

LATENT SPACE INFERENCE FOR SPATIAL TRANSCRIPTOMICS

Sayem Nazmuz Zaman (68943182), David Wang (99030033), and Daniel Chen (43419712)

FINAL REPORT

for

2255 - Spatial Transcriptomics

Capstone Project for ENPH 479 Course at University of British Columbia

2023 April

Sponsor of Capstone Project

Professor Jiarui Ding, Assistant Professor at UBC Computer Science Department

Supervisors of Capstone Project

Dylan Gunn, Project Lab Director

Miti Isbasescu, Head of Software Systems

Bernhard Zender, Student Support

## EXECUTIVE SUMMARY

In order to understand the complexities of cellular biology, researchers are interested in two important metrics: the genetic expression information of cells and their spatial coordinates within a tissue sample. However, state-of-the art methods, namely single-cell RNA sequencing and image based spatial transcriptomics can only recover a subset of this information, either full genetic expression with loss of spatial information, or spatial information with loss of resolution in sequencing data.

In this project, we investigate a probabilistic machine learning method to obtain the full genetic expression information for tissues samples while also preserving their spatial coordinates. This is done through mapping both datasets to a joint latent space representation with the use of variational machine learning methods. From here, the full genetic and spatial information can be decoded and to give us greater insights on the understanding of cellular processes and pathways.

All results and code for this project is documented at our github repository:

<https://github.com/2255-Spatial-Transcriptomics>

# TABLE OF CONTENTS

EXECUTIVE SUMMARY . . . . .	ii
LIST OF TABLES . . . . .	v
LIST OF FIGURES . . . . .	vi
SECTION 1 : COMMONLY USED TERMS . . . . .	1
SECTION 2 : INTRODUCTION . . . . .	2
2.1 Information about the Sponsor . . . . .	2
2.2 Background and Significance of the Project . . . . .	2
2.3 Project Objectives . . . . .	3
2.4 Scope and Limitations . . . . .	3
SECTION 3 : DISCUSSION . . . . .	5
3.1 Approach and System Overview . . . . .	5
3.2 Data Collection . . . . .	5
3.3 Selecting Machine Learning Frameworks: VAE and VGAE . . . . .	9
3.4 Theory and Model Functionality . . . . .	13
3.5 Inference . . . . .	20
SECTION 4 : IMPLEMENTATION . . . . .	22
SECTION 5 : RESULTS . . . . .	24
5.1 Benchmarking Results . . . . .	24
5.2 Discriminator Implementation . . . . .	26
5.3 Results for SEDR . . . . .	27
SECTION 6 : Conclusions . . . . .	30
SECTION 7 : Recommendations . . . . .	31

7.1	Perform more Quantitative and Qualitative Evaluations . . . . .	31
7.2	Hyperbolic and Hyperspherical Latent Prior and Batch Correction . . . . .	31
7.3	Pipeline Validation . . . . .	31
SECTION 8 : Deliverables . . . . .		32
APPENDICES . . . . .		33
BIBLIOGRAPHY . . . . .		46

## LIST OF TABLES

TABLE 1.1	Commonly Used Terms (Also See Table 3.1) . . . . .	1
TABLE 3.1	Commonly Used Terms for Project Pipeline . . . . .	14
TABLE 5.1	scPhere kNN accuracy using 4-fold cross-validation. . . . .	24
TABLE 5.2	scDHA kNN 4-fold cross-validation accuracy with ARI score . . . . .	24
TABLE 5.3	VASC kNN accuracy using 4-fold cross-validation. . . . .	25
TABLE 5.4	scVI kNN accuracy using 4-fold cross-validation. . . . .	25
TABLE A.1	scPhere Results for 2D, 10D, 20D for Dataset A . . . . .	36
TABLE A.2	scPhere Results for 2D, 10D, 20D for Dataset B . . . . .	37
TABLE A.3	scPhere Results for 2D, 10D, 20D for Dataset C . . . . .	38
TABLE A.4	scDHA Results for 15D for All Datasets . . . . .	39
TABLE A.5	VASC Results for 2D, 10D, 20D for Dataset A . . . . .	43
TABLE A.6	VASC Results for 2D, 10D, 20D for Dataset B . . . . .	44
TABLE A.7	VASC Results for 2D, 10D, 20D for Dataset C . . . . .	45

## LIST OF FIGURES

FIGURE 3.1	Data collection pipeline for scRNA sequencing . . . . .	6
FIGURE 3.2	Data collection pipeline for spatial transcriptomics . . . . .	7
FIGURE 3.3	Simple Architecture of a Variational Autoencoder . . . . .	9
FIGURE 3.4	The red circle indicates the VAE portion of the pipeline, which is the purpose of this benchmarking process. . . . .	10
FIGURE 3.5	The blue circle indicates the VGAE portion of the pipeline, which inherits the architecture from SEDR [Chen et al., 2021]. . . . .	13
FIGURE 3.6	<b>Step 1</b> - Obtaining ideal latent space representation for 2000 Highly Variable genes of scRNA data . . . . .	15
FIGURE 3.7	<b>Step 2.1</b> - Obtaining ideal latent space representation for 500 Highly Variable genes of scRNA data (shared with Spatial Transcriptomics) by reducing loss between $\bar{Z}_{sc, exp}^{2000}$ and $Z_{sc, exp}^{500}$ . . . . .	16
FIGURE 3.8	<b>Step 2.2</b> - Obtaining ideal latent space representation for 500 Highly Variable genes of ST gene expression data (shared with scRNA data) by reducing adversarial loss between $\bar{Z}_{sc, exp}^{500}$ and $\bar{Z}_{sc, exp}^{500}$ . . . . .	18
FIGURE 3.9	<b>Step 3</b> - Obtaining ideal latent space representation for full ST data . . .	19
FIGURE 3.10	<b>Inference</b> - We can infer the spatial information of genes from scRNA dataset which do not have spatial information . . . . .	21
FIGURE 5.1	Top three VAE Comparisons . . . . .	26
FIGURE 5.2	Results of discriminator prediction results after training for 10 epochs. The dataset consists of the latent representations of a single cell RNA sequencing dataset (collected from Allen Merfish) and a spatial transcriptomics dataset (collected from 10x Genomics) for a human brain tissue. We note that the discriminator is able to successfully distinguish between the two latent spaces with high accuracy. When projecting the first two principal components of both datasets, we see that there is significant overlap between the latent distributions. . . . .	27

FIGURE 5.3	Comparing the original SEDR framework (top) [Chen et al., 2021] to our implementation (bottom). After concatenation of the latent spaces of spatial and expression data, we then merge the two into a single latent space through the use of a fully connected layer. . . . .	28
FIGURE 5.4	Comparing the results of SEDR_v1 (left) versus SEDR_v2 (right) results for a breast cancer dataset. Both models were trained for 500 epochs with DEC loss. Results obtained from original model achieved a loss of 212.9 while our modified model achieves a loss of 222.6 . . . . .	29
FIGURE 5.5	Comparison between the mapping of cells from the latent space of our model (right) as compared to the original tissue (left). We see that the combined latent space of spatial and expression data provides a good representation of the ground truth cell groupings in the tissue. . . . .	29
FIGURE A.1	Dataset A . . . . .	35
FIGURE A.2	Dataset B . . . . .	35
FIGURE A.3	Dataset C . . . . .	35
FIGURE A.4	Confusion Matrix for Max kNN ( $k = 13$ ) . . . . .	40
FIGURE A.5	Confusion Matrix for Max kNN ( $k = 9$ ) . . . . .	41
FIGURE A.6	Confusion Matrix for kNN at max accuracy ( $k = 16$ ) . . . . .	42

## SECTION 1

### COMMONLY USED TERMS

Term	Description
VAE	Variational Autoencoder
VGAE	Variational Graph Autoencoder
sc-RNA	Single-cell RNA sequencing (method)
ibST	Image-based Spatial Transcriptomics (method)
Single-cell RNA-sequencing dataset (scRNA)	$X_{sc}$
Image-based spatial transcriptomics dataset (ST)	$X_{st}$
Latent Representation of $X_{sc}$	$Z_{sc}$
Latent Representation of $X_{st}$	$Z_{st}$
Reconstructed data from $X$ using VAE/VGAE	$\tilde{X}$
Highly Variable Genes	HVG

Table 1.1: Commonly Used Terms (Also See Table 3.1)



## SECTION 2

### INTRODUCTION

#### 2.1. Information about the Sponsor

Professor Jiarui Ding is an assistant professor at the Computer Science Department at the University of British Columbia. He was a postdoctoral associate at the Broad Institute of MIT and Harvard. His research interest mainly lies in the interception between probabilistic machine learning, visualization, and bioinformatics. He is experienced in various machine learning theories and approaches that could be applied to genetic sequencing data. In particular, he aims to understand how variational autoencoders can bridge the advantages between different RNA sequencing methods, which generates a novel approach to analyzing commercially available gene expression data.

#### 2.2. Background and Significance of the Project

A better understanding of cell and tissue biology can provide powerful insights into treating diseases and cancer. State-of-the-art research focuses on two main methods to dive deep into the genomics information contained within cells and tissues. Single-cell RNA sequencing (scRNA-seq) technologies measure the mRNA expression information at a single-cellular resolution. However, this technology involves separating cells from the tissue into individual cells for measurement, which results in the loss of spatial information about the cells' location within the tissue microenvironment. Retaining this spatial information can provide powerful insights into tissue architecture and cellular pathways. Image-based spatial transcriptomics can preserve the spatial information of sequenced genes. However, this method only allows the profiling of a smaller set of predefined genes (e.g. 500 compared to all 20,000 human genes) and at coarser resolutions.

Our project seeks to investigate a methodology for retaining both sets of information (full gene expression and spatial coordinates) when presented with only gene expression information by imputing the missing information through a probabilistic machine learning framework. This pipeline can provide a more comprehensive understanding of inter-cellular interactions utilizing gene expression and spatial information.

Note that more advanced methods, such as MERFISH, can preserve spatial information while

sequencing at an exceptionally high resolution. However, one major limitation of these methods is their relatively high cost, making them difficult for large-scale studies. Our project aims to maximize the use of existing data collection methods and provide an alternative analysis pipeline to those more advanced methods.

### **2.3. Project Objectives**

Our research aims to create a probabilistic machine learning model to associate gene expression data with spatial information. This would enable us to estimate the spatial position of cells based on the genetic information of RNA molecules in each cell. We will use two datasets - a single-cell RNA-sequencing dataset ( $X_{sc}$ ) and an image-based spatial transcriptomics dataset ( $X_{st}$ ) to train this statistical model and generate a shared latent space. This will facilitate the 'decoding' of the latent code of a cell from image-based spatial transcriptomics and allow us to obtain a comprehensive transcriptomic-wide measurement or fill in any unmeasured genes in image-based spatial transcriptomics.

### **2.4. Scope and Limitations**

#### **2.4.1. Scope**

The input of our project will be the scRNA sequencing data and the spatial transcriptomics data. The output of our project will be the inferred spatial location for a specific cell.

In this project, we use existing VAE and VGAE algorithms that are readily available in the literature and have a working codebase. We slightly modified those algorithms to make them fit into our pipeline, but we are not reinventing the wheel. We also use existing data collected by research institutes such as 10x Genomics and Allen Brain Institute, which are processed cell-gene or spot-gene matrices instead of raw FASTQ files.

#### **2.4.2. Limitations**

##### **Morphological differences between tissues of different patients**

The primary limitation of the project is that the two datasets ( $X_{sc}$  and  $X_{st}$ ) come from the same type of tissue type but might be from different patients. Although our model might infer spatial information of genes, there might be morphological differences between the tissues of different

patients, so the spatial information inferred might need to be more accurate.

### **Our Model does not generalize to unknown cell types**

If there are cell types which our model does not observe, then our model might fail when inferring the spatial location of these cell types. This might be very common in certain types of cancers where tumour populations are unique to a patient and might not have been observed before.

### **VAE assume a multidimensional normal prior for low-dimensional latent variables**

Our VAE model, scVI, maps high-dimensional data into low-dimensional Cartesian latent space, which causes points to be "squeezed" closer to one another, resulting in an inaccurate latent representation of the data. Suppose a hyperbolic latent space distribution was used with uniform prior distributions on a hypersphere, or a hyperbolic latent space representation was used with a wrapped normal distribution in hyperbolic space as the prior. In that case, the latent representation might have been more representative based on the results of Ding and Regev [2021].

### **VAE uses cosine distances to approximate proximity of cells when embedding data on a hypersphere**

Based on the results of Ding and Regev [2021], a hypersphere latent space representation seems to represent scRNA data better. Our VAE, scVI, uses cosine distance to measure the proximity between two cells when embedding in a lower dimensional space and embedding data on a hypersphere in Euclidean space distorts dimensional reduction.

### **VAE does not use batch correction**

In addition to this, single-cell profiles are impacted highly by diverse biological factors such as age, sex, diseases and batch effects in different experiments with different lab protocols. These batch corrections are handled outside of our VAE, scVI.

