**Gerstein Lab's Experience in Integrative Deep-Learning Models for Brain Traits Using Genomics and Imaging (Final RPPR)**

This project aimed to develop advanced computational tools that connect multi-omics data (such as genetic and brain imaging data) to various traits related to human brain function, including cognitive abilities and mental health conditions such as Alzheimer’s Disease, schizophrenia, bipolar disorder, and autism. The ultimate goal was to improve our understanding of core brain functions (like memory & attention) and disease mechanisms to help the development of personalized treatments and interventions.

Our team first combined large-scale datasets encompassing genomics, transcriptomics, and brain imaging information as integrated dataset: we processed and harmonized extensive data from various public sources, such as the BRAIN Initiative (BI), NeMO, GTEx, BrainSpan, PsychENCODE, the Allen Human Brain Atlas (AHBA), the UK Biobank, OpenNeuro, the Human Connectome Project (HCP), and the Bipolar and Schizophrenia Network for Intermediate Phenotypes (BSNIP). Such a comprehensive approach allowed us to integrate diverse biological information and thus enhance the scale and scope of our analysis.

Our team then developed novel and robust predictive models that take the integrated dataset as input and estimate individual genetic risk for different cognitive and psychiatric conditions. One key innovation of our project was the development and refinement of the Linear Network of Cell Type Phenotypes (LNCTP) model, which is designed to process genetic and single-cell gene expression information. This model enabled researchers to not only predict disease risk but also gain insights into specific cellular mechanisms and pathways involved in brain-related conditions.

To add interpretability for research insights, we integrated advanced statistical techniques such as Graphical LASSO and Gaussian Markov Random Fields (GMRFs). Graphical LASSO helped identify important gene interactions by enforcing sparsity in the model, thus pinpointing the most critical genetic features. While the GMRFs approach further improved the precision and robustness of the model with high-dimensional data (i.e. genomic). These improvements resulted in both the accuracy of predictions and the interpretability of the biological relevance of predictions, which helped the identification of key genes, GRNs, and brain regions that are related to disease processes.

We extensively validated our models through multiple experiments, demonstrating their high predictive performance compared to traditional methods. Our model also demonstrated robustness via its generalizability across independent datasets, indicating the potential application in neuroscience research and clinical settings.

This project supported the training and development of early-career scientists. Our interdisciplinary and collaborative environment allowed trainees, including researchers and students to acquire valuable skills in neuroscience, computational biology, statistical modeling, and data integration. Team members received consistent mentoring through weekly meetings and benefited from feedback and guidance from senior investigators. In addition, our team members have also participated in both national and international scientific conferences, where they present findings, engage in discussions, and expand professional networks.

To maximize the project’s impact and facilitate future research advancements, all computational tools, predictive models, software, datasets, and analytical pipelines developed during this project were made openly available on platforms such as GitHub. This commitment to open science ensures that our contributions can be easily accessed, replicated, and built by the community.

Finally, significant scientific findings resulting from this project have been published through various peer-reviewed scientific journals, which includes rigorous validation studies, methodological innovations, and novel insights.