10/12/2025 Presentation date: Dec 02, 2025

BIOINF 593 Project Proposal

Project Members: Andre Gala-Garza, Prema Immadisetty, Jason(Jingqiao) Zhao

Problem Statement:
The central challenge in single-cell cancer analysis is that raw gene expression counts, while quantitative, lack direct functional context. They tell us *which* genes are active but not necessarily ***how*** their protein products are functioning to drive the cell's behavior.

Our project aims to solve this by creating a new, **function-aware embedding for each individual cell**. By fusing a cell's transcriptomic profile with the rich, functional information pre-trained into **ProteinBERT**, we will generate a vector that captures the cell's biological state.

Ideally if the idea of cell level embedding works, in the high-dimensional feature space defined by these embeddings, cancerous and healthy cells will form distinct, more separable clusters and **might have a universal general direction that represents cancer**. Our final goal is to learn a **decision boundary** within this "functional space" to classify individual cells and rigorously evaluate how effectively our protein function-informed representations distinguishes cell states.

Datasets:

- **Tumor data:** “A single-cell and spatially resolved atlas of human breast cancers”: https://cellxgene.cziscience.com/collections/dea97145-f712-431c-a223-6b5f565f362a
- **Healthy data:** “Human breast cell atlas”: https://cellxgene.cziscience.com/collections/48259aa8-f168-4bf5-b797-af8e88da6637?utm
- **ProteinBERT (specialized in function feature capture):** https://huggingface.co/GrimSqueaker/proteinBERT

Project Challenges & Tentative Solutions:

1. **How to Encode Abundance Into the Protein Vector:**

   We'll multiply each protein embedding by its transcript count. Or normalized transcript count.

2. **How do we approach the final cell-level aggregation:**

   We want to use Attention-Based Pooling to create a smart, weighted average of all protein vectors in a cell.

3. **Gene Selection:**

   We'll filter for the top **~2,000 Highly Variable Genes (HVGs)** to reduce noise and focus on the most informative signals.

4. **How to design an appropriate Loss function for End-to-End Training:**

   We'll use **Triplet Loss** to train the model, forcing it to create a feature space where cancer and healthy cell embeddings are pushed far apart