A Graph Convolutional Neural Network for Inhibitor Screening and Pharmacophore Modeling of Prolyl-tRNA Synthetase

Matias Vantilburg, Sanchita Hati, and Sudeep Bhattacharyay
Department of Chemistry and Biochemistry, UW-Eau Claire, Wisconsin-54701

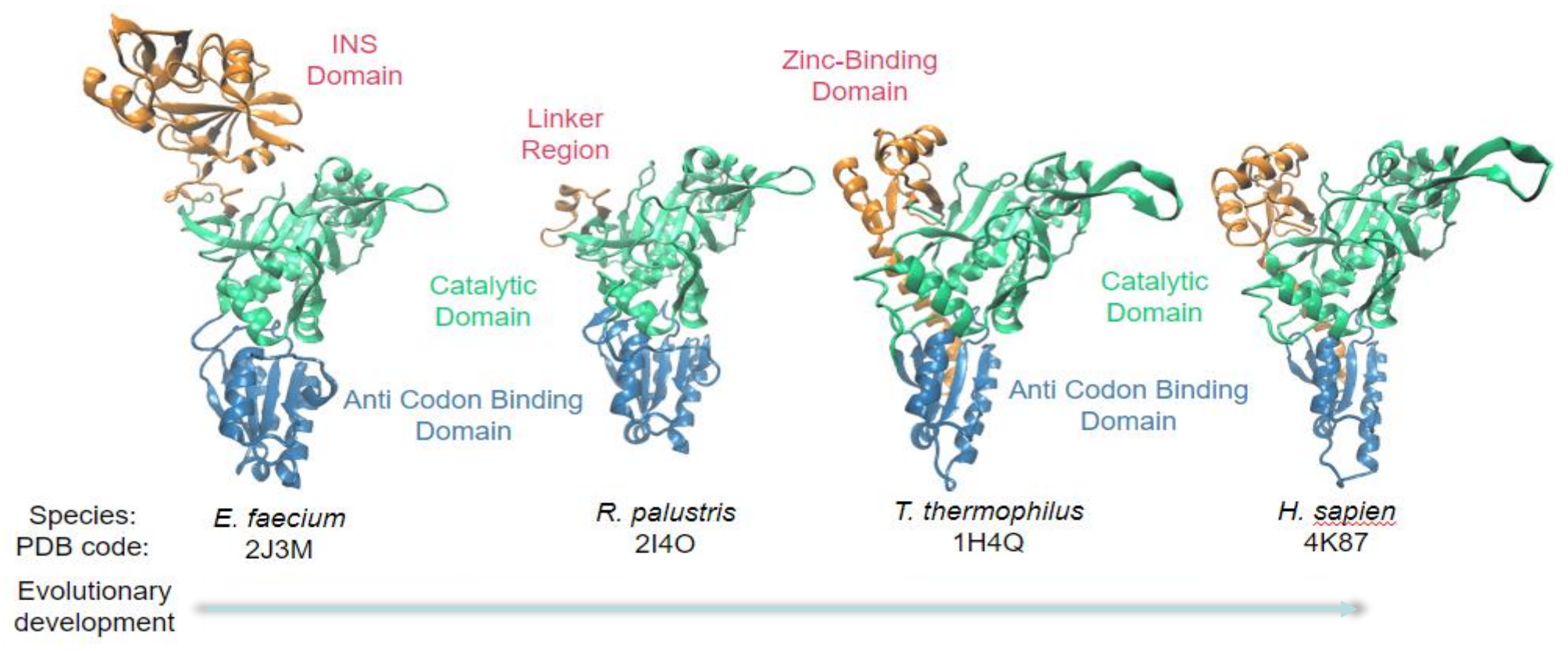
Abstract

Chemical knowledge traditionally operates in terms of important groups relevant to a molecule's properties. This meshes poorly with current machine learning approaches, which are difficult to interpret or incorporate by chemical engineers.

Some methods exist for interpretation of predictive models, but these are incapable of the exploration that is necessary to find selective binders, a notoriously difficult domain. Generative models, while are among the most promising in cheminformatics for the ability to generate small molecule drug leads while optimizing for specific properties, are particularly guilty of this.