## SECTION 3

### DISCUSSION

#### 3.1. Approach and System Overview

The project aims to develop a probabilistic machine learning model to map the gene expression data to spatial information. However, associating high-dimensional gene expression data with another high-dimensional spatial transcriptomics data is significantly challenging. It is a common practice to reduce the dataset to its lower dimensional latent representation for further analysis. Variational autoencoders (VAE) have been proven effective in obtaining the optimal latent space for high-dimensional RNA sequencing data. On the other hand, we can use Variational graph autoencoders (VGAE) to obtain the latent spaces for spatial transcriptomics data [Chen et al., 2021]. We will develop a probabilistic machine learning pipeline, called the "latent mapping pipeline," that consists of VAEs and VGAEs to associate those latent spaces with their original data. The latent mapping pipeline consists of three steps, and their simplified descriptions are as follows:

1. Obtain optimal latent space representation for scRNA dataset (Section 3.4.2)
2. Map the gene expressions in the scRNA dataset to the gene expressions in the spatial transcriptomics dataset (Section 3.4.2)
3. Map the gene expressions to the spatial information in the spatial transcriptomics dataset (Section 3.4.2)

The in-depth explanation of the three steps are explained in the Project Pipeline section. Once this pipeline is trained on datasets containing specific cell types, we will be able to achieve our project goal: to impute the spatial information of a specific gene from the scRNA dataset, or vice versa.

#### 3.2. Data Collection

We have two main techniques to study gene expression patterns in cells: Single Cell RNA Sequencing and Spatial Transcriptomics. These two methods rely on RNA sequencing to capture and analyze gene expression data. Both techniques have become increasingly popular in the field of transcriptomics and are often used together to provide a more comprehensive understanding of gene

expression in tissues. However, there are respective drawbacks for each method due to technical limitations.

### 3.2.1. Single Cell RNA Sequencing

Single-cell RNA sequencing (scRNA-seq) is a technique that enables researchers to study gene expression in individual cells. The procedure involves isolating single cells from a sample and capturing the RNA molecules in each cell. The RNA is then converted into DNA sequenced using a high-throughput method shown in Figure 3.1. We can analyze the resulting data and identify which genes are expressed in each cell to gain insights into how different cells function and interact.

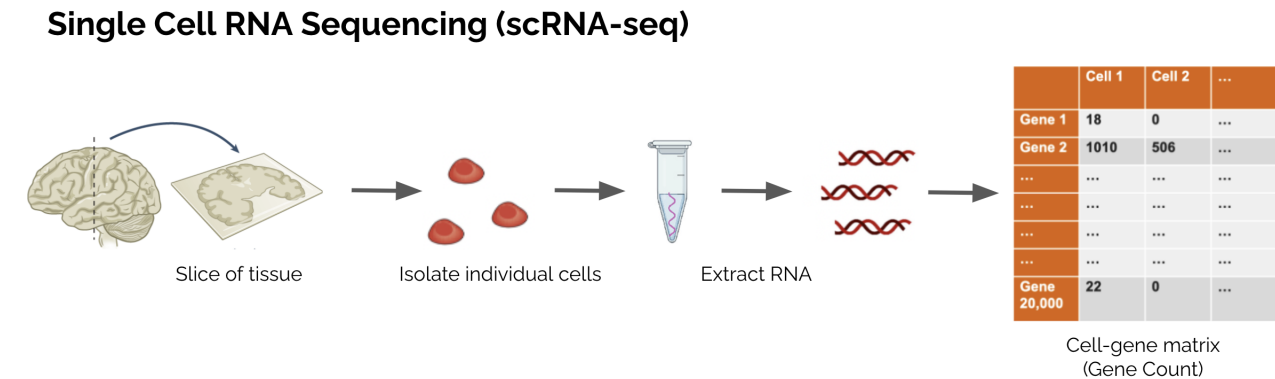


Figure 3.1: Data collection pipeline for scRNA sequencing

We use public datasets from companies and research institutions, such as 10x Genomics and Allen Institute for Brain Science. These datasets are already processed and are in the cell-gene 2D matrix form that contains the gene count for a specific gene in a specific cell type. Typically, the rows represent the gene types, and the columns represent the individual cells sequenced.

### Data Processing

It is worth noting that the data is often subject to different sources of noise that can affect its reliability and usability. One noise source is biological variability, which arises from differences in gene expression levels between individual cells or tissues due to genetic, epigenetic, or environmental factors. Another noise source is technical variability, which arises from errors introduced during the experimental and sequencing steps, such as PCR amplification bias, sequencing errors, or batch effects. Moreover, housekeeping genes are highly expressed across many different cell types, but

their expression levels may not necessarily reflect the cells' actual biological functions or traits. Therefore, we must ensure the data is reliable and perform preliminary data processing to remove housekeeping genes. The detailed data processing procedure is described in Data Processing section.

The processed cell-gene matrix is denoted as  $\mathbf{X}_{sc}$ , which will be the input datasets for our pipeline.

### Drawbacks for scRNA Sequencing Data

The main drawback of scRNA sequencing data is that the spatial information of the genes would be lost while isolating individual cells from a tissue sample. This procedure could damage some cells and prevent us from knowing where the genes were located prior to the sequencing process.

### 3.2.2. Spatial Transcriptomics

Spatial transcriptomics is a technique that enables researchers to study gene expression in cells within their natural tissue context. The procedure involves collecting RNA molecules from a tissue sample using a unique slide with probes designed to detect specific RNA molecules. The RNA molecules are then converted into cDNA and sequenced, allowing researchers to map the spatial distribution of gene expression patterns in the tissue sample. This technique provides valuable information about how different cells are arranged and interact with each other in their natural environment, which is essential for understanding complex biological systems, cellular pathways and evolutionary trajectories.

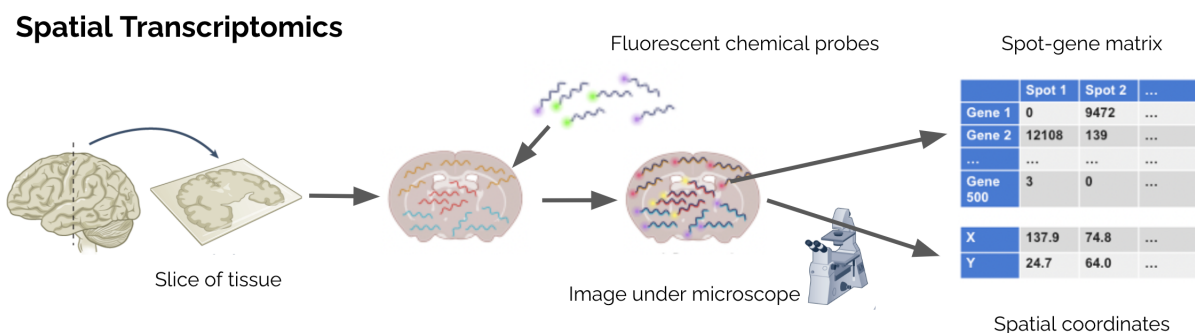


Figure 3.2: Data collection pipeline for spatial transcriptomics

We also gathered publicly-released spatial transcriptomics datasets that contain similar cell types as  $\mathbf{X}_{sc}$ . The dataset consists of two components: the spots' spatial location ( $x, y$  coordinates) and the spot-gene matrix. The latter is a 2D matrix representing the gene expression levels for each spot on

a tissue slice. "Spot" refers to a defined area on the slice where RNA molecules have been captured and sequenced. Similar to analyzing scRNA sequencing data, we perform data pre-processing on the raw spatial transcriptomics data described in Data Processing section and obtain  $\mathbf{X}_{st}$ .

### **Drawbacks for Spatial Transcriptomics Data**

Although we can obtain the spatial information of the spots in spatial transcriptomics, we can only profile a small number of genes per spot due to chemical limits (number of chemical probes per assay and finite RNA molecule capturing efficiency for each probe). Generally speaking, we can only capture around 500 genes in spatial transcriptomics.

#### **3.2.3. Challenges in Associating Two Datasets**

Now, we have two data collection methods that produce the datasets we are working with. However, associating high dimensional gene expression data to another high dimensional spatial transcriptomics data is significantly challenging due to a number of reasons. First, the noise or variability in the data can obscure core patterns or features, making it difficult to distinguish between random fluctuations and true biological differences. Secondly, high-dimensional datasets are often noisy, complex and nonlinear, making it challenging to visualize or interpret the data in a meaningful way. Last but not least, they are also confounded by technical factors such as batch effects or differences in experimental conditions. The simplest and most common way to solve these issues is to obtain each high-dimensional dataset's low-dimensional latent representation using VAEs and VGAEs and perform further analysis on them.

### 3.3. Selecting Machine Learning Frameworks: VAE and VGAE

#### 3.3.1. Benchmarking optimal VAEs

##### Overview

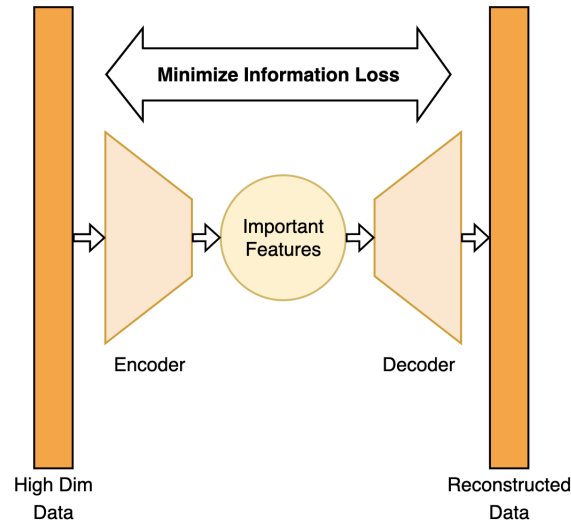


Figure 3.3: Simple Architecture of a Variational Autoencoder

We utilize Variational Autoencoders (VAE) to obtain the optimal latent space that accurately represents the high dimensional sc-RNA data ( $X_{sc}$ ). Traditional VAEs is a neural network that contains an encoder and a decoder. The encoder produces the latent representation of the input data, while the decoder tries to reconstructs the latent representation back to the original data. The encoder and decoder network jointly minimizes the loss between the original data and the reconstructed data, as shown in Figure 3.3. Many VAE architectures have been implemented in literature and to not reinvent the wheel, we focused on benchmarking five VAE methods that were designed for analyzing single-cell RNA sequencing data:

1. scPhere: Deep Generative Model Embedding on Hyperspheres and Hyperbolic Spaces [Ding and Regev, 2021]
2. scDHA: Single-Cell Decomposition using Hierarchical Autoencoder [Tran et al., 2019]
3. VASC: Deep Variational Autoencoder [Wang and Gu, 2018]



4. DRA: Deep Adversarial Variational Autoencoder [Lin et al., 2020]

5. scVI: Deep generative modeling for single-cell transcriptomics [Lopez et al., 2018]

All of these methods remove gene data with an insignificant contribution to the core features that represent the dataset and minimize the loss when reconstructing the dataset from the latent features. To analyze the performance of each method, we extracted the latent representations from the same set of data using these methods. Then, we select the best performing VAE by evaluating the quality of their corresponding latent spaces (see Benchmark Approach and Evaluation Metrics).

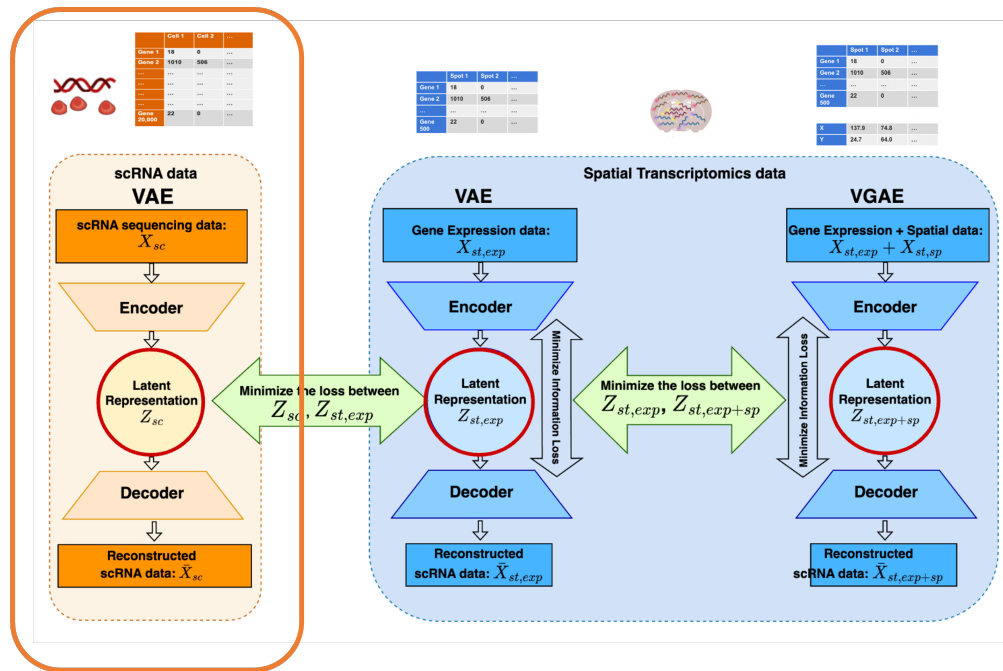


Figure 3.4: The red circle indicates the VAE portion of the pipeline, which is the purpose of this benchmarking process.

