Predicting Drug Effects on Single Cell State

Zeinab Navidi

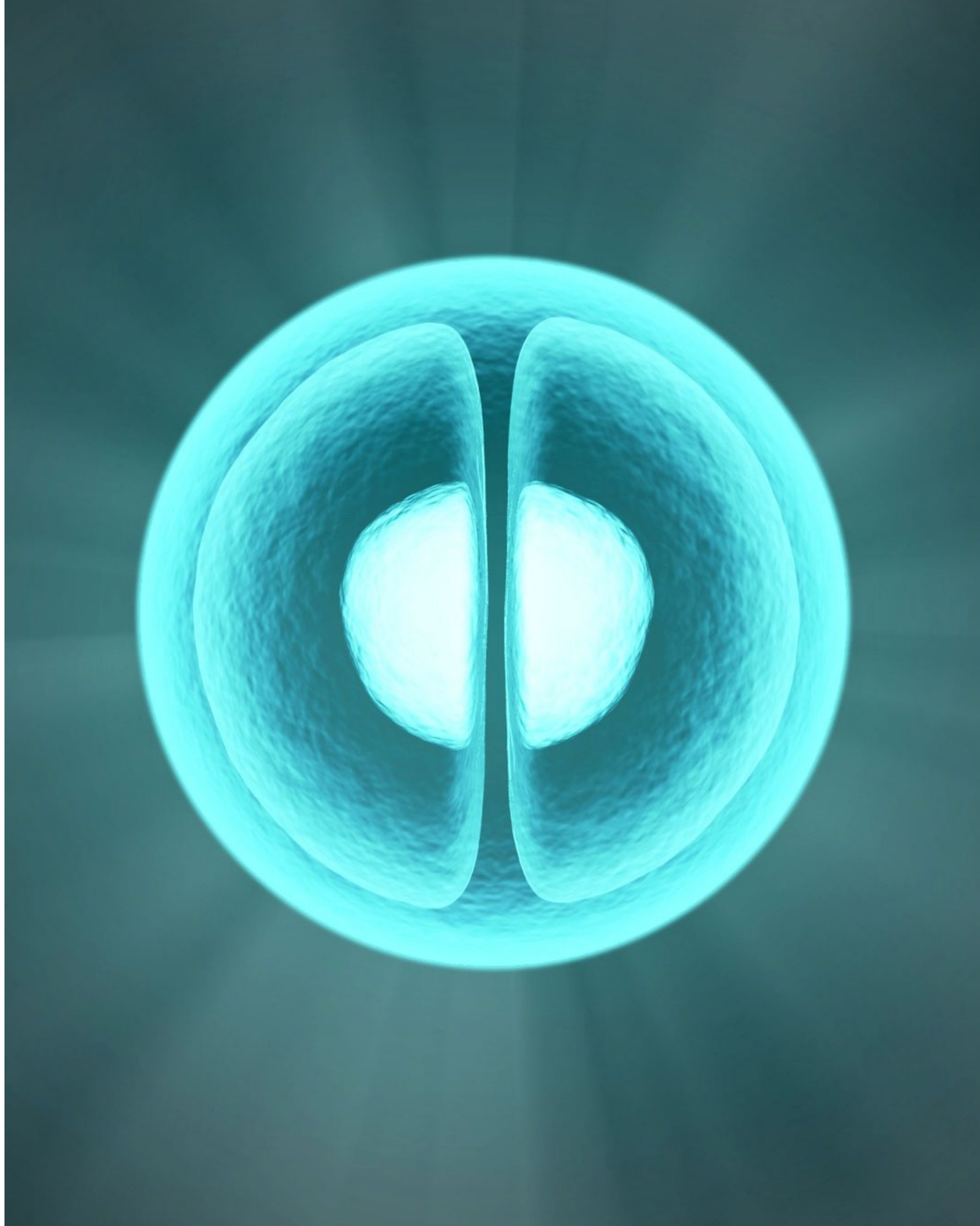
CSC2431 Project Proposal

University of Toronto

Supervised: Dr. Bo Wang,

Co-supervised: Dr. Benjamin Haibe-kains

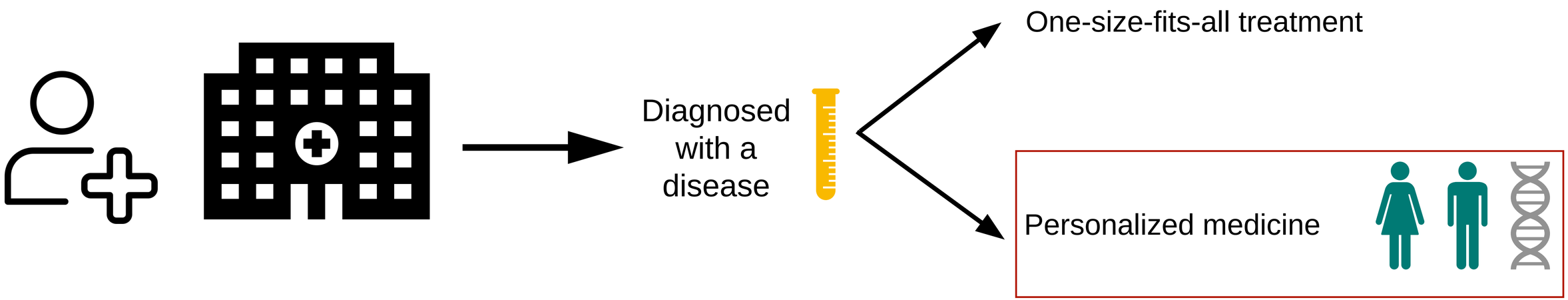
October 12, 2022



Overview

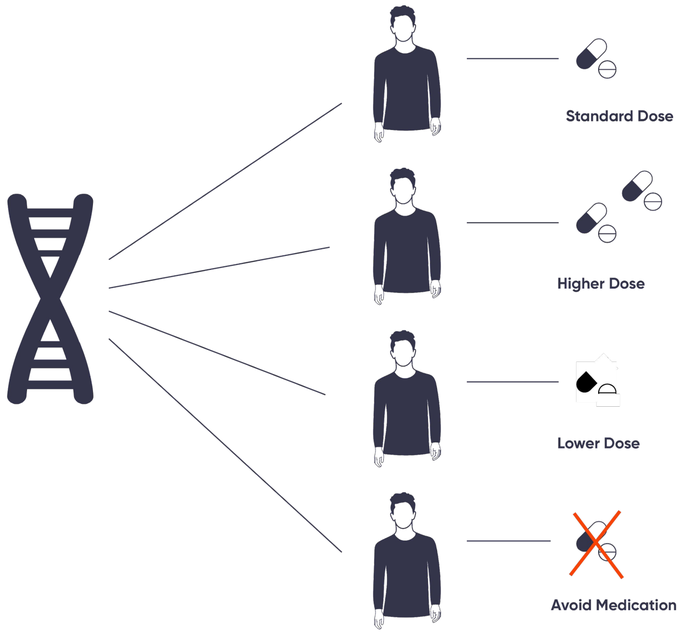
- Motivation
- Introduction to single-cell sequencing data
- Project proposal:
  - Optimal dimensionality reduction of single-cell data
  - Robust drug response prediction
- Who can join this project? Learning opportunities!

Motivation


Diagnosed with a disease