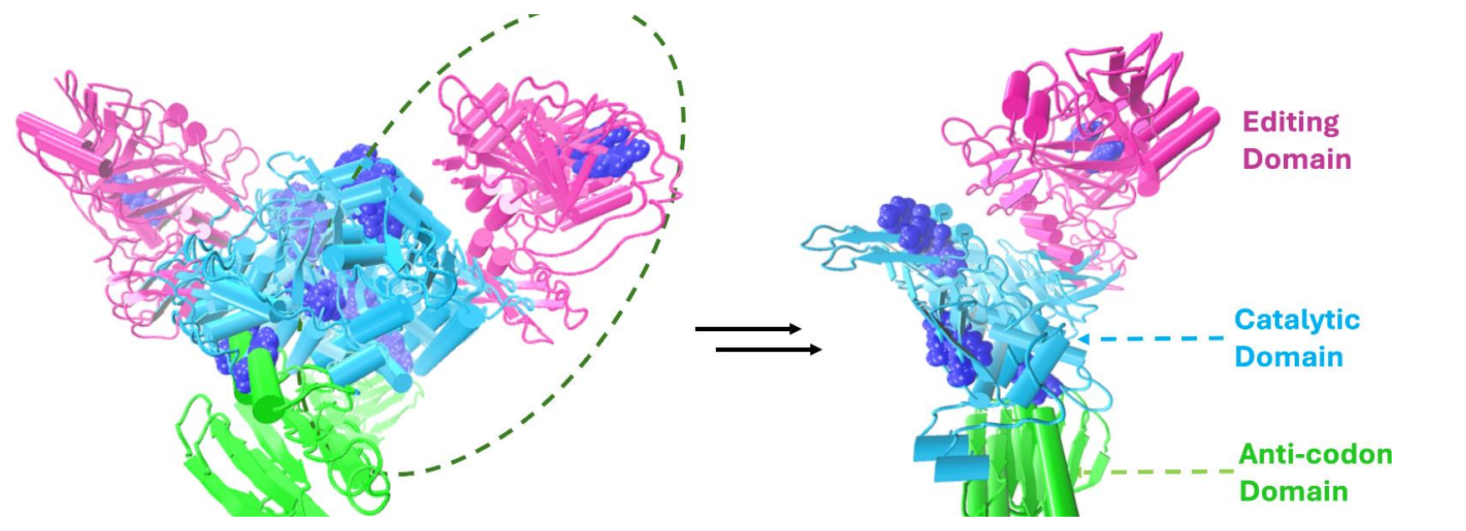
This project aims to combine these two approaches by first identifying the 'most relevant pharmacophores' for an arbitrary molecular prediction task, and then generating a chemically valid molecule incorporating that pharmacophore.

This project aims to apply to this approach to the difficult task of docking selectivity among similar tRNA synthases, a common target for antibiotic drug discovery research. Validation is through running classical docking simulations on the generated molecules.



