Conducting Phylogenetic Inference on B Cells with Language Transformers

Danielle Justo (Mentor: Dr. Sanket Rane, 2026 IICD SRP)

Combining modern single-cell sequencing technology with phylogenetic inference can lead to a better understanding of immune responses and aid in the development of vaccines and immunotherapy treatments. Common approaches to the reconstruction of phylogenetic trees include algorithmic approaches, like distance-based methods, and statistical approaches, like maximum likelihood and Bayesian inference, however less explored is the integration of machine learning in these frameworks. Transformers have proven capable of extracting contextual dependencies between variables in complex, high-dimensional data, such as that derived from DNA sequencing. Additionally, prior work in the lab demonstrates that a vision transformer can learn latent representations of Spatiotemporal data generated from dynamic models, aiding in mechanistic inference such as the estimating parameters of the data generation process. We seek to build on this research, investigating whether sequential data processing transformers, such as language models, are capable of uncovering similarly useful latent representations from DNA sequencing data that can be used in the reconstruction of B cell phylogenetic ancestry.

References

Park, J. W., Zhao, K. & Rane, S. Spatiodynamic inference through vision-based generative modeling, *arXiv* (2025).