### Benchmark Approach and Evaluation Metrics

There are numerous metrics to evaluate a VAE's performance, including reconstruction loss, performance in downstream tasks (generation and classification), and the evidence lower bound (ELBO). We performed **K-nearest neighbours (kNN) with k-fold cross validation** on the latent embeddings, then compared those embeddings of the nearest neighbours to the original input data to evaluate the quality of the latent space. We used a 75/25 split for the training/validation dataset and calculated the final accuracy using the following formula

$$accuracy(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \mathbb{I}^{\hat{y}_i=y_i}, \text{ where } \mathbb{I}^{\hat{y}_i=y_i} \text{ is 1 if } \hat{y}_i = y_i \text{ and 0 otherwise}$$

There are a few key points to consider while using kNN to evaluate the performance of VAEs:

1. **Quality of Input Data:** We need to ensure that the input data is not overly noisy, does not contain an unreasonable amount of outliers, and has a roughly equal distribution of different cell types for kNN evaluation to be accurate. In our analysis, we use datasets provided by the Broad Institute of MIT and Harvard.
2. **Number of  $k$ :** We need an appropriate number of nearest neighbours  $k$  to obtain an accurate evaluation of the model. Generally, the model obtains the highest accuracy at around  $k = \{\text{number of cell types}\}$ . The higher the  $k$ , the smoother the classification boundary, and the lower the misclassification loss, but the more computationally expensive. In our analysis, we vary  $k$  to obtain the latent representation that can most accurately express the initial data distribution.
3. **Choice of distance metric:** We chose to use the **Euclidean distance** for our kNN algorithm for our project.

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

### K-Fold Cross Validation

In our analysis, 4-fold cross-validation was used: each dataset was split into 4 groups, and the model was trained and tested 4 separate times, with each group being the test set during different iterations. This ensures uniformity when training and testing the data as we ensure that our model is not only sensitive to a subset of the data and we are interested in observing how our model performs regardless of the data subset being used to train and test. We then compare the average model performance and accuracy along with the standard deviation across the 4 different folds.

## Datasets

All the datasets that were used were from the Single Cell Portal provided by Broad Institute of MIT and Harvard. We particularly worked with 3 main datasets:

- **Dataset A:** lung\_human\_ASK440

Description: A sample tissue of the upper left lobe of the lung from a male donor aged 58 who was a smoker.

- **Dataset B:** Adipose

Description: Loosely connective tissue that forms body fat.

- **Dataset C:** UC\_Stromal

Description: This dataset consists of stromal cells from a Human umbilical cord.

See Dataset Cell Type Distribution for the cell type distributions of the datasets.

## Source Code Arrangement

The source code for the benchmarking pipeline is published on GitHub. The code base consists of 4 submodules that are forked from their respective repositories.

### 3.3.2. Selecting optimal VGAE

Similar to variational autoencoders, variational graph autoencoders are able to obtain the latent representation of a given high dimensional dataset. However, the encoder of the VGAE consists of graph convolutional networks (GCNs) that is able to encode and decode graph-structured data, which is precisely useful for our spatial transcriptomics datasets. It takes in a adjacency matrix and a feature matrix, in our case, the spot-gene matrix, and outputs the learned latent representation of both input matrices.

The method SEDR (Unsupervised spatially embedded deep representation of spatial transcriptomics) developed a VGAE specifically for analyzing spatial transcriptomics data [Chen et al., 2021]. We decided to make use of their work and incorporate their VGAE architecture into our latent mapping pipeline.

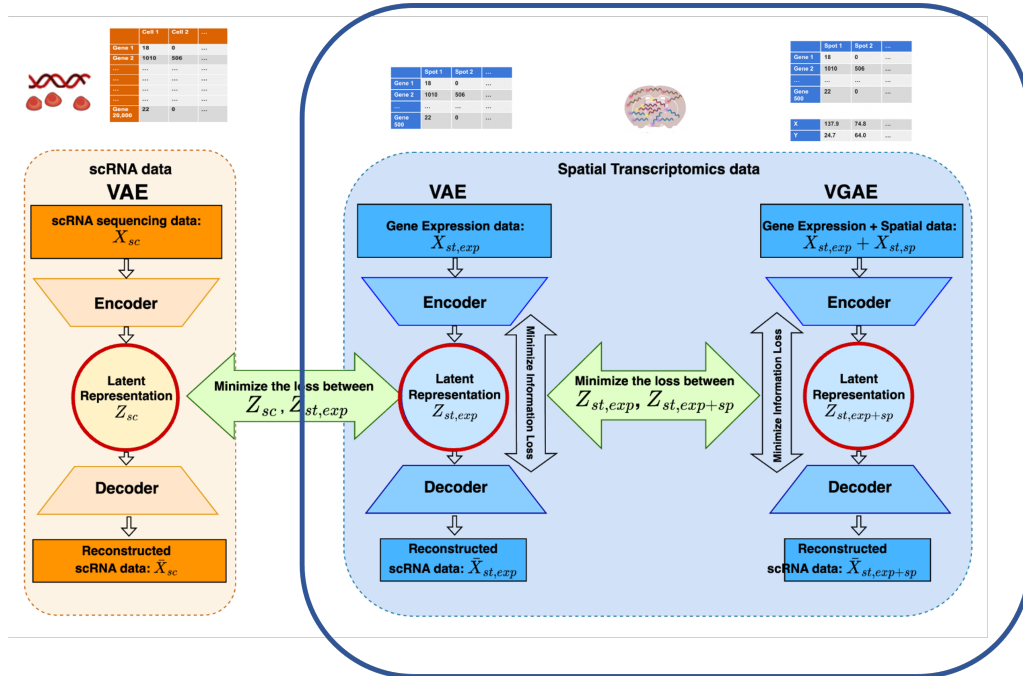


Figure 3.5: The blue circle indicates the VGAE portion of the pipeline, which inherits the architecture from SEDR [Chen et al., 2021].

### 3.4. Theory and Model Functionality

#### 3.4.1. Data Processing

Our first task with both the datasets is to apply some quality control measures. This helps to reduce the size of our datasets. We first remove cells which have less than 200 genes expressed and remove genes which are expressed in less than 60 cells. These numbers were suggested by our capstone supervisor and are an extrapolation of the values used for analyzing a dataset of Peripheral Blood Mononuclear Cells (PBMC)[Satija [2015]].

Next, we remove cells with high mitochondrial and ribosomal genes. High proportion of these cells indicate loss of cytoplasmic RNA from perforated cells.

Since we are given two extremely high-dimensional datasets, we would want to select only the genes which are relevant and have high probability of being mapped. These genes correspond to genes with high variability. The discovery of the highly variable gene (HVG), allows detection of genes that correspond to cell-to-cell variation within a homogeneous cell population. We have the underlying assumption that the highly variable genes are more likely to be mapped between the two datasets.

The top 2000 highly variable genes were selected from the scRNA dataset ( $X_{sc, exp}^{2000}$ ). In addition to this, the top 500 highly variable genes shared between the scRNA dataset and the Spatial Transcriptomics dataset were also processed for later use ( $X_{sc, exp}^{500}, X_{st, exp + sp}^{500}$ ).

Term	Description
$X_{sc, exp}^n$	The cell gene matrix of single-cell RNA sequencing data, with n genes
$X_{st, exp}^n$	The spot gene matrix of spatial transcriptomics data, with n genes
$X_{st, sp}^n$	The spatial locations for spots for spatial transcriptomics data, with n genes
$X_{st, exp+sp}^n$	The gene expression and spatial locations for spots for spatial transcriptomics data, with n genes
$Z_i^n$	The latent space of dataset i, for n genes
$\bar{Z}_i^n$	The finalized, <b>fixed</b> latent space of dataset i, for n genes
$\tilde{X}_i^n$	The reconstructed data for dataset i, with n genes

Table 3.1: Commonly Used Terms for Project Pipeline

### 3.4.2. Project Pipeline

Since we have two datasets, scRNA dataset and Spatial Transcriptomics dataset containing two different types of information (scRNA data contains only gene expression data and Spatial Transcriptomics data contains both gene expression data and spatial information of the cells), it is difficult to map these two datasets together. We solve this problem by mapping the latent spaces generated by VAEs and VGAEs of the two datasets. We do this in three steps.

#### Step 1: Find ideal Latent Space representation of the top 2000 highly variable genes of the scRNA dataset

In step 1, a VAE framework is implemented to generate an ideal, compact latent space representation for the 2000 HVG of the scRNA data. The encoder network of the VAE,  $f_{vae}$ , takes the original data and passes it through a series of neural networks to find a latent space representation:

$$Z_{sc, exp}^{2000} = f_{vae1}(X_{sc, exp}^{2000})$$

The decoder network of the VAE,  $f_{vae}^{-1}$ , takes the latent space representation and tries to reconstruct the original data from the latent space representation:

$$\tilde{X}_{sc, exp}^{2000} = f_{vae}^{-1}(Z_{sc, exp}^{2000})$$

The ideal latent space is found by minimizing both the KL divergence between the latent space obtained and the original data and the reconstruction loss between the original data and the reconstructed data. This ideal latent space representation of the 2000 HVG of the scRNA data is then **fixed**, and we represent this as  $\bar{Z}_{sc, exp}^{2000}$

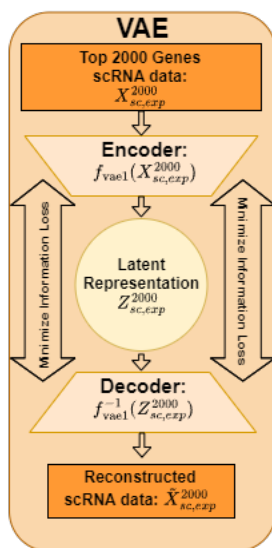


Figure 3.6: **Step 1** - Obtaining ideal latent space representation for 2000 Highly Variable genes of scRNA data

## Step 2: Map the 500 shared genes between the 2000 HVG scRNA data and the Spatial Transcriptomics data

Once we have fixed the latent space representation of the 2000 HVG of the scRNA data,  $\bar{Z}_{sc, exp}^{2000}$ , we try to map the shared 500 genes between the 2000 HVG of the scRNA data and the spatial transcriptomics data. This is done in two sub steps:

1. **Step 2.1:** We want to find the ideal Latent Space representation of the shared 500 genes of the scRNA data. This ideal representation is found by minimizing the KL divergence between the latent space and the original data (shared 500 genes of the scRNA data) and minimizing

the loss between the original data and the reconstructed scRNA data of the shared 500 genes from its latent space representation.

$$Z_{sc, exp}^{500} = f_{vae1}(X_{sc, exp}^{500})$$

$$\tilde{X}_{sc, exp}^{500} = f_{vae1}^{-1}(Z_{sc, exp}^{500})$$

In addition to this, to ensure that the latent space of the 500 genes of the scRNA data,  $Z_{sc, exp}^{500}$ , is **similar** to the fixed latent space representation of the 2000 HVG of the scRNA data,  $\bar{Z}_{sc, exp}^{2000}$ , from **Step 1** we implement a euclidean loss,  $L_1$ :

$$L_1 = \text{dist}(Z_{sc, exp}^{500}, \bar{Z}_{sc, exp}^{2000})$$

Minimising this loss ensures that the two latent spaces are similar to each other. Once we have found an ideal latent space representation for the shared 500 genes for the scRNA data by minimizing KL divergence, reconstruction loss and euclidean loss between the latent space representation of the 500 shared genes and the 2000 HVG of the scRNA dataset, this latent space representation is then **fixed** and is represented by  $\bar{Z}_{sc, exp}^{500}$ . This is shown in Figure 3.7.

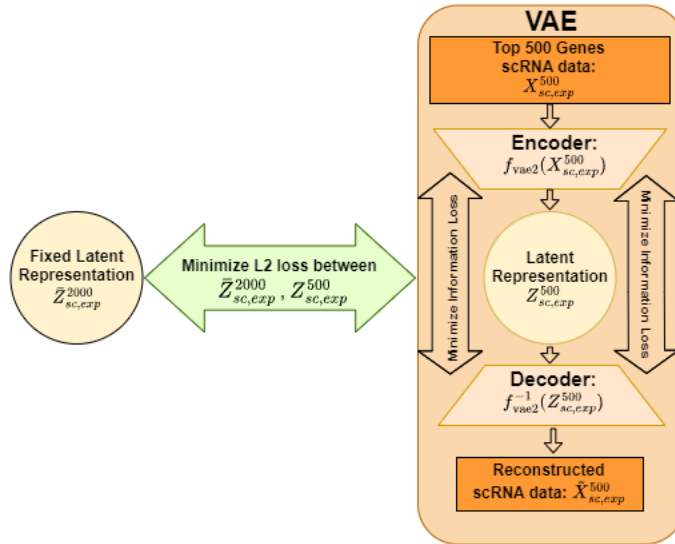


Figure 3.7: **Step 2.1** - Obtaining ideal latent space representation for 500 Highly Variable genes of scRNA data (shared with Spatial Transcriptomics) by reducing loss between  $\bar{Z}_{sc, exp}^{2000}$  and  $Z_{sc, exp}^{500}$

2. **Step 2.2:** We also want to find the ideal Latent Space representation of the shared 500 genes of the ST gene expression data. This ideal representation is found by minimizing the KL divergence between the latent space and the original data (shared 500 genes of the ST gene expression data), and minimizing the loss between the original data and the reconstructed ST gene expression data of the shared 500 genes from its latent space representation. Since the gene expression of the ST data does not contain any spatial information, we can use a VAE for this purpose:

$$\begin{aligned} Z_{\text{st, exp}}^{500} &= f_{\text{vae2}}(X_{\text{st, exp}}^{500}) \\ \tilde{X}_{\text{st, exp}}^{500} &= f_{\text{vae2}}^{-1}(Z_{\text{st, exp}}^{500}) \end{aligned}$$

Now, since the ST gene expression data for 500 genes contain the same **type** (gene-expression) of information as the scRNA shared 500 genes, we can map their respective latent spaces together but in order to do so, we have to minimize an additional loss function. This loss function is adversarial instead of plain euclidean since we are trying make two latent spaces from different datasets (and thus different distributions) similar. Here we introduce a discriminator network  $\mathbf{D}$ , which we will train to differentiate the two latent spaces: the **fixed** Latent space of the 500 shared genes for the scRNA data,  $\bar{Z}_{\text{sc, exp}}^{500}$ , and the latent space of the 500 shared genes for the ST gene expression data,  $Z_{\text{st, exp}}^{500}$ .