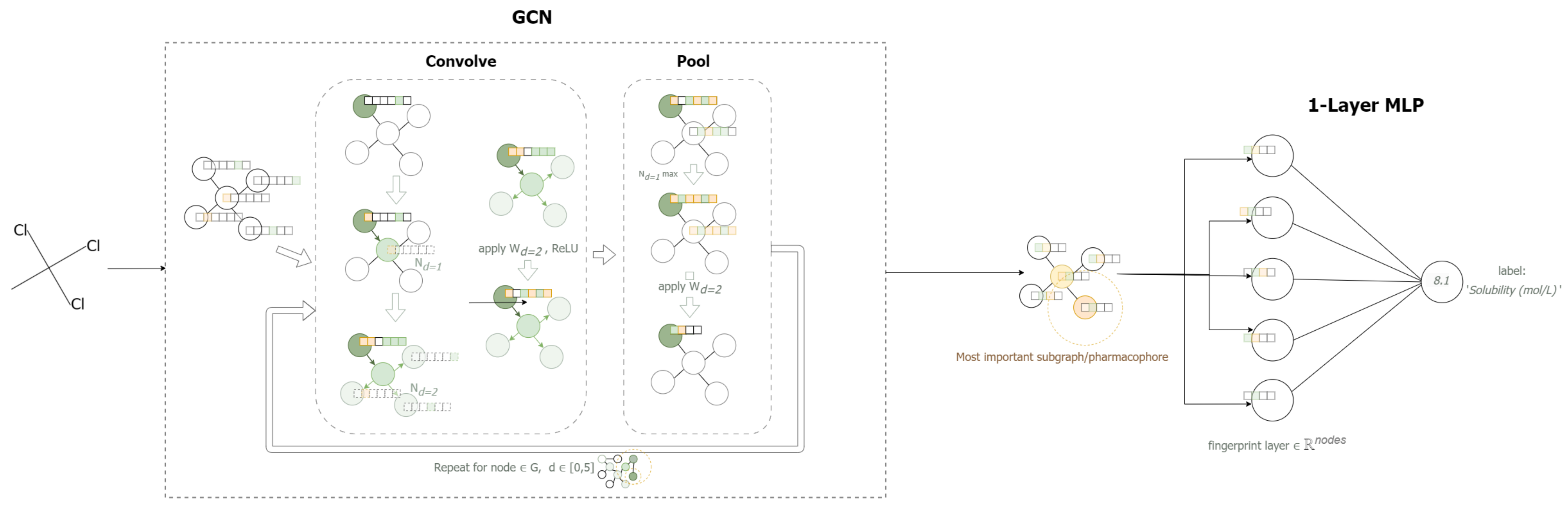
Motive

- AARSs are a common target for antibiotic drug discovery research
- Structural variations common in AARSs from different species
- Possibility for pathogen-specific AARS inhibitors
- Prolyl-tRNA Synthetases (ProRSs) - a specific example of an aminoacyl tRNA synthetases (AARS), chosen for it's complexity



Methods

- **Model A: Pharmacophore Identifier**
  - Training:
    - Graph convolutional layers producing a final set of node embeddings per atom-degree pair
    - Fed into a deep layer to predict labels (ex: binding affinity, solubility, etc)
      - 1-layer to maintain correlation between embeddings & labels
  - Evaluation:
    - Compare each fingerprint feature vs labels, calculate R², return most anti/correlated features
    - Atom-degree pair with highest activation for most correlated feature is the 'most correlated pharmacophore'


GCN
Convolve
Pool
Cl
Cl
Cl
$N_{d=1}$
apply $W_{d=2}$, ReLU
$N_{d=2}$
$N_{d=1}$ max
apply $W_{d=2}$
Repeat for node ∈ G, d ∈ [0,5]
Most important subgraph/pharmacophore
1-Layer MLP
8.1
label: 'Solubility (mol/L)'
fingerprint layer ∈ $\mathbb{R}^{nodes}$

- **Model B: Small Molecule Generator**
  - Training:
    - Graph convolutional layers producing a final set of node embeddings per atom-degree pair
    - Aggregated (summed) to produce subgraph embedding
    - Next-Node Prediction:
      - Subgraph embedding fed into two multilayer perceptron's, predicting the Gaussian mean & variance of the feature labels (e.g. C,N,O,P,… one-hot encoded; dequantized)
    - Next-Edge Prediction:
      - Subgraph embedding, origin node embedding, and destination node embedding all fed into two MLPs predicting Gaussian mean & variance of the bond labels (e.g. single bond, double,… one-hot encoded; dequantized)
      - Reject unless valid # of bonds
    - To train, iteratively feed in subgraph i, predicting node i+1 bond features and (node i->node j) edge features for all j in subgraph I
    - Loss is negative log likelihood with penalty term for variance