One-size-fits-all treatment
Personalized medicine

Motivation

- How to find the most effective treatment for a patient?
  - Apply many possible treatment on the patient sample to select the best approach
- Challenge:
  - Testing many different treatments are impractical



- **There is a need for computational tools for predicting cell's response to different treatments**

1. Patient pre-treatment genetic information

2. Treatment information

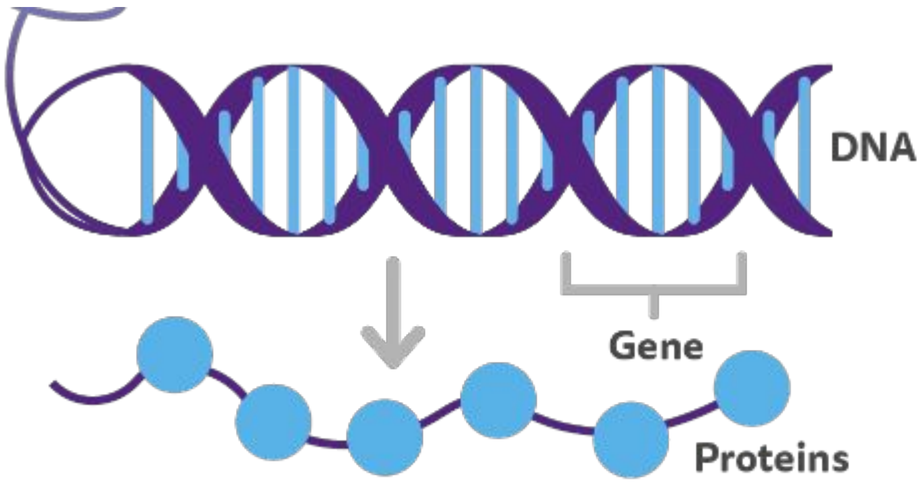
→ AI Model → Effect of input treatment on the patient sample → Shortlist the most effective treatments to be tested in clinic