This can help in finding the relationships between highly complex distributions. We define the loss function

$$L_2 = \mathbf{D}(\bar{Z}_{\text{sc, exp}}^{500}, Z_{\text{st, exp}}^{500})$$

as the prediction accuracy the network has on differentiating between the labels of the respective latent spaces.

We first train the discriminator to maximize  $L_2$  so that our discriminator can correctly differentiate the two latent spaces and identify their labels correctly. Once the discriminator is trained, we then try to minimize  $L_2$  when training the VAE,  $f_{\text{vae2}}$ , so that we generate a



latent space,  $Z_{st,exp}^{500}$ , which our discriminator fails to differentiate from the **fixed** latent space,  $\bar{Z}_{sc,exp}^{500}$ . Once this loss is minimized, the latent space representation for the 500 shared genes of the ST gene data is **fixed**,  $\bar{Z}_{st,exp}^{500}$

All of this is shown in Figure 3.8.

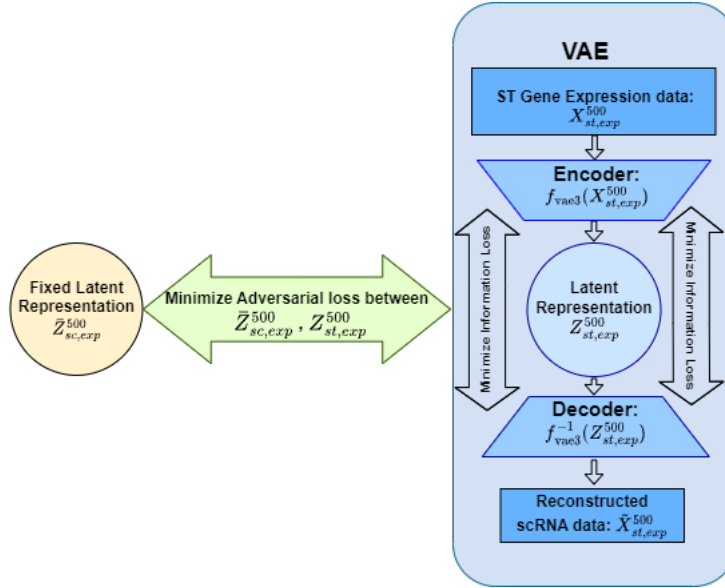


Figure 3.8: **Step 2.2** - Obtaining ideal latent space representation for 500 Highly Variable genes of ST gene expression data (shared with scRNA data) by reducing adversarial loss between  $\bar{Z}_{sc,exp}^{500}$  and  $Z_{st,exp}^{500}$

### Step 3: Map the ST Gene Expression latent space to the a generated latent space

Now that we have the **fixed** latent space representation for the 500 shared genes of the ST gene data is **fixed**,  $\bar{Z}_{st,exp}^{500}$ , we try to map this to the latent space representation of the full ST data containing both the gene expression and spatial information. In order to do so we first need to generate a latent space for the ST data and this latent space must encapsulate both the gene expression and spatial information into one complete latent space.

To do so we introduce a Variational Graph Autoencoder (VGAE), which embeds both expression and spatial information into a graph neural network, and forms a latent space representation for the combined information. The encoder network and decoder network for a VGAE works in the same

way as that for the VAE:

$$Z_{st, \text{exp}+sp}^{500} = f_{\text{vgae}}(X_{st, \text{exp}+sp}^{500})$$

$$\tilde{X}_{st, \text{exp}+sp}^{500} = f_{\text{vgae}}^{-1}(Z_{st, \text{exp}+sp}^{500})$$

Just like the VAEs, the ideal representation is found by minimizing the KL divergence between the latent space and the original data (ST gene expression and spatial information) and minimizing the loss between the original data and the reconstructed ST data.

We also ensure that the latent space of the full ST data (gene expression and spatial information),  $Z_{st, \text{exp}+sp}^{500}$ , is **similar** to the fixed latent space representation of the 500 shared genes of the ST gene expression data obtained from **Step 2.2**,  $\bar{Z}_{st, \text{exp}}^{500}$ . This can be done by implementing another euclidean loss,  $L_3$  since both the latent spaces are from the same dataset thus have similar distributions:

$$L_3 = \text{dist}(Z_{st, \text{exp}+sp}^{500}, \bar{Z}_{st, \text{exp}}^{500})$$

Minimising this loss ensures that the two latent spaces are similar to each other. Step 3 is shown in Figure 3.9

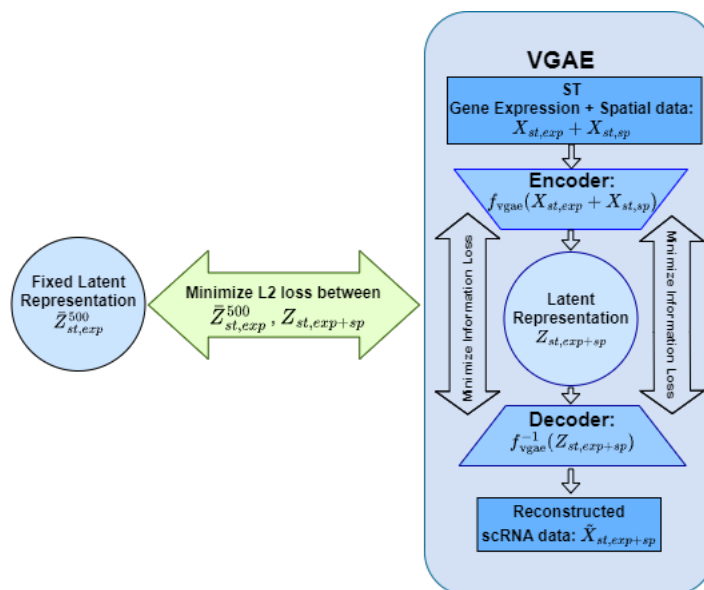


Figure 3.9: **Step 3** - Obtaining ideal latent space representation for full ST data

### 3.5. Inference

Once all our networks are trained, and losses have been minimized, we can perform inference to infer the spatial information of genes from scRNA dataset.

From **Step 2** and **Step 3**, we have mapped various latent spaces to each other by making their distributions similar. We can define these mappings as function, which takes a latent space, with some loss and generates the mapped latent space:

$$\text{Step 2.2: } \bar{Z}_{\text{st, exp}}^{500} \sim \bar{Z}_{\text{sc, exp}}^{500} \Rightarrow \bar{Z}_{\text{st, exp}}^{500} = f_1(\bar{Z}_{\text{sc, exp}}^{500})$$

$$\text{Step 3: } \bar{Z}_{\text{st, exp+sp}} \sim \bar{Z}_{\text{st, exp}}^{500} \Rightarrow \bar{Z}_{\text{st, exp+sp}} = f_2(\bar{Z}_{\text{st, exp}}^{500})$$

Similarly, we have trained our VAE encoder network from **Step 2.1** and our VGAE decoder network from step 3:

$$\text{Step 2.1: } \bar{Z}_{\text{sc, exp}}^{500} = f_{\text{vae2}}(X_{\text{sc, exp}}^{500})$$

$$\text{Step 3: } \tilde{X}_{\text{st, exp+sp}} = f_{\text{vgae}}^{-1}(X_{\text{st, exp+sp}})$$

If we isolate the trained encoder network for the VAE in **Step 2.1**, the mapping functions for **Step 2.2**, **Step 3** and the trained decoder function from **Step 3**, we can build our inference model. This is shown in Figure 3.10.

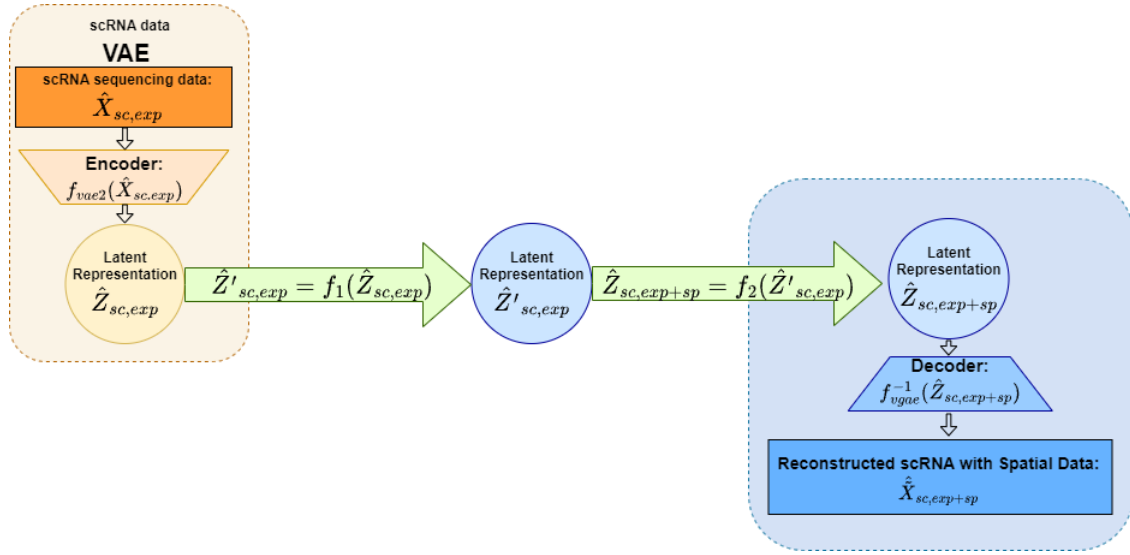


Figure 3.10: **Inference** - We can infer the spatial information of genes from scRNA dataset which do not have spatial information

Following the inference pipeline in Figure 3.10, we get:

$$\hat{X}_{st, exp+sp} = f_{vgae}^{-1}(f_2(f_1(f_{vae2}(X_{sc, exp}))) = f_{model}(X_{sc, exp})$$

Thus, our inference model allows us to infer the spatial information of the gene expression data for the scRNA data.

## SECTION 4

### IMPLEMENTATION

Two popular libraries in machine learning are PyTorch and Tensorflow. After a preliminary analysis, our team decided to utilize PyTorch for the implementation of our models due to its ease of modification and pythonic implementation style. Based on prior experience and sponsor recommendations, we also determined that PyTorch is the better framework to implement custom models and make changes to existing models. Our team researched current implementations which overlap with our proposed architecture, and sought to integrate the existing codebase with our implementation.

For implementation of the Variational Graph Autoencoder, we utilized scVI due to its compatibility with PyTorch. For the implementation of the Variational Graph Autoencoder, we followed the implementation of SEDR. Both these repositories were integrated and we established modules and the appropriate class definitions to integrated these modules into our code. We followed a modular, object-oriented approach to ensure organization and scalability of our framework. All related implementation and tests can be found on our team's github repository.

The pseudo-code for our implementation to train the latent spaces using three steps described in Section 3.4.2 is shown in the next page.

