BioE 145/245 Final Project, 2025

We have discussed applications of single cell RNA-seq and the computational challenges that arise from this complex data. In this project, you will work directly with scRNA-seq data in a real-world application.

The project is due **Friday May 16th (the last day of finals week) at 11:59pm**, on Gradescope. Remember that you cannot use slip days on the final project. Many students prefer to finish the project by the last day of classes so that you can keep your RRR week free for studying for other classes! We will have some office hours during RRR week.

This project has **two parts that you will hand in at the same time**, but you are strongly encouraged to finish part 1 soon so that you can use it for part 2 during lab sections.

Peripheral blood mononuclear cells (PBMCs) are a group of different cell types in your blood, a subset of what we call white blood cells. They include many important parts of your immune system: T cells, B cells, natural killer cells, and so on. These cells are responsible for your innate and adaptive immune responses. When you get infected or vaccinated, they kick into high gear. In different situations, or with different diseases, the mix of different cell types shifts and so does the gene expression within one type of cell. PBMCs can be isolated very easily from patient blood samples, just by spinning the blood in a centrifuge that separates different cell types by weight. So, looking at these cells can be very informative.

Part 1: Lower-dimension representations of the cells

Starting from scRNA-seq data from PBMCs, you will look at the cells in a lower-dimension space. You will implement an autoencoder to find a latent space representation of your data. Then, you will compare two-dimensional representations of your data using t-SNE, PCA, and the latent space defined by your autoencoder. We will give you the scRNA-seq data as a count matrix (already normalized and cleaned up) with cell type labels that were assigned by the original researchers, and guidelines on how to implement and use the autoencoder.

For part 1, you will hand in **figures of the data from your three different dimensionality reduction approaches, with cells colored by the cell type labels**, plus a **one paragraph justification describing which one you think performed best and why**. You will also submit a **python notebook with the relevant code**.

Part 2: Classifier

You will implement a classifier to classify unlabeled cells as one of the cell types labeled in the original data. In lab section, we will discuss classifier methods (random forests, ensemble methods, etc.) and more specific guidelines on what accuracy to aim for and how to evaluate. For part 2, you will hand in a **short writeup (one paragraph) about your classifier choice and implementation**, and **figures or data showing the performance of your classifier**. You will also submit a **python notebook with the relevant code**.

Again, the full assignment will be due **May 16th**, with a short writeup of part 1 and part 2, plus the notebooks. For each part, ***the figures and writeup must be submitted together as a pdf, not embedded in the notebook.***