Final Project Proposal

scGCL: an imputation method for scRNA-seq data based on Graph Contrastive Learning

## Team 5 Members:

1. Angela Mei: amei5@illinois.edu
2. Joe Hou: jueh4@illinois.edu
3. Chetan Dindukurthi: chetand3@illinois.edu

## Citations to the original paper:

Zehao Xiong, Jiawei Luo, Wanwan Shi, Ying Liu, Zhongyuan Xu, Bo Wang, scGCL: an imputation method for scRNA-seq data based on graph contrastive learning, *Bioinformatics*, Volume 39, Issue 3, March 2023, btad098, <https://doi.org/10.1093/bioinformatics/btad098>

## General problem:

The sparse nature of single-cell RNA sequencing (scRNA-seq) data arises from the biological and technical challenges encountered when measuring gene expression at the individual cell level. This sparsity indicates that numerous genes go undetected in the examined cells within the dataset. Several factors contribute to this phenomenon, including the complexities of capturing and analyzing the minute amounts of RNA present in single cells, such as:

(1) Low RNA content- Low content makes it challenging to detect all mRNA molecules present; some cells naturally contain more mRNA molecules than others, which have more chance to detect the genes.

(2) Low gene expression or gene expression variance- Natural differences in gene expression levels between cells, even within the same cell type, can contribute to the observed sparsity. Some genes are only expressed in specific cell states or conditions, leading to a large number of zeros in the data matrix where a particular gene is not active in most cells.

(3) Technical variance- The process of isolating single cells, reverse transcribing RNA into cDNA, and amplifying the cDNA before sequencing introduces technical variability. Some mRNA molecules may be lost or degraded during these steps, resulting in incomplete detection of the transcriptome.

(4) Dropout events- The most important factor contributing to sparsity, dropout events occur when mRNA molecules present in the cell are not detected, leading to a zero count for genes that are actually expressed but missed during sequencing. This is often due to inefficiencies in reverse transcription, amplification, or sequencing steps. Addressing dropout events through imputation in scRNA-seq data can significantly streamline the analysis process by filling in gaps where gene expression values are missing or undetected.

This approach not only diminishes noise in the data but also enhances the precision of clustering and classification efforts, bolsters the accuracy of differential expression analyses, and aids in the integration of data from various sources, thereby simplifying the overall complexity of handling scRNA-seq datasets.

## Specific approach:

Given the scarcity of cell type labels, contrastive learning has emerged as a powerful tool in self-supervised learning, achieving enhanced feature representation by amplifying the similarity between positive samples and diminishing that between negative ones. When applied within the context of graph theory, this approach evolves into graph contrastive learning, which is theoretically capable of discerning the intricate relationships between cells and subsequently reconstructing missing gene expression values. This integration harnesses the strength of contrastive learning, tailored to the graph domain, to effectively capture the cellular interactions and dynamics inherent in scRNA-seq data.

We introduce a method called scGCL, a single-cell Graph Contrastive Learning approach for imputing missing values in scRNA-seq data, which combines graph contrastive learning with the Zero-inflated Negative Binomial (ZINB) distribution for estimating dropout events. scGCL leverages contrastive learning to encapsulate both global and local semantic information, and it employs a strategy for selecting positive samples that enhance the representations of target nodes. To effectively model the global probability distribution of gene expression data, scGCL utilizes an autoencoder framework based on the ZINB distribution. This framework is designed to reconstruct scRNA-seq data by leveraging a prior distribution, offering a nuanced approach to addressing dropout imputation challenges.

## Hypotheses to be tested:

We hypothesize that scGCL will exhibit better clustering performance and imputation performance than other existing methods. Additionally, we hypothesize that scGCL can enhance differential gene expression analysis in disease datasets, such as Alzheimer’s disease or Huntington’s disease.

## Ablations planned:

In the ablation study conducted on the Adam and Romanov datasets, two key components were systematically removed to assess their impact on the system's performance:

1. The first scenario involved the removal of the ZINB-based encoder while maintaining the AFGRL and graph convolutional models. This allowed us to observe how the system performs without the contribution of the ZINB encoder, focusing solely on the effects of the AFGRL framework and the graph convolutional approach.
2. In the second scenario, the ZINB-based encoder was retained, but the graph convolution and AFGRL models were excluded. This setup aimed to isolate and evaluate the significance of the ZINB-based encoder by removing the influence of the other two components, offering insight into its standalone contribution to handling the datasets.

## Description of how you will access the data:

The dataset for this experiment is available to download from the [PyTorch Geometric library](https://pytorch-geometric.readthedocs.io/en/latest/). This library is built upon the normal PyTorch library for a specific reason: to write and train GNNs easily. The data is then downloaded and stored in an H5 file, to be used as an input for our model.

## Feasibility of the computation:

In the preprocessing phase, only 2048 genes with high variability were chosen for further analysis and used in the training process. The number of cells in the dataset is also a crucial element that influences computational requirements. To facilitate the implementation of this algorithm and minimize computational load, a smaller dataset (with fewer cells) can be employed. This approach helps in reducing the computational burden significantly. For the purposes of our study, we will only use one dataset (Adam) as opposed to the full 14 datasets in the original paper.

## Whether you will use the existing code or not:

Yes, we will use the existing code available from the github repository and presented in the paper. This code will start as our baseline and we will adjust accordingly.