---

**Algorithm 1** The 3 step training procedure

---

**Prepare:**  $X_{sc, exp}^{2000}$ ,  $X_{sc, exp}^{500}$  ▷ expression datasets for SC  
 $X_{st, exp}^{500}$ ,  $X_{st, sp}^{500}$  ▷ expression and spatial datasets for ST

**for** e in s1\_epochs **do**

**Train:** VAE  $f_{vae1}$  s.t. reconstruction loss of  $\tilde{X}_{sc, exp}^{2000} = f_{vae1}^{-1}(f_{vae1}(X_{sc, exp}^{2000}))$  is minimized

**end for**

**Fix:**  $f_{vae1}(X_{sc, exp}^{2000}) = Z_{sc, exp}^{2000} \rightarrow \bar{Z}_{sc, exp}^{2000}$

**for** e in s2\_epochs **do**

$Z_{sc, exp}^{500} = f_{vae2}(X_{sc, exp}^{500})$

$Z_{st, exp}^{500} = f_{vae3}(X_{st, exp}^{500})$

**while**  $\mathbf{D}(Z_{sc, exp}^{500}, Z_{st, exp}^{500}) < \alpha$  for a max of T iterations **do**

**Train:** discriminator  $\mathbf{D}$  to improve accuracy ( $L_2$  of differentiating between latent spaces)

$$L_2 = \mathbf{D}(Z_{sc, exp}^{500}, Z_{st, exp}^{500})$$

**end while**

**for** i in s2b\_epochs **do**

$L_1 = \text{dist}(Z_{sc, exp}^{500}, \bar{Z}_{sc, exp}^{2000})$  ▷ Similarity between SC expression latent spaces

$L_2 = \mathbf{D}(Z_{sc, exp}^{500}, Z_{st, exp}^{500})$  ▷ Similarity between SC and ST latent space

**Train:** VAE  $f_{vae2}$  by minimizing weighted sum of  $L_1, L_2$ ,

and reconstruction loss of  $\tilde{X}_{sc, exp}^{500} = f_{vae2}^{-1}(f_{vae2}(X_{sc, exp}^{500}))$

**Train:** VAE  $f_{vae3}$  by minimizing weighted sum of  $L_2$

and reconstruction loss of  $\tilde{X}_{st, exp}^{500} = f_{vae3}^{-1}(f_{vae3}(X_{st, exp}^{500}))$

**end for**

**end for**

**Fix:**  $f_{vae3}(X_{st, exp}^{500}) = Z_{st, exp}^{500} \rightarrow \bar{Z}_{st, exp}^{500}$

**for** e in s3\_epochs **do**

$Z_{st, exp+sp}^{500} = f_{vgae}(X_{st, exp}^{500}, X_{st, sp}^{500})$

$L_3 = \text{dist}(Z_{st, exp+sp}^{500}, \bar{Z}_{st, exp}^{500})$  ▷ Similarity to ST expression latent space

**Train:** VGAE  $f_{vgae}$  by minimizing weighted sum of  $L_3$

and reconstruction loss of  $\tilde{X}_{st, exp}^{500}, \tilde{X}_{st, sp}^{500} = f_{vgae}^{-1}(f_{vgae}(X_{st, exp}^{500}, X_{st, sp}^{500}))$

**end for**

---

## SECTION 5

### RESULTS

#### 5.1. Benchmarking Results

We present the benchmarking results for each method in the following tables. For scPhere, we observed the latent representations for 2, 10 and 20 dimensions on a hyper-sphere. We assumed a negative-binomial distribution for the observation distribution, and we computed the latent space distribution, which is assumed to be a von Mises-Fisher distribution. Avg CV Acc stores the average 4-fold cross-validation accuracy.

Datasets	scPhere Accuracy					
	2D		10D		20D	
	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc
Dataset A	88.90%	89.86%	89.87%	89.87%	71.18%	76.61%
Dataset B	95.94%	95.07%	94.78%	95.56%	95.36%	95.79%
Dataset C	81.81%	76.33%	86.28%	75.77%	86.46%	76.20%

Table 5.1: scPhere kNN accuracy using 4-fold cross-validation.

For scDHA, the latent dimension  $m$  is predetermined in the library to be 25 if the dataset has more than 50000 genes and 15 otherwise since varying  $m$  between 10 and 20 does not alter the analysis results [Tran et al., 2019, p.8]. All our datasets have less than 50000 genes, thus the following analysis has a latent dimension of 15. For latent space distribution plots, kNN accuracy v.s.  $k$  plots, and confusion matrices, please see scDHA Supplemental Figures.

Datasets	scDHA Accuracy	
	15D	
	Avg CV Acc	ARI
Dataset A	84.17%	0.54
Dataset B	94.2%	0.32
Dataset C	86.7%	0.17

Table 5.2: scDHA kNN 4-fold cross-validation accuracy with ARI score

For VASC, we observed the normal latent representations for 2, 10 and 20 dimensions. We assumed a multi-dimensional normal prior distribution for the observation distribution, and we computed the posterior distribution.

Datasets	VASC Accuracy					
	2D		10D		20D	
	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc
Dataset A	64.41%	70.57%	72.74%	77.37%	69.96%	75.41%
Dataset B	84.06%	87.01%	86.09%	87.88%	86.67%	87.74%
Dataset C	30.64%	37.51%	54.24%	55.61%	54.98%	56.82%

Table 5.3: VASC kNN accuracy using 4-fold cross-validation.

For scVI, we assume a normal distribution for the latent space representation. We observe the latent space representation for 2, 10 and 20D on a normal distribution. We see that scVI performs similar to scPhere for 2D and 10D representations for Datasets A and B. However, scVI’s performance seems to decrease as we increase the number of latent dimensions. This is not problematic for us since we want a low dimensional representation of our data and 10D representation seems ideal for this case. In addition to this, scVI is implemented in PyTorch and thus is compatible with our VGAE models.

Datasets	scVI Accuracy					
	2D		10D		20D	
	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc	KNN Acc	Avg CV Acc
Dataset A	89.51%	89.86%	83.96%	86.12%	25.934%	32.95%
Dataset B	91.88%	94.19%	93.33%	93.105%	44.637%	53.628%
Dataset C	68.74%	67.17%	85.77%	75.256%	70.20%	54.31%

Table 5.4: scVI kNN accuracy using 4-fold cross-validation.

DRA uses the most novel approach, generative adversarial networks, out of the four methods. Similar to other methods, we can obtain the latent matrix  $z$  for all datasets. However, the results for the latent representations were not ideal. We obtained an average 4-fold cross-validation accuracy of 38% for Dataset A and sub-optimal performances for Datasets B and C. Therefore, we did not move forward with this method.

### 5.1.1. VAE Model Selection

The results of our top three performing models are shown in Figure 5.1.



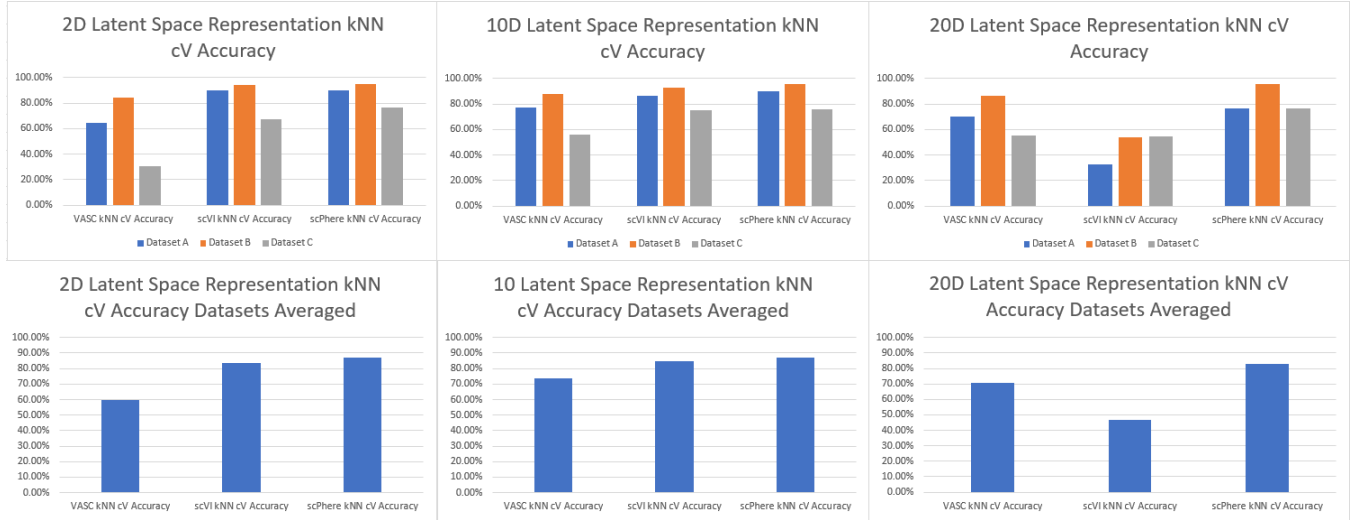


Figure 5.1: Top three VAE Comparisons

We see that overall, scVI tends to perform better compared to VASC for 2D and 10D latent representation but performs significantly worse for 20D latent representation. scSphere performs the best on average for all the dimensions on average. We picked **scVI** for our VAE since it performs about the same as scSphere for 2D and 10D latent space representation and in addition to this, scVI is implemented in PyTorch so it is easier to integrate with our VGAE model which is also implemented in PyTorch.

## 5.2. Discriminator Implementation

For the discriminator, we implement a standard binary classifier in PyTorch from scratch. We implemented a model with 3 hidden layers, each with 128 nodes, RELU activation, Binary Cross Entropy with Logits as the loss function, and the Adam optimizer. The discriminator network was implemented as its own class in our code repository to allow easy modifications to model hyperparameters for future fine-tuning. To verify the performance of our discriminator, on two latent spaces generated from scVI, this time using 10 dimensional latent space features, as shown in figure ???. We visualize the results by plotting the 2 principal components of both datasets, and note that the discriminator learns to distinguish the high dimensional data with great precision. Here the discriminator was trained also for 30 epochs.

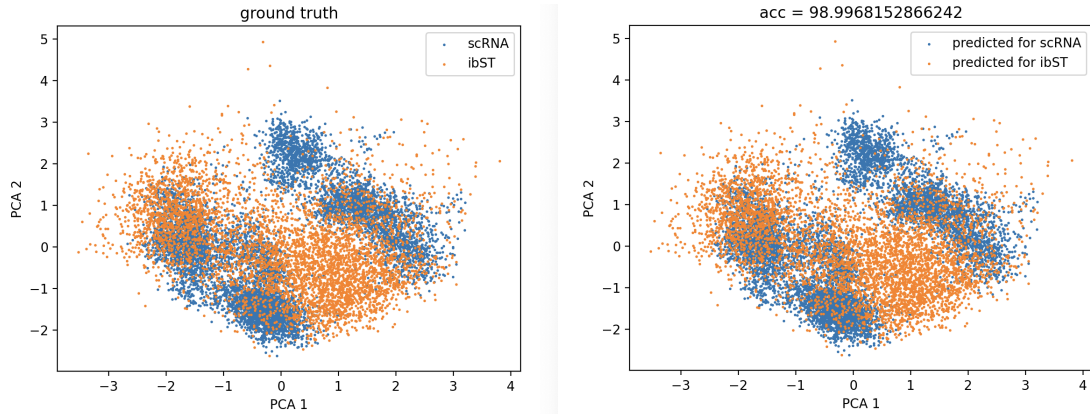


Figure 5.2: Results of discriminator prediction results after training for 10 epochs. The dataset consists of the latent representations of a single cell RNA sequencing dataset (collected from Allen Merfish) and a spatial transcriptomics dataset (collected from 10x Genomics) for a human brain tissue. We note that the discriminator is able to successfully distinguish between the two latent spaces with high accuracy. When projecting the first two principal components of both datasets, we see that there is significant overlap between the latent distributions.

### 5.3. Results for SEDR

The SEDR framework performs latent mapping of spatial transcriptomics datasets and produces a latent representation of both the spatial graph and gene expression. We make modifications on top of this architecture and introduce an additional hidden layer within the network, which combines the latent spaces of the spatial information and gene expression information into a single latent space. We denote the original SEDR as SEDR\_v1 and our modified model as SEDR\_v2.