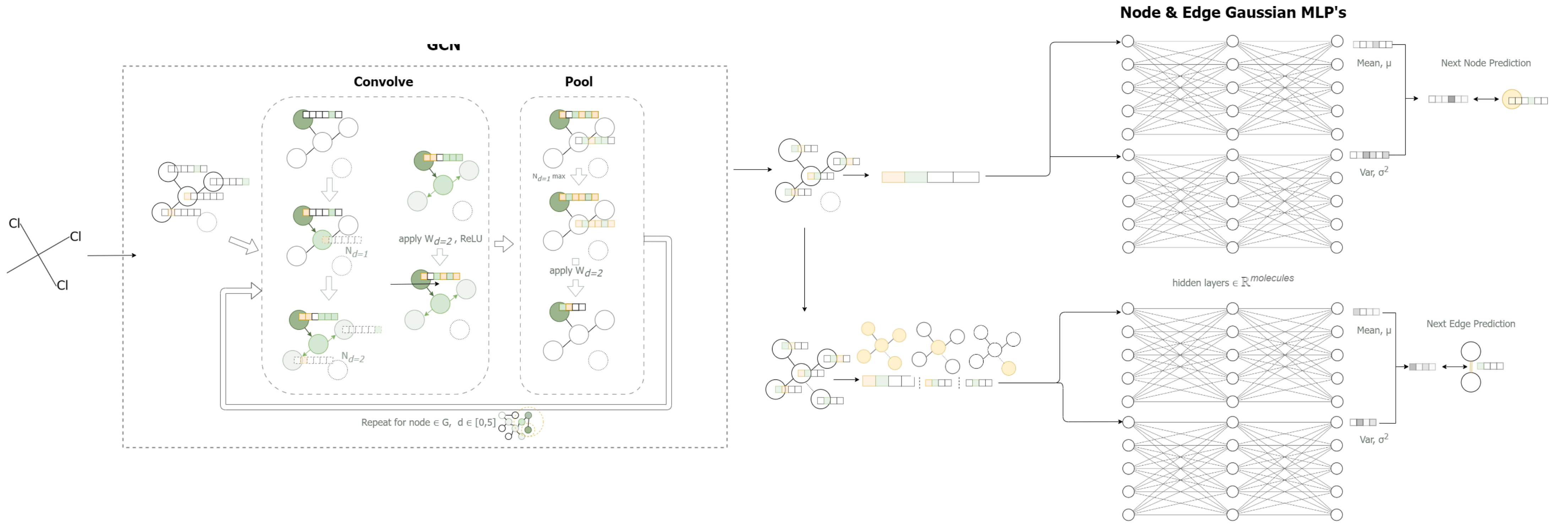
ϵ = Standard normal error = (true – mean)/var

$$\epsilon_i = \left(z_i^X - \mu_i^X\right) \odot \frac{1}{\alpha_i^X}$$

Loss = error term (combined gaussian densities of ϵ) + penalty term for variance

$$\mathcal{L}_i^X = -\log(\mathrm{Prod}(p_{\mathcal{E}}(\epsilon_i))) - \log(\mathrm{Prod}(\frac{1}{\alpha_i^X}))$$

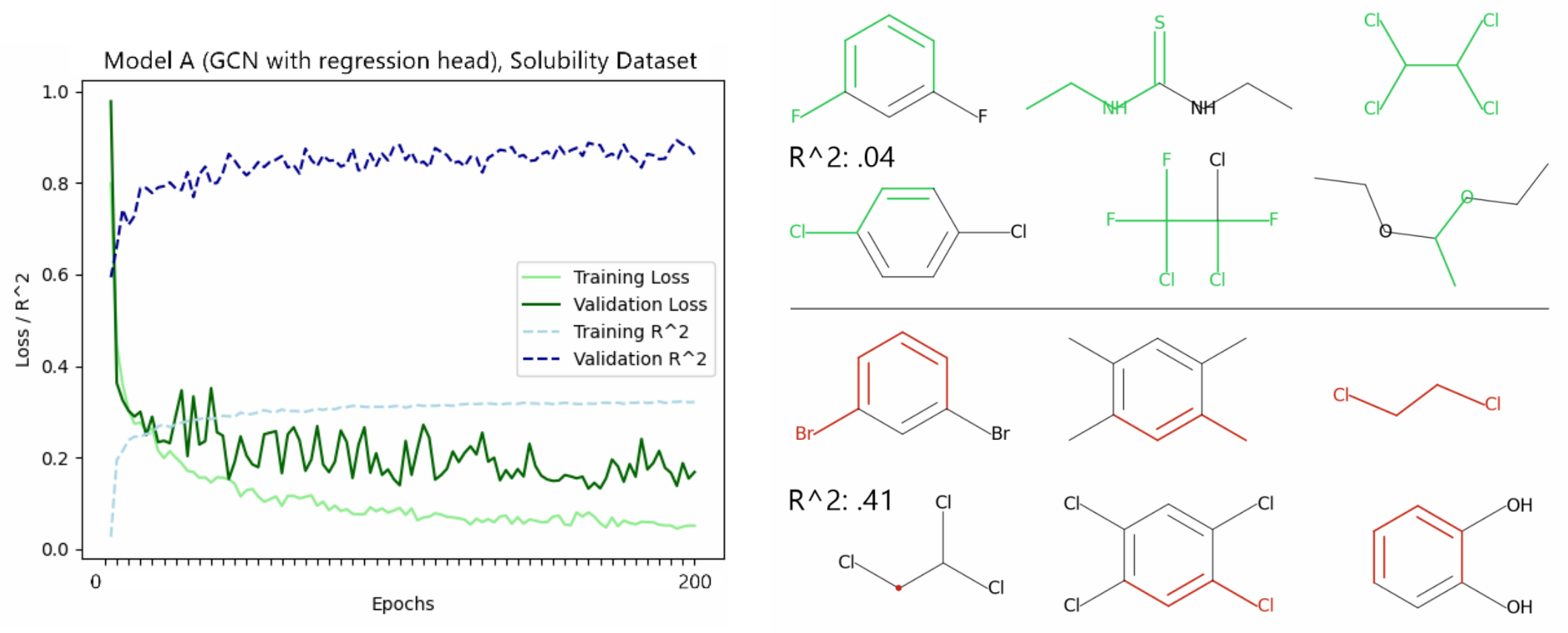
  - Generation:
    - Begin predicting from final atom in pharmacophore
    - Predict next node/edge mean, var vectors; sample var, add & argmax to get predicted type