Introduction to single-cell sequencing data

- **DNA** (or deoxyribonucleic acid) is a long molecule that contains our unique genetic code. Like a recipe book it holds the instructions for making all the **proteins** in our bodies.

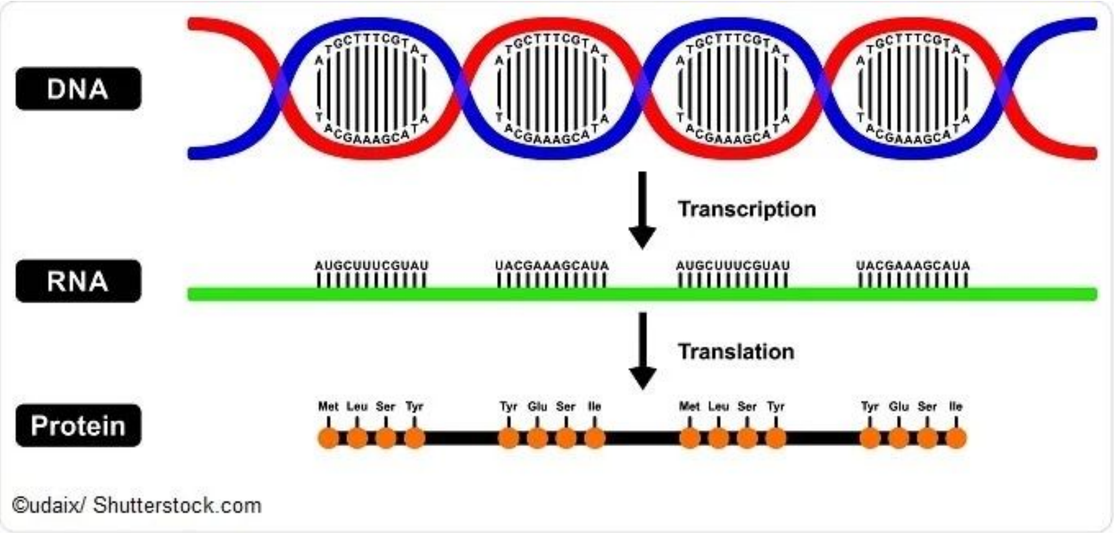
- **Proteins** are the functional units of cells!

- How proteins are created? **genes** are segments of the DNA that encodes the information for making these proteins!



Introduction to single-cell sequencing data

- **Genes** are transcribed into **RNA**
- **RNA** (intermediate message) are translated into **proteins**