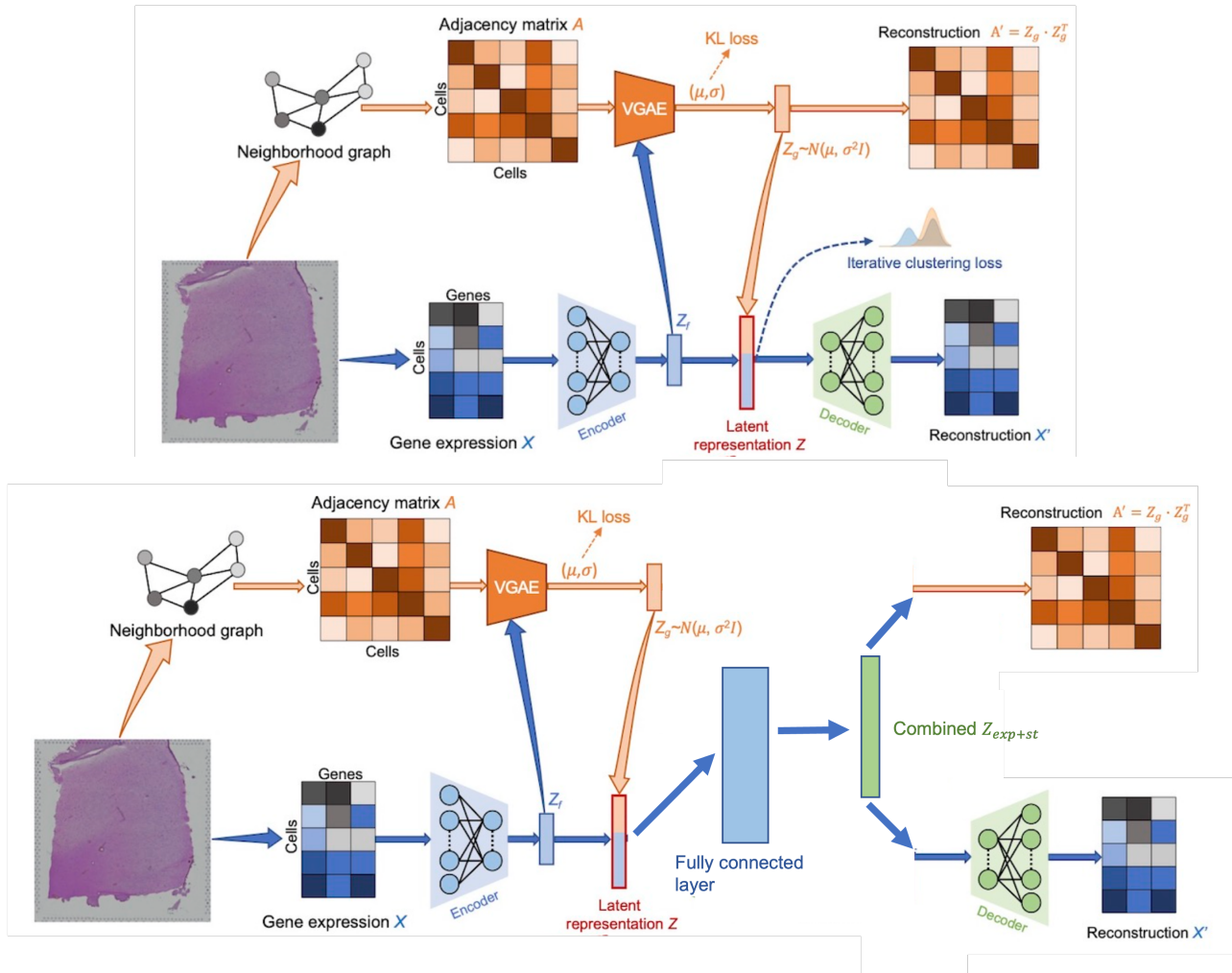


Figure 5.3: Comparing the original SEDR framework (top) [Chen et al., 2021] to our implementation (bottom). After concatenation of the latent spaces of spatial and expression data, we then merge the two into a single latent space through the use of a fully connected layer.

We compare the results of SEDR\_v2 with the original model and conclude that the combined latent representation via the fully connected layer is feasible for retaining the learned information. The results can be seen in figure 5.4. Further tests are performed which confirm the feasibility of this model and can be seen on our github repo.

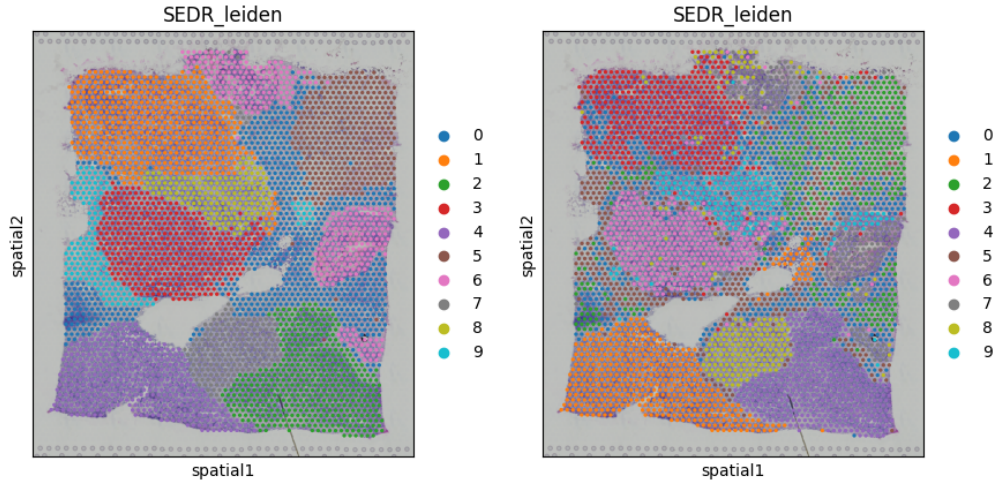


Figure 5.4: Comparing the results of SEDR\_v1 (left) versus SEDR\_v2 (right) results for a breast cancer dataset. Both models were trained for 500 epochs with DEC loss. Results obtained from original model achieved a loss of 212.9 while our modified model achieves a loss of 222.6

Next we compare the results generated using our modified SEDR architecture with the original tissue sample. Our test for a human brain tissue sample is shown in figure 5.5.

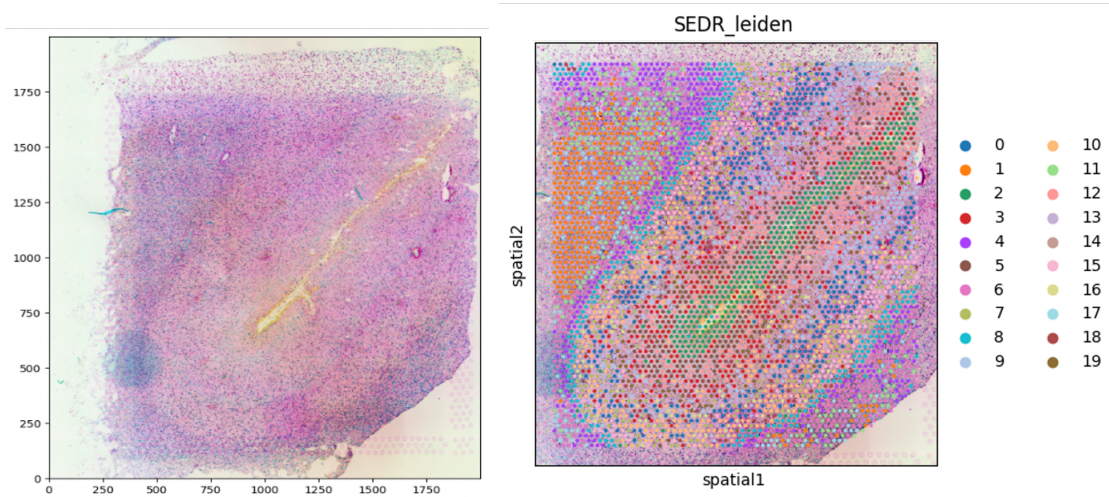


Figure 5.5: Comparison between the mapping of cells from the latent space of our model (right) as compared to the original tissue (left). We see that the combined latent space of spatial and expression data provides a good representation of the ground truth cell groupings in the tissue.

## SECTION 6

### CONCLUSIONS

Overall, we explored various Variational Autoencoders in this project to obtain a good latent space representation of scRNA data. We explored the architecture of these different VAEs and compared their performances on similar datasets. Based on the results and also for easier implementation, we decided to pick the VAE, scVI, to find a latent space representation for the scRNA data (Section 5.1.1). For our spatial transcriptomics dataset, we used a well-known VGAE, SEDR, to find the latent space representation of the spatial transcriptomics dataset. We modified the existing SEDR architecture to introduce additional hidden layers in SEDR’s encoder network to generate a shared latent space for the gene expression and spatial information of spatial transcriptomics data (Section 5.3). We obtained good latent space representations for scRNA and spatial transcriptomics datasets using these architectures. In addition, we were able to develop a discriminator model and utilize it to minimize the adversarial loss between the latent space of the spatial transcriptomics gene expression data and the scRNA data. (Section 5.2)

Although we could implement the two pipelines of our final model, the latent space of the scRNA dataset and the latent space of the Spatial Transcriptomics dataset, due to time constraints, we needed more time to finish mapping the two latent spaces. Once the mapping of the latent spaces is established, we can proceed to implement the final stage of our project, which is to infer the spatial information of the cells based on their gene expression information.

## SECTION 7

### RECOMMENDATIONS

We believe that this project has the potential to branch out in multiple directions and lead to exciting discoveries. We recommend future researchers to explore the following implementations and ideas:

#### **7.1. Perform more Quantitative and Qualitative Evaluations**

Although we investigated in common VAE/VGAE evaluation metrics (kNN, ARI) and machine learning performance assessments from literature, there are many more robust analysis that can be done on the overall performance and accuracy of the pipeline.

#### **7.2. Hyperbolic and Hyperspherical Latent Prior and Batch Correction**

According to our benchmarking results, scPhere [Ding and Regev, 2021] has the best performance among all VAE methods due to its enforcement of the prior distribution and batch correction considerations. Due to implementation constraints (see VAE Model Selection), we selected scVI over scPhere because the former was implemented in PyTorch, which is consistent with SEDR. Enforcing a hyperbolic or hyperspherical latent prior distribution will reduce distortion and cluttering of the latent representation [Ding and Regev, 2021]. Additionally, batch correction considers different biological or external factors when training the VAE. Exploring how the whole pipeline would perform if these considerations were included would be a valuable investigation.

#### **7.3. Pipeline Validation**

Currently, we only validated our pipeline using a small MERFISH dataset we obtained from Allen Brain Institute. It is necessary that future researchers validate this pipeline using larger datasets and experiment with data from different tissues, different number of cells/genes and different machine learning parameters (epoch, layers, activation functions, etc.).

## SECTION 8

### DELIVERABLES

We agreed to deliver the following deliverables to our project sponsor:

1. Proof of concept "latent mapping" architecture:
  - (a) Fully functional probabilistic machine learning pipeline utilizing VAE and VGAE
  - (b) Encoding of scRNA sequencing data and spatial transcriptomics data into a learned latent space
  - (c) Ability to perform latent mapping between latent spaces of our data
  - (d) Ability to infer spatial information from latent space
2. Published code-base and documentation of our research
3. Project proposal, presentations, and final report

# APPENDIX A

## BENCHMARKING VAE

### A.1. Benchmarking Methods

#### A.1.1. **scPhere: Deep Generative Model Embedding on Hyperspheres and Hyperbolic Spaces**

Current dimension reduction techniques tend to map high-dimensional data into a low-dimensional Cartesian latent space, which results in a cluttered, inaccurate latent representation of the data. **scPhere**, a deep generative model designed by Ding and Regev, is dedicated to mitigating this issue by mapping the RNA sequencing data into a hyperspherical or hyperbolic latent space. The resulting latent representation would be more naturally distributed and more accurately represent the high dimensional data. **scPhere** also implements multi-level batch correction. Ding and Regev [2021] Single-cell profiles in datasets are usually impacted by diverse factors including technical batch effects in different experiments with different lab protocols and also biological factors which include inter-individual variations, sex, diseases and tissue location. **scPhere** can learn models of data with multiple variables. In addition to this, **scPhere** preserves the structure of sc-RNA data in low dimensional spaces even. With all of these advantages, **scPhere** can help to find an optimal latent space representation for single-cell RNA sequencing data.

#### A.1.2. **scDHA: Single-Cell Decomposition using Hierarchical Autoencoder**

**scDHA**, single-cell decomposition using hierarchical autoencoder designed by Tran et al., uses a two-stage autoencoder to reduce noise and dimensions. The first stage uses a non-negative kernel autoencoder to obtain the part-based representation of the initial data by removing data that has a small contribution to the latent features. The second stage uses a stacked Bayesian self-learning network to project the data into multiple low-dimensional latent spaces and obtain their latent embeddings. This research implemented an R library and mainly focused on the performance of cell segregation through unsupervised learning, transcriptome landscape visualization, cell classification, and pseudo-time inference. Tran et al. [2019]



### A.1.3. VASC: Deep Variational Autoencoder

VASC, dimensionality reduction with deep Variational autoencoder is a multi-layer generative model for dimensionality reduction. In general, VASC models the distribution of high-dimensional original data  $P(X)$  by a set of latent variables  $z$ . The primary goal of VASC is to find the optimal  $z$  by capturing the intrinsic information of the input data. VASC tries to determine the posterior distribution  $P(z|X)$  by designing another common distribution family  $Q(z|X)$ , the variational distribution, to approximate  $P(z|X)$ .  $P(z|X)$  is then approximated from  $Q(z|X)$  using the Kullback-Leibler (KL) divergence between the two distributions. Deep neural networks are used in VASC to model the variational distribution  $Q(z|X)$ . The structure of VASC can be summarized as follows:

- **Input Layer:** This layer takes in the expression matrix from sc-RNA data and applies pre-processing such as log-transform for robust results. It also rescales the expression of every gene by normalizing it with the maximum expression value of an individual gene from the same cell.
- **Dropout Layer:** This layer sets some features to zero during the encoding phase. This introduces random noise in the data.
- **Encoder Network:** This is a three-layer fully connected neural network which carries out dimensionality reduction, reducing the dimension by 4 in each layer.
- **Latent sampling layer:** Latent variables  $z$  were modelled as a normal distribution with standard normal prior  $N(0, I)$ . This was used to estimate the posterior parameters,  $\mu$ ,  $\Sigma$ .
- **Decoder network:** The decoder network used the generated latent variables  $z$  to recover the original expression matrix, which was designed as a three-layer neural network fully connected with hidden unit dimensions of 32, 128 and 512 and an output layer.
- **Zero Inflated layer:** The zero-inflated (ZI) layer, which models the dropout events by the distribution  $e^{-\tilde{y}^2}$ , where  $\tilde{y}$  is the recovered expression value.