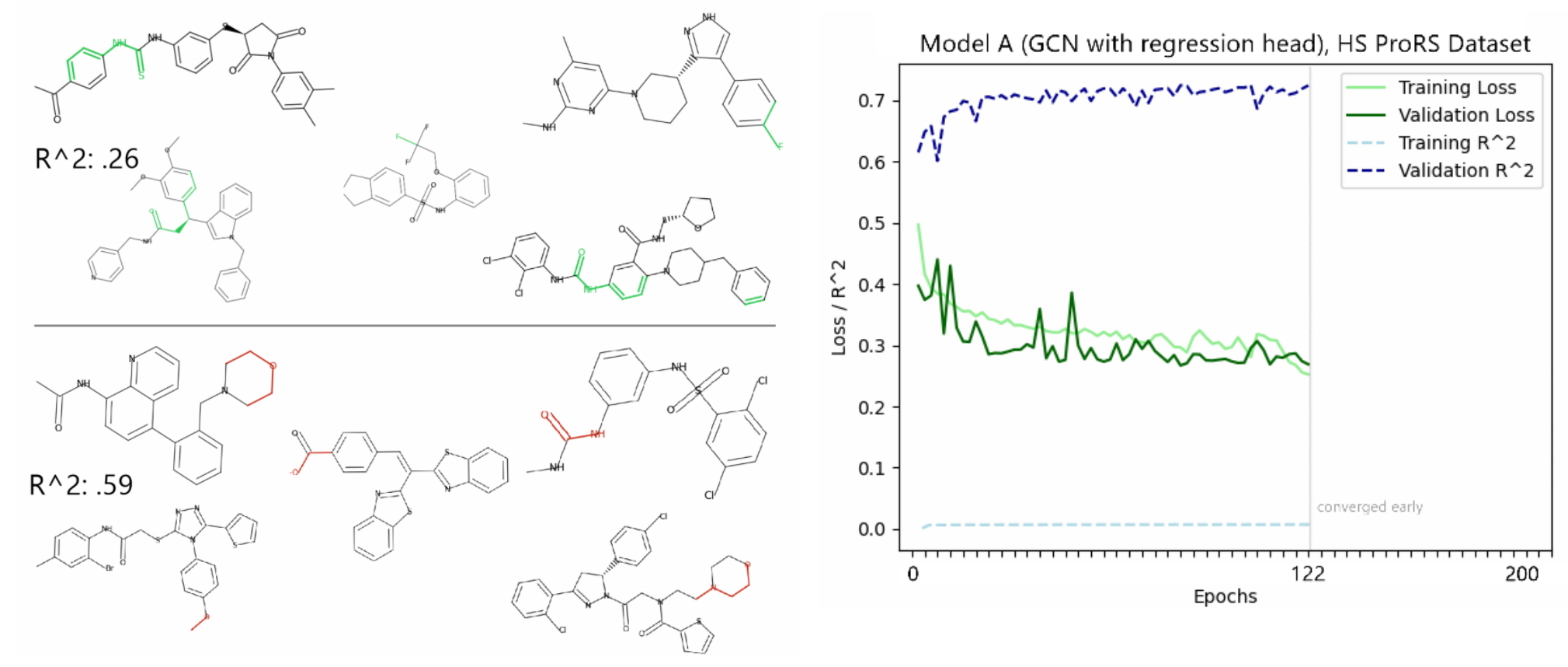
Preliminary Results

Below are the pharmacophore's produced by Model A on various datasets: solubility as an easily interpretable case example to