TITLE: Prediction-based inference for large-scale scientific inquiry

SUPERVISOR: Jessica Gronsbell

From structural biology to epidemiology, predictions from machine learning (ML) models increasingly complement costly gold-standard data to enable faster, more affordable, and scalable scientific inquiry. For example, DeepMind's AlphaFold, a generative AI model, predicts protein structures from amino acid sequences in seconds, whereas gold-standard experimental methods take months and cost over $100,000. AlphaFold's predictions have accelerated proteomic research and deepened our understanding of how protein mutations contribute to disease. However, the scalability of answering scientific questions with ML predictions comes at a cost. Predictions inevitably contain errors relative to gold-standard data, which can bias scientific conclusions if not properly accounted for in downstream statistical inference. This risk is especially concerning in big data settings, where increased precision and potentially high bias can lead the scientific community to be more confident in the wrong answers.

This project builds on our emerging research program in prediction-based (PB) inference, that is, statistical methods that leverage a large volume of predictions of outcomes and/or covariates together with a small amount of gold-standard data to reliably answer scientific questions. This project will focus on the development of a PB inference method that optimally combines predictions from multiple ML models, with theoretical guarantees for bias mitigation and improved efficiency. We will also rigorously evaluate our method with analyses of protein structures and release open source software.