### A.1.4. DRA: Deep Adversarial Variational Autoencoder

DRA, dimensionality reduction with adversarial variational autoencoder, is a GAN-based architecture that combines an adversarial autoencoder and a variational autoencoder to reduce dimensions for single-cell RNA sequencing data. Wang and Gu aims to use a novel Dual Matching (AAVE-DM) system, which consists of a generator and two discriminators, to obtain the optimal latent representation of the data. Wang and Gu [2018]

### A.1.5. scVI: Deep generative modeling for single-cell transcriptomics

scVI is a popular software package that uses a variational autoencoder (VAE) for the analysis of single-cell RNA sequencing (scRNA-seq) data. The scVI model learns a low-dimensional representation of the gene expression data that captures the underlying biological variability. The model incorporates a number of design features that make it particularly well-suited to the analysis of scRNA-seq data, including the ability to handle sparse data, batch effects, and missing data. scVI has been shown to outperform existing methods for clustering, visualization, and imputation of scRNA-seq data, making it a valuable tool for researchers in the field. In addition, scVI is open source and available for use by the broader scientific community.

## A.2. Benchmarking Results

### A.2.1. Dataset Cell Type Distribution

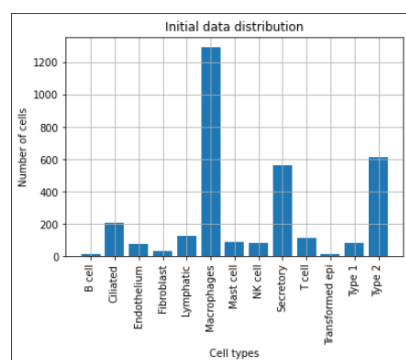


Figure A.1: Dataset A

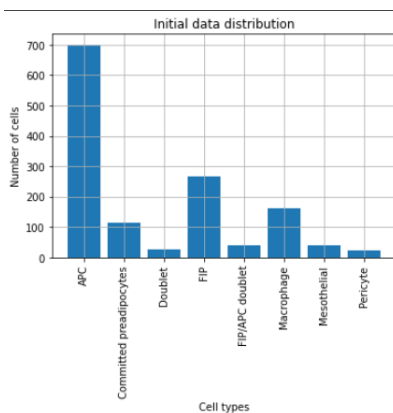


Figure A.2: Dataset B

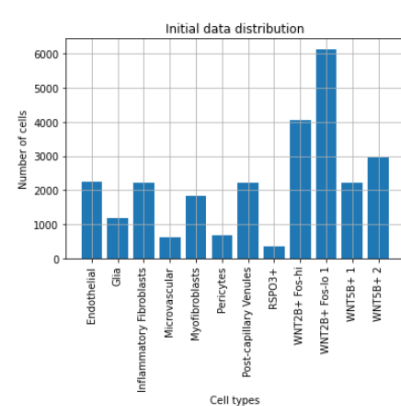


Figure A.3: Dataset C

### A.2.2. scSphere Supplemental Figures

Table A.1: scSphere Results for 2D, 10D, 20D for Dataset A

Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy	Confusion Matrix
2-D	<p>K value vs KNN Accuracy for ScSphere 2D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 2D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 11, accuracy = 88.902</p>
10-D	<p>K value vs KNN Accuracy for ScSphere 10D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 10D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 15, accuracy = 89.747</p>
20-D	<p>K value vs KNN Accuracy for ScSphere 20D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 20D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 1, accuracy = 71.17</p>

Table A.2: scSphere Results for 2D, 10D, 20D for Dataset B

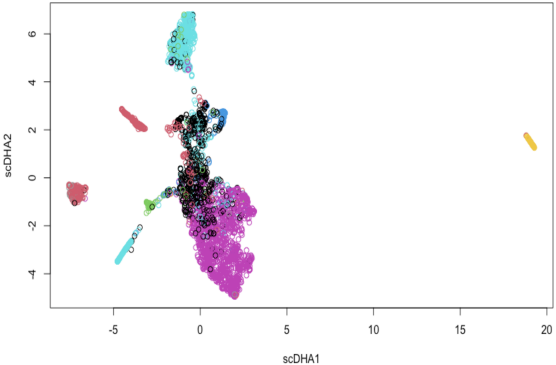
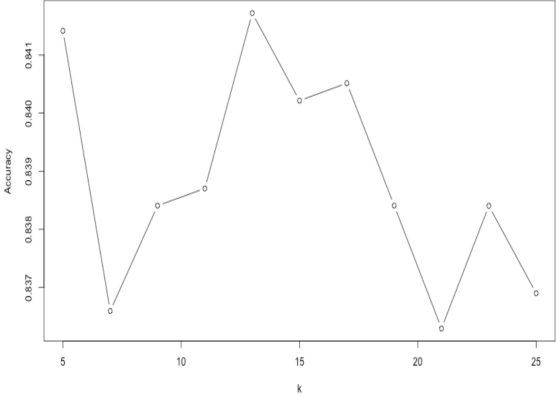
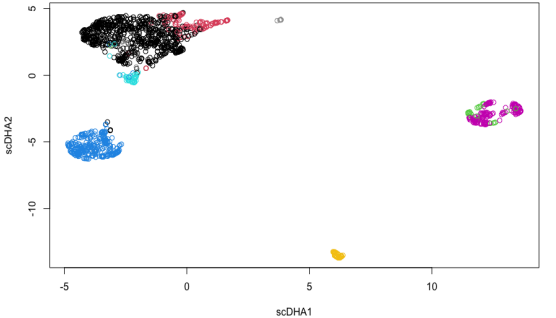
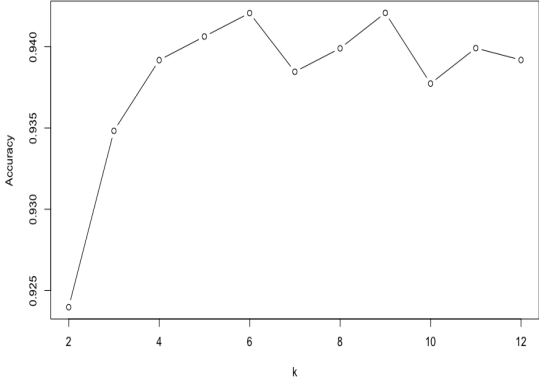
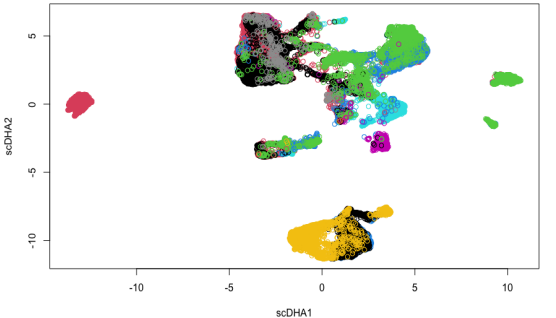
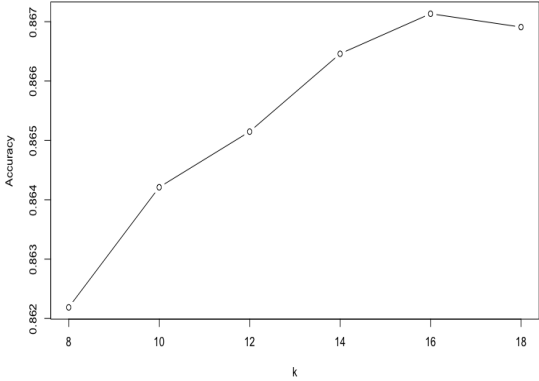
Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy	Confusion Matrix
2-D	<p>K value vs KNN Accuracy for ScSphere 2D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 2D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 17, accuracy = 95.942</p>
10-D	<p>K value vs KNN Accuracy for ScSphere 10D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 10D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 3, accuracy = 94.783</p>
20-D	<p>K value vs KNN Accuracy for ScSphere 20D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 20D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 1, accuracy = 95.362</p>

Table A.3: scSphere Results for 2D, 10D, 20D for Dataset C

Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy	Confusion Matrix
2-D	<p>K value vs KNN Accuracy for ScSphere 2D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for ScSphere 2D with vmf latent distribution</p>	<p>Confusion Matrix, k = 17, accuracy = 81.814</p>
10-D	<p>K value vs KNN Accuracy for ScSphere 10D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for 10D ScSphere, vmf latent distribution</p>	<p>Confusion Matrix, k = 15, accuracy = 89.747</p>
20-D	<p>K value vs KNN Accuracy for ScSphere 20D with vmf latent distribution</p>	<p>4-Fold Cross Validation KNN Accuracy for ScSphere 20D with vmf latent distribution</p>	<p>Confusion Matrix, k = 1, accuracy = 84.618</p>

### A.2.3. scDHA Supplemental Figures

Table A.4: scDHA Results for 15D for All Datasets

Data	Latent Distribution	4-fold cross-validation Accuracy																								
A		<p data-bbox="1094 386 1321 407">K=5-25 vs. Accuracy for lung_human_ASK440</p>  <table border="1" data-bbox="911 422 1463 814"> <caption>Accuracy for lung_human_ASK440</caption> <thead> <tr> <th>k</th> <th>Accuracy</th> </tr> </thead> <tbody> <tr><td>5</td><td>0.841</td></tr> <tr><td>7</td><td>0.827</td></tr> <tr><td>9</td><td>0.834</td></tr> <tr><td>11</td><td>0.835</td></tr> <tr><td>13</td><td>0.841</td></tr> <tr><td>15</td><td>0.840</td></tr> <tr><td>17</td><td>0.841</td></tr> <tr><td>19</td><td>0.834</td></tr> <tr><td>21</td><td>0.827</td></tr> <tr><td>23</td><td>0.834</td></tr> <tr><td>25</td><td>0.827</td></tr> </tbody> </table>	k	Accuracy	5	0.841	7	0.827	9	0.834	11	0.835	13	0.841	15	0.840	17	0.841	19	0.834	21	0.827	23	0.834	25	0.827
k	Accuracy																									
5	0.841																									
7	0.827																									
9	0.834																									
11	0.835																									
13	0.841																									
15	0.840																									
17	0.841																									
19	0.834																									
21	0.827																									
23	0.834																									
25	0.827																									
B		<p data-bbox="1094 848 1305 869">K= 2 ~ 12 vs. Accuracy for adipose</p>  <table border="1" data-bbox="911 890 1446 1268"> <caption>Accuracy for adipose</caption> <thead> <tr> <th>k</th> <th>Accuracy</th> </tr> </thead> <tbody> <tr><td>2</td><td>0.925</td></tr> <tr><td>3</td><td>0.935</td></tr> <tr><td>4</td><td>0.939</td></tr> <tr><td>5</td><td>0.940</td></tr> <tr><td>6</td><td>0.941</td></tr> <tr><td>7</td><td>0.938</td></tr> <tr><td>8</td><td>0.939</td></tr> <tr><td>9</td><td>0.941</td></tr> <tr><td>10</td><td>0.938</td></tr> <tr><td>11</td><td>0.939</td></tr> <tr><td>12</td><td>0.939</td></tr> </tbody> </table>	k	Accuracy	2	0.925	3	0.935	4	0.939	5	0.940	6	0.941	7	0.938	8	0.939	9	0.941	10	0.938	11	0.939	12	0.939
k	Accuracy																									
2	0.925																									
3	0.935																									
4	0.939																									
5	0.940																									
6	0.941																									
7	0.938																									
8	0.939																									
9	0.941																									
10	0.938																									
11	0.939																									
12	0.939																									
C		<p data-bbox="1094 1310 1305 1331">K= 8 ~ 18 vs. Accuracy for uc_stromal</p>  <table border="1" data-bbox="911 1352 1446 1730"> <caption>Accuracy for uc_stromal</caption> <thead> <tr> <th>k</th> <th>Accuracy</th> </tr> </thead> <tbody> <tr><td>8</td><td>0.862</td></tr> <tr><td>10</td><td>0.864</td></tr> <tr><td>12</td><td>0.865</td></tr> <tr><td>14</td><td>0.866</td></tr> <tr><td>16</td><td>0.867</td></tr> <tr><td>18</td><td>0.867</td></tr> </tbody> </table>	k	Accuracy	8	0.862	10	0.864	12	0.865	14	0.866	16	0.867	18	0.867										
k	Accuracy																									
8	0.862																									
10	0.864																									
12	0.865																									
14	0.866																									
16	0.867																									
18	0.867																									

