**Project: Identifying disease markers of Parkinson’s Disease**

Team Leads: Shreya Menon and Alex Silva

**Elevator Pitch**

Using machine learning to discover markers of Parkinson’s Disease in multi-omic data

**Abstract**

Parkinson’s disease (PD) is the second-most common neurodegenerative disorder and is the fastest growing neurological disorder in terms of cost to disability-adjusted life-years. While there exist therapies that improve quality of life, no treatment currently reverses and completely stops disease progression, presenting an unmet need for a deeper understanding of disease pathology. The underlying etiology of PD is likely due to a complex interaction between environmental, genetics and aging-related factors. The molecular pathophysiology of PD is complex and involves the disruption of multiple different cell types.

This project aims to train and implement machine learning models on chromatin accessibility and gene expression data to find potential markers of PD and genes that may mediate disease. The data that we will use are cell type specific gene expression count values for and accessibility values for genomic regions. This project offers an opportunity to gain exposure to key steps in common machine-learning practice: feature engineering, modeling, and interpretation of model weights. We will first leverage feature selection techniques, such as principal components analysis and LASSO, to find combinations of gene counts and accessibility values associated with PD. We will then apply machine learning methods ranging from simple logistic regression to artificial neural networks to probe the extent to which PD diagnosis and progression can be predicted from our genetic features. Finally, we will look at the effect sizes for linear models and saliency mapping for non-linear neural networks to investigate the genetic features most important for predicting PD disease status and aging.

If time permits, we will also attempt to integrate protein expression data in our overall aim to identify markers of PD.

**Zoom link:** [**https://ucsf.zoom.us/meeting/tJwkdOytpzojH90t-8AUEaNe05CwE\_gPjJHp/ics?icsToken=98tyKuCgpzgvH92UuRyORow-HYjCZ-\_xmH5EgvoNiFK9IhV2Yy3zZrdhJZdAF-ja**](https://ucsf.zoom.us/meeting/tJwkdOytpzojH90t-8AUEaNe05CwE_gPjJHp/ics?icsToken=98tyKuCgpzgvH92UuRyORow-HYjCZ-_xmH5EgvoNiFK9IhV2Yy3zZrdhJZdAF-ja)

**Meeting ID: 989 3559 1959**

**Password: 105818**

**Data Source:**

* 69,289 nuclei with matched chromatin accessibility and gene expression data from 31 individuals (9 young donors, 8 aged donors, 14 PD patients)
* <https://www.medrxiv.org/content/10.1101/2022.01.18.22269350v4.full-text>

**Key Learning Objectives:**

Clustering:

* Use unsupervised clustering algorithms to cluster cells based on off gene expression and chromatin accessibility
  + Which clusters emerge? Do certain patients cluster together?

Classification:

* Can you classify aged vs control cells based on gene expression values? PD vs control?