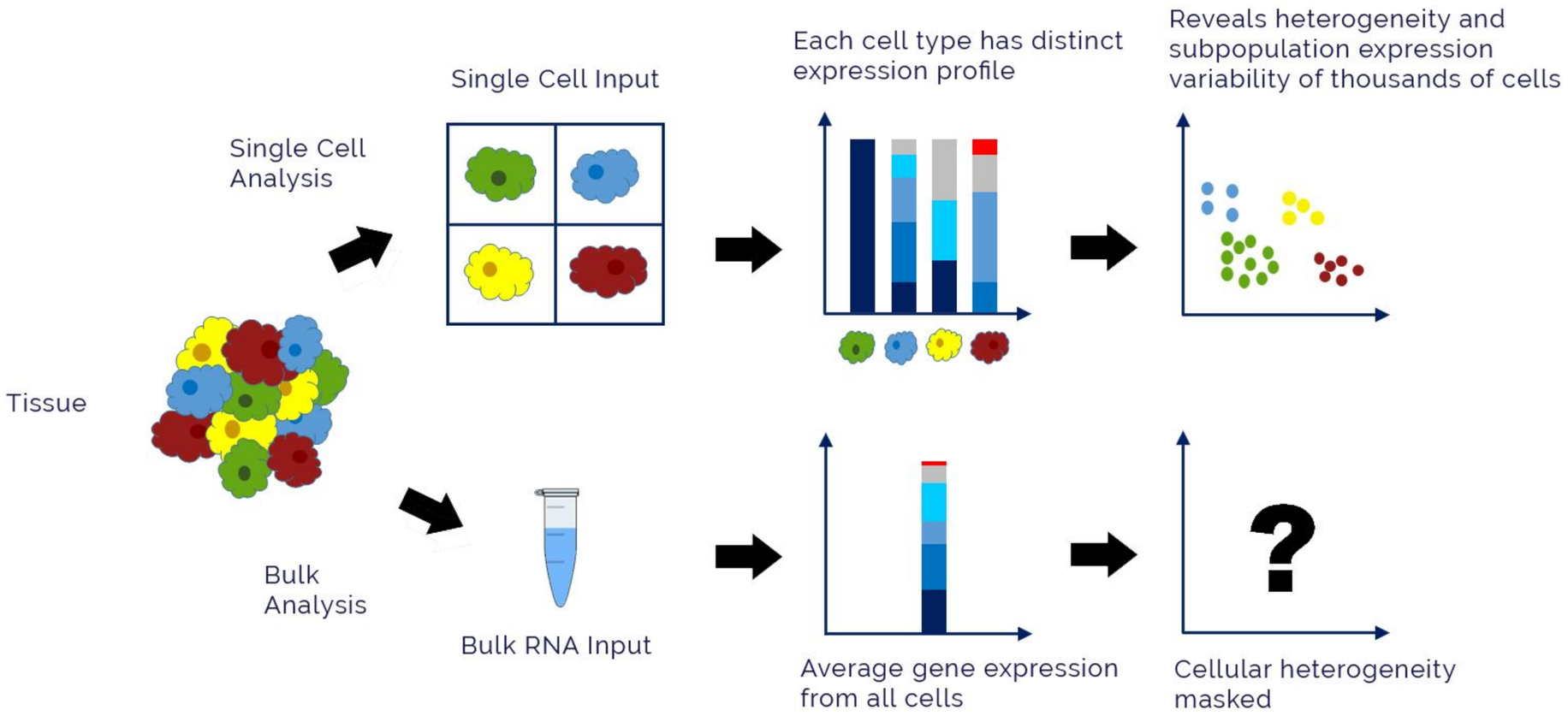


- We focus on RNA data, which provide information about **gene expression**

- **Gene expression** controls how much RNA is made

Introduction to single-cell sequencing data

RNA-sequencing technology measures the number of RNA for each gene (gene expression profile)