judge performance, as well as three different species-specific AARS target binding affinity datasets intended to be analyzed.

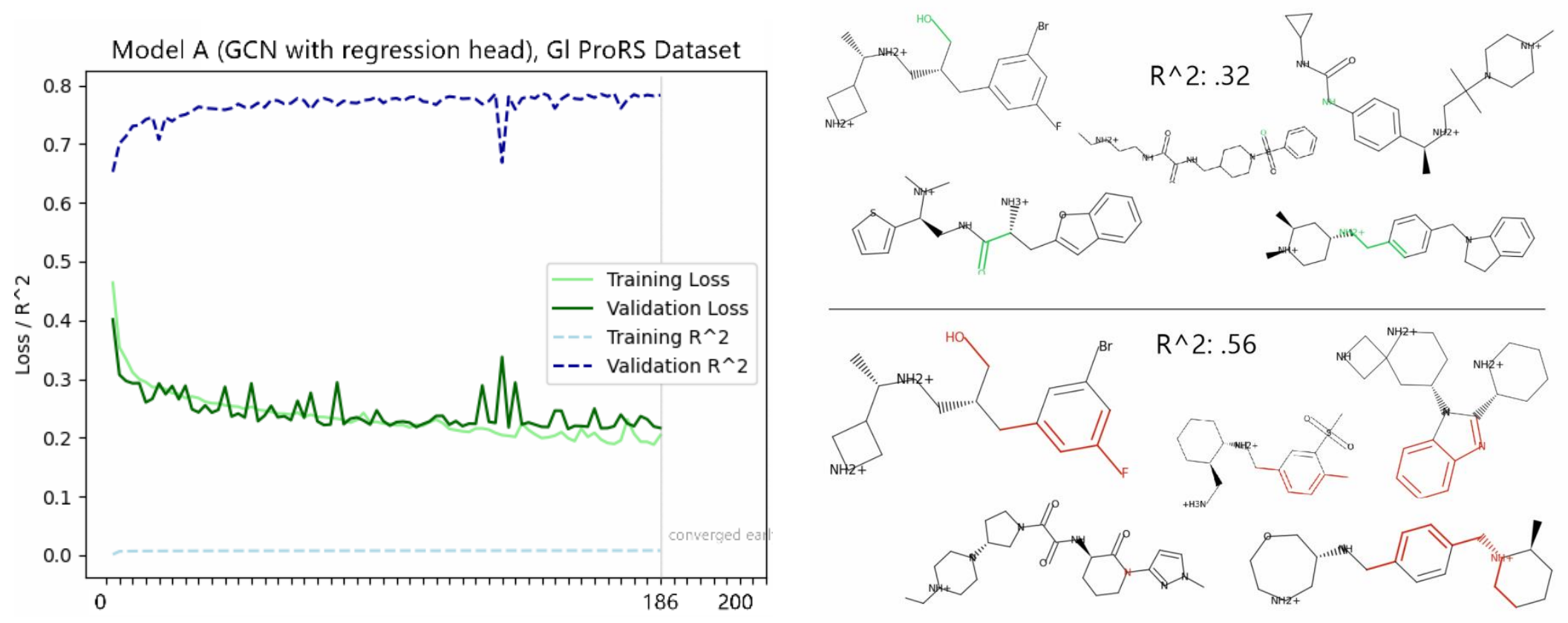
- Solubility Dataset:



