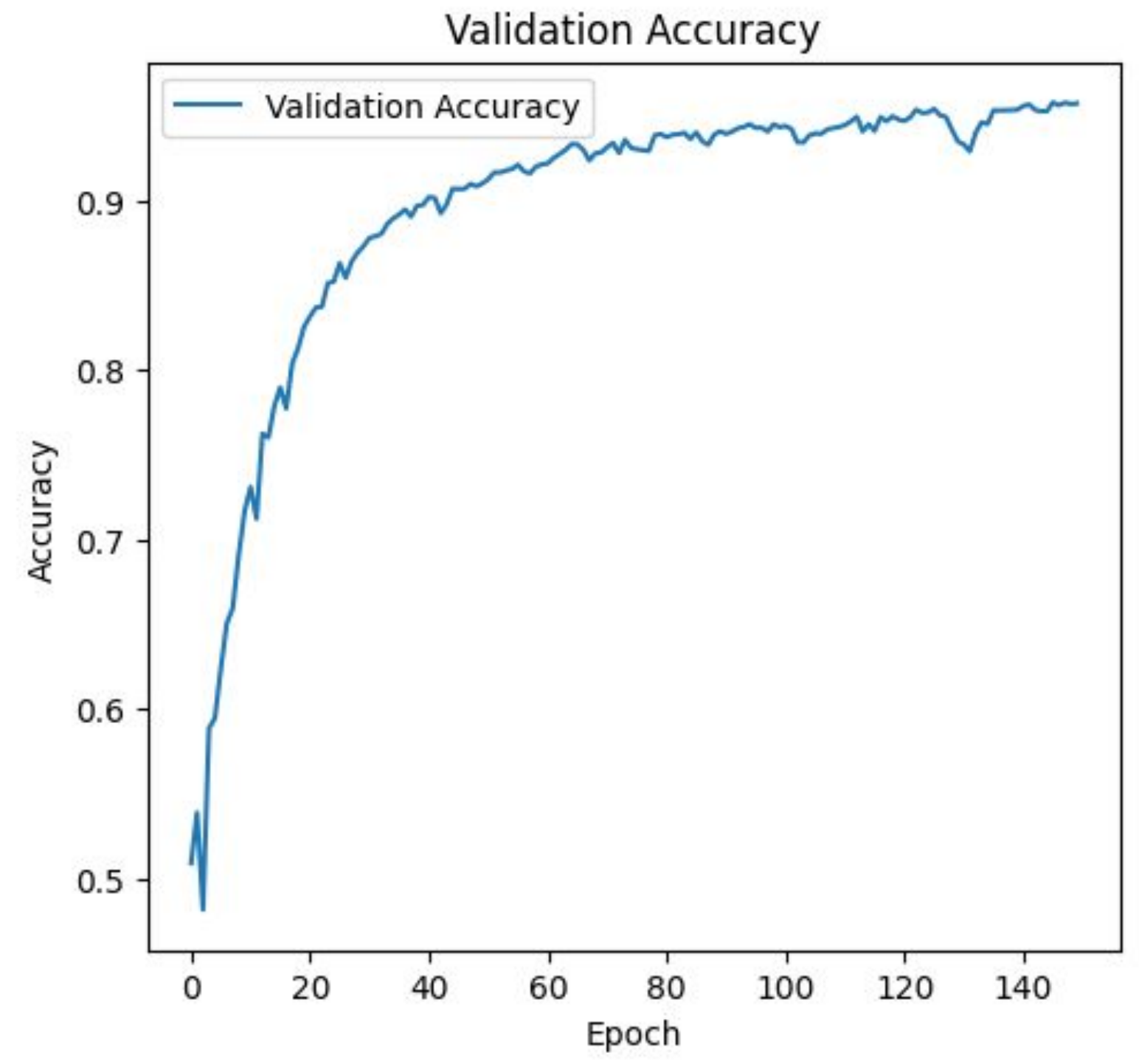
- Heterogeneous layer handles covalent and non-covalent interactions separately

PPI Model Results

- Stats after 150 epochs
  - **Train loss: 0.0294**
  - **Validation loss: 0.1289**
  - **Validation accuracy: 0.9572 (95.72%)**
  - **Training time: 11 min on laptop**
- Some overfitting, yet high validation accuracy
- Correctly classified nodes in green; incorrectly classified nodes in red





