GREMLIN: Graph Estimation from MR images Leading to Inferences in Neuroscience

A theme in modern scientific research that is increasingly shifting into the spotlight is reproducibility - are studies and findings developed in such a way that their claims can be meaningfully leveraged for future scientific work or applied across different data populations. In computational neuroscience, this is regularly referred to as a "reproducibility crisis," as inflated p-values and studies which fail to generalize are unfortunately common. During my Master's degree in The Johns Hopkins University Department of Biomedical Engineering, I developed software tools and statistical methods to combat this crisis in connectomics, the study of developing and understanding connectivity networks within the brain. I developed a one-click open-source pipeline for structural connectome estimation from Magnetic Resonance Imaging (MRI), and released the largest-ever open-access database of human brain connectivity maps (available at http://m2g.io). The pipeline I developed, ndmg (available on GitHub at neurodata/ndmg), was optimized for four main qualities: accuracy, reliability, robustness, and usability.

To ensure accuracy, I leveraged many off-the-shelf algorithms which have demonstrated their efficacy at producing results that are neuroanatomically valid. I implemented a variety of quality control figures within the pipeline which produced summaries, both numeric and visual, of the derivatives being produced so that users would be able to interact with the data being processed and evaluate if errors may have occurred at a given step.

To ensure reliability, I developed a statistic, named discriminability, and a corresponding statistical test which evaluated the consistency of derivatives within populations. In essence, this statistic represents the strength of the relationship between all samples or observations in a study, and compares that to the expected relationships. Considering the case of brain-derived networks/graphs, one would expect that two graphs from different observations of a single participant would be much more strongly related than between two different individuals. This is a particularly necessary test as no ground-truth data is available for brain graphs. Optimizing over this statistic, I was able to verify that there is participant-specific signal in the derived graphs.

To ensure robustness, I jointly optimized the discriminability statistic described above over many datasets which were produced across widely varying parameters, scanner types, geographical locations, and resolutions. This joint optimization ensured that the pipeline was able to serve as a one-click solution that could be generally used across disparate datasets.

To ensure usability, I packaged my pipeline in Python, Docker, Singularity, and integrated into a variety of high performance computing environments and platforms across Canada and the United States, as well as the Amazon commercial cloud. This lowers the barrier to entry for any neuroscientist or researcher who wishes to use my pipeline, to simply require the access of a web-portal rather than requiring computational experience or expertise.