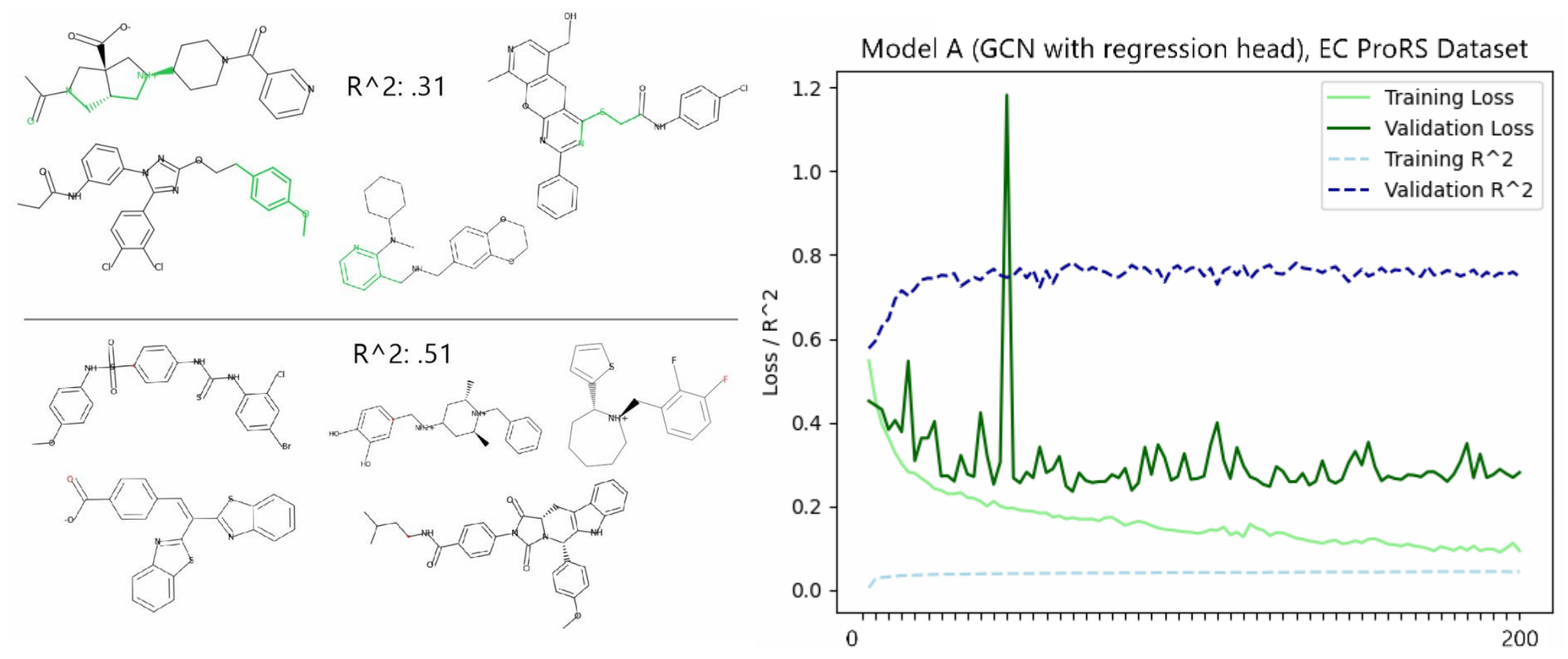
- Human variant ProRS-binding Dataset



- Bacterium (*E. coli*) variant ProRS Dataset



- Cyanobacteria (*G. violaceus*) variant ProRS Dataset



Bibliography

1. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv. https://arxiv.org/abs/1509.09292
2. Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C.-Y., & Hou, T. (2023). Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-38192-3
3. Delaney, J. S. (2004). ESOL: estimating aqueous solubility directly from molecular structure. Journal of Chemical Information and Computer Sciences, 44(3), 1000–1005. https://doi.org/10.1021/ci034243x
4. Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., & Tang, J. (2020). GraphAF: A Flow-based Autoregressive Model for Molecular Graph Generation. arXiv. https://arxiv.org/abs/2001.09382

Acknowledgments

- We acknowledge that the University of Wisconsin-Eau Claire occupies the sacred and ancestral lands of Indigenous Peoples. We honor the land of the Ojibwe and Dakota Nations.
- Office of Research and Sponsored Programs, UW-Eau Claire
- Blugold Center for High-performance Computing, UW-Eau Claire
- The National Science Foundation, Award number: OAC 2150191