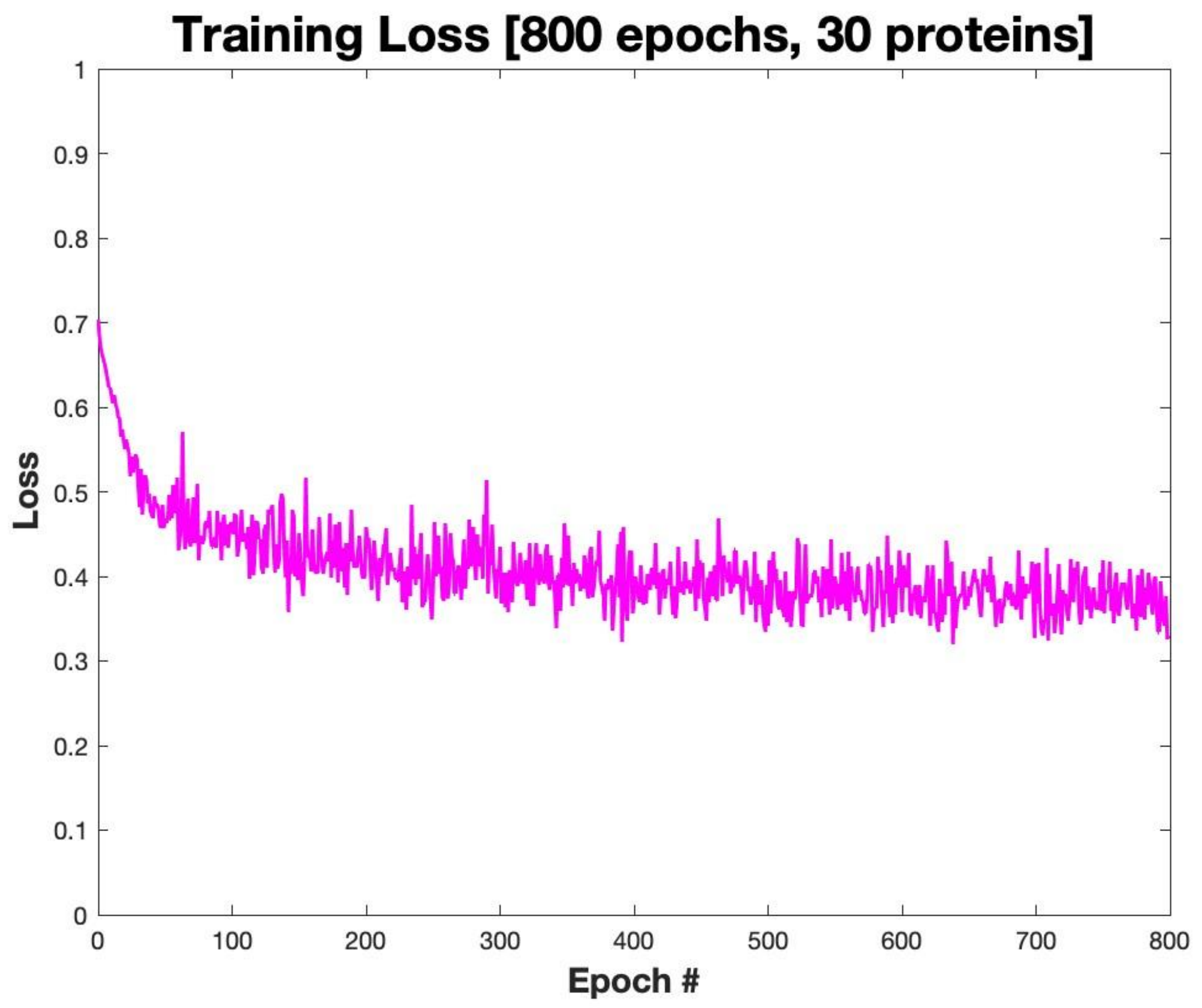


Druggability Classifier Results

- Stats after 800 epochs using 30 protein subset:
  - **Train loss: ~0.35** – gradually decreasing but stagnant (overfitting)
  - **Test accuracy: 0.818 (81.8%)** – poor metric due to heavily class-imbalanced dataset (<10% positive label nodes)
  - **Recall: 0.175 (17.5%)** – very poor due to very small protein subset
  - **Training time: 20 min on supercomputing cluster**
- Very long time to train with full 3,600+ protein dataset (>2 days for 200 epochs)
  - 1m+ nodes, 1b+ edges; in progress

CONFUSION MATRIX		**Actual labels**	
		Positive	**Negative**
Predicted labels	**Positive**	197 (TP)	**2,476 (FP)**
	Negative	929 (FN)	15,552 (TN)

- Druggability of a protein cavity concluded from its constituent atom predicted labels
- **False positives could indicate unidentified druggable cavities**
  - High false negative rate and false positive rate, but both druggable and undruggable predictions are being made
  - Focus on hyperparameter and model architecture tuning and visualization



Conclusions & Future Work

- **Due to the intrinsic graph-like structure of biochemical data, GNNs are well-suited for handling related predictive tasks**
- Adding more hidden channels and hidden layers could increase the accuracy of the GNN applied to the PPI dataset by *enabling longer-range message passing*
- Training with full protein-ligand dataset for druggability classifier will increase recall score and overall performance; *investigate positive unlabeled learning and self-supervised representation learning of atom embeddings*
- **Validated drug target candidates from this model could accelerate preclinical studies & lower treatment costs, particularly for rare diseases**

References & Acknowledgements

- Dr. Phillip Gingrich, Dr. Bissan Al-Lazikani (U.T. M.D. Anderson Cancer Center)
- Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P., & Al-Lazikani, B. (2013). Objective assessment of cancer genes for drug discovery. *Nature Reviews. Drug Discovery, 12*(1), 35–50. https://doi.org/10.1038/nrd3913
- Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021). A gentle introduction to graph neural networks. *Distill, 6*(9), e33. https://doi.org/10.23915/distill.00033
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks* (arXiv:1710.10903). arXiv. https://doi.org/10.48550/arXiv.1710.10903
- Yang, Z., Zhong, W., Lv, Q., Dong, T., & Yu-Chian Chen, C. (2023). Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures(Gign). *The Journal of Physical Chemistry Letters, 14*(8), 2020–2033. https://doi.org/10.1021/acs.jpclett.2c03906
- https://github.com/shuowang-ai/graph-neural-network-pyg/blob/master/ppi.py