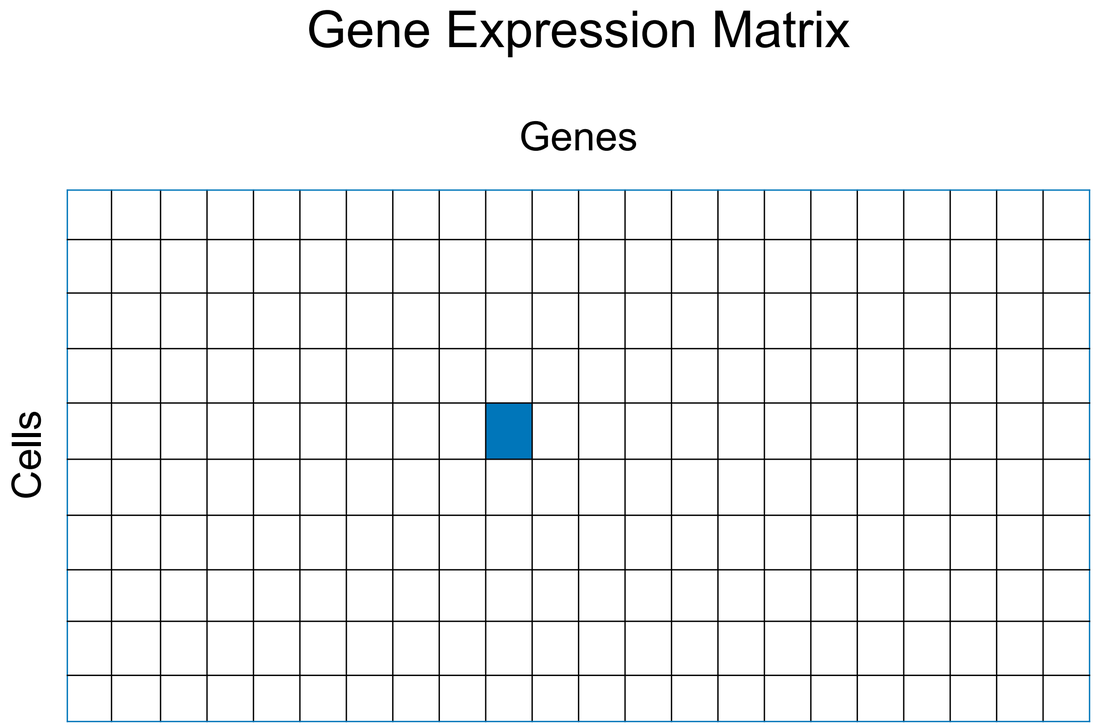


Introduction to single-cell sequencing data

Data:

- We will work with single-cell gene expression data for pre-treatment and post-treatment samples

- Each element of this matrix indicate the number of RNA sequenced for a gene in each individual cell



Project Proposal

Problem: Developing a novel Machine Learning model for predicting a cell's post-treatment gene expression from pre-treatment gene expression

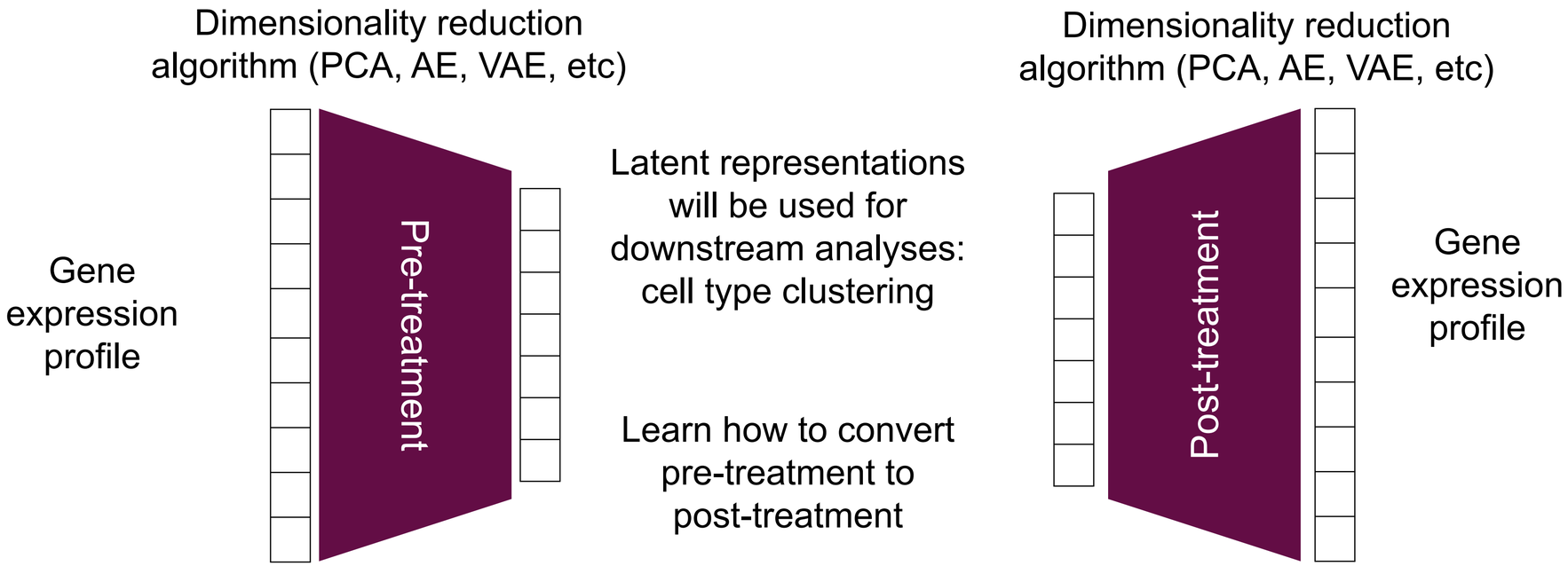
Two interesting proposed tasks:

1. Optimal dimensionality reduction of post-treatment gene expression profile
2. Effective post-treatment gene expression prediction from pre-treatment data

These analyses will give insight about what factors help training more robust and effective machine learning models for post-treatment cell state prediction

Project Proposal

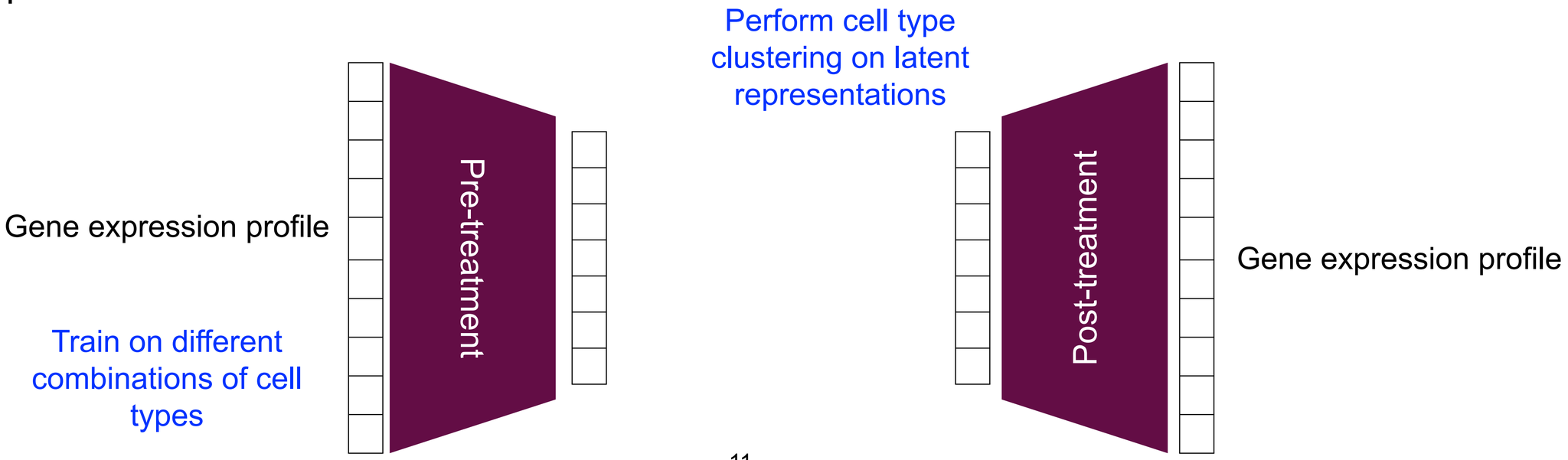
- Dimensionality reduction is a common task in high dimensional single-cell gene expression analysis
- **First step in both tasks:** Deploy an existing tool OR implement a Machine Learning model for post-treatment gene expression prediction from pre-treatment expression profile



Project Proposal

Task 1: Optimal dimensionality reduction of high dimensional gene expression profiles

- **Objective**: To investigate the effect of cell type diversity and proportion in training samples on maintaining cell-type-specific characteristics in post-treatment lower dimension representation
- **Proposed approach**: Train ML model (Variational Autoencoder) on data with different cell type combinations and compare their cell type clustering performance on post-treatment latent representation



Project Proposal

Task 1: Optimal dimensionality reduction of high dimensional gene expression profiles

- **Data:**
  - Kang et al.
  - Includes two groups of pre-treatment and post-treatment cells
  - 8 cell types
  - 18,868 cells, 6,998 genes
  - 23M

Project Proposal

Task 2: Effective post-treatment gene expression profile prediction from pre-treatment data

- **Objective**: To investigate how the effectiveness of drugs would affect learning drug response prediction
- A metric was recently introduced for each drug, indicating how strong the effect of each drug is (Peidli et al.)
- **Proposed approach**: Train ML model (Variational Autoencoder) on data treated with different drugs and effectiveness, and compare the performance on post-treatment expression prediction

Project Proposal

Task 2: Effective post-treatment gene expression profile prediction from pre-treatment data

- **Data:**
  - Srivatsan et al.
  - Includes two groups of pre-treatment and post-treatment cells
  - Treated with 4 different drugs
  - 14811 cells, 58347 genes
  - 173.7M

Resources Estimation

- One GPU
- 8 CPUs
- 32G memory

Who can join this project?

- No biology background is required!
- Experience with Python (preferably PyTorch) is super helpful!
- Understanding Variational Autoencoder is helpful!
- Any motivated student interested in the drug response prediction can join!

Learning opportunities

- Learn about single-cell data and its features
- Get hands-on experience with Pytorch and using Python packages for running existing tools
- Learn about dimensionality reduction algorithms, generative model (Variational Autoencoder), clustering algorithms
- Learn and practice teamwork and collaboration!

Thanks for listening!

Any questions?