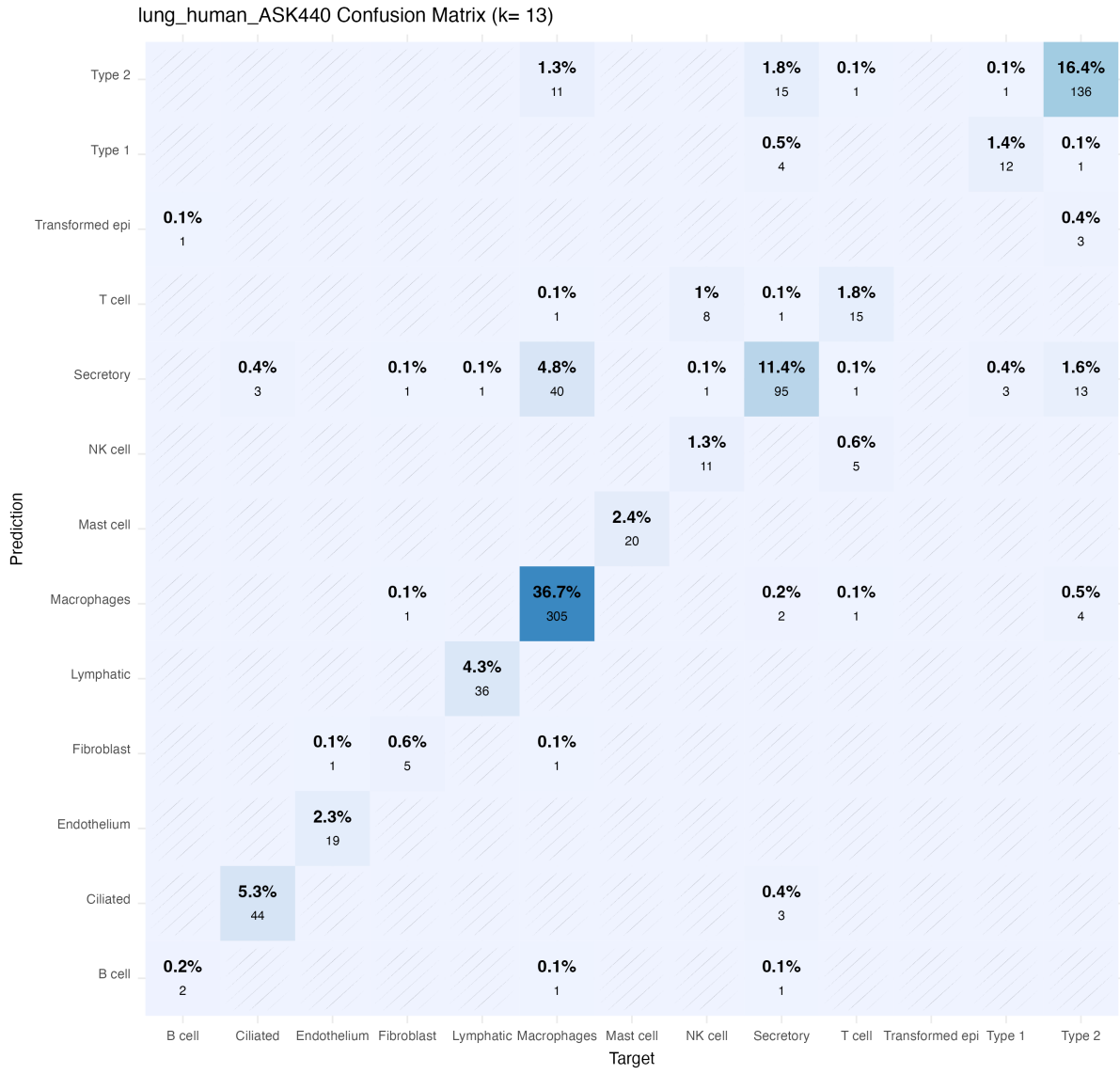


Figure A.4: Confusion Matrix for Max kNN ( $k = 13$ )

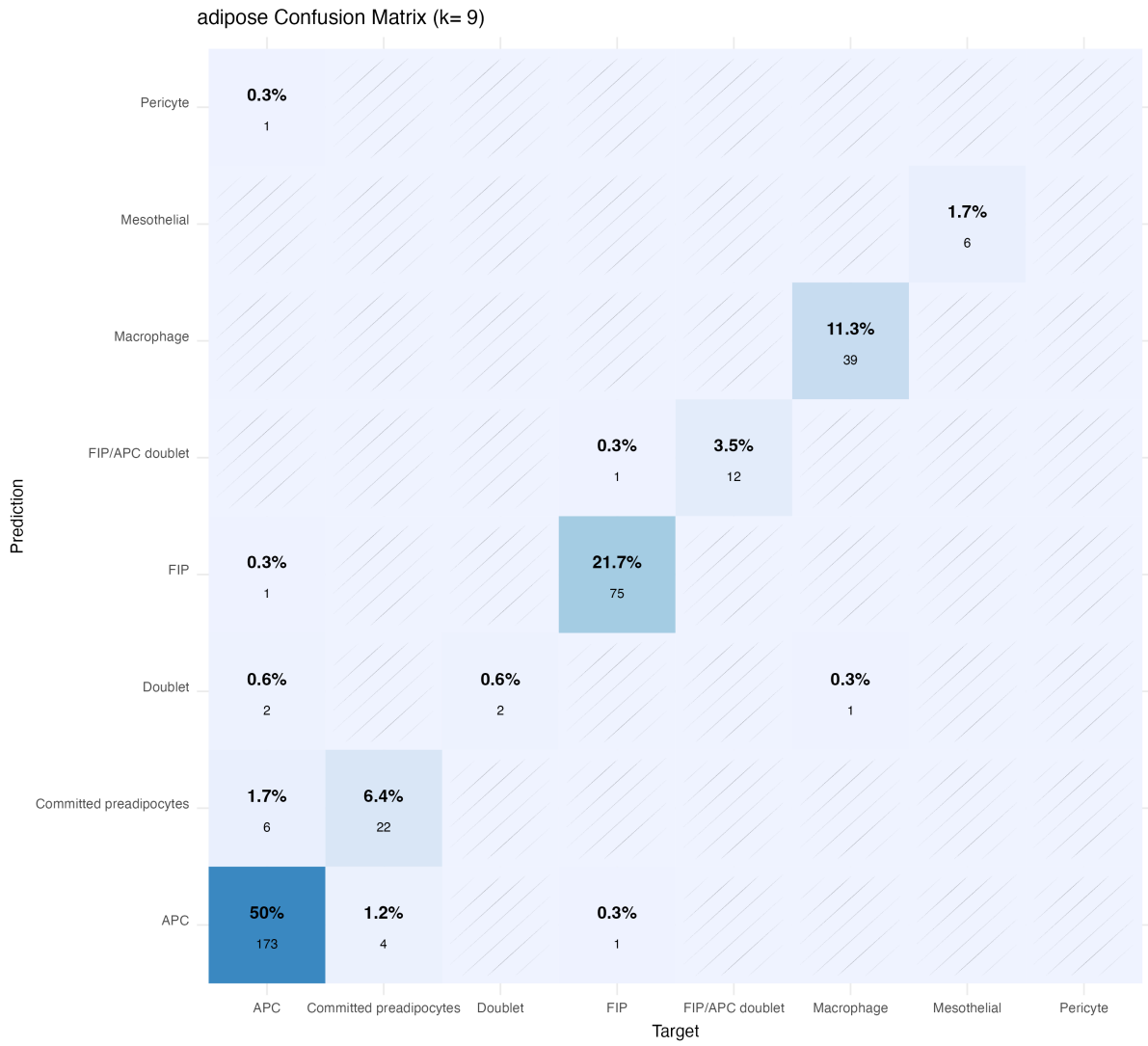


Figure A.5: Confusion Matrix for Max kNN ( $k = 9$ )



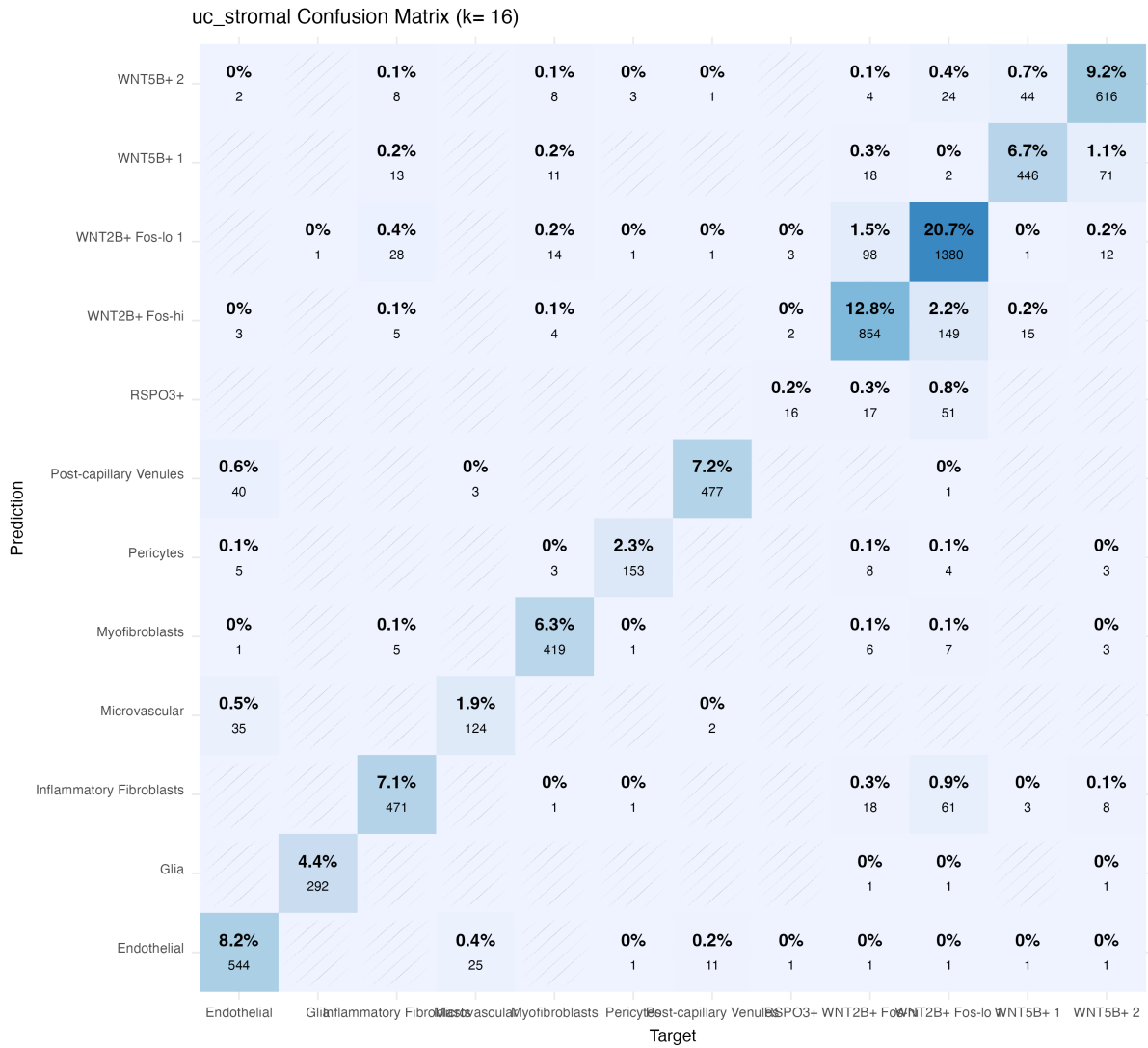


Figure A.6: Confusion Matrix for kNN at max accuracy ( $k = 16$ )

### A.2.4. VASC Supplemental Figures

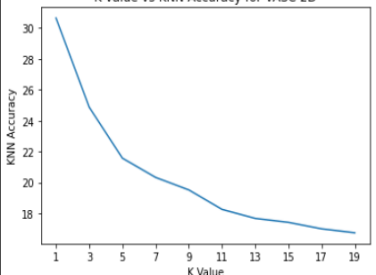
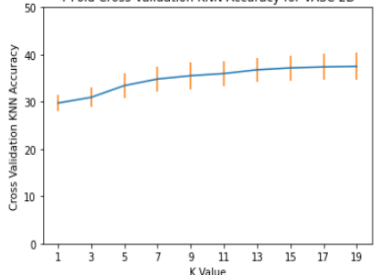
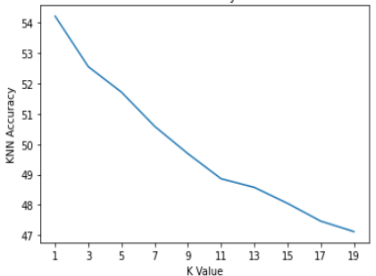
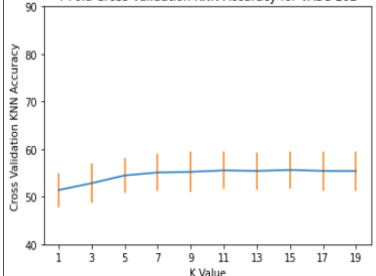
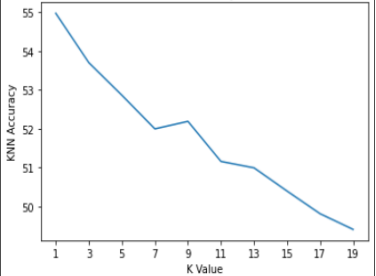
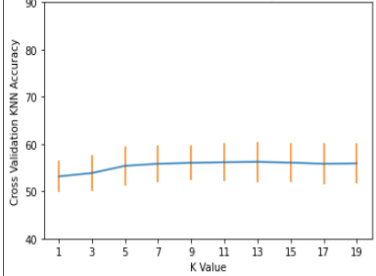
Table A.5: VASC Results for 2D, 10D, 20D for Dataset A

Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy	Confusion Matrix
2-D	<p>K value vs KNN Accuracy for VASC 2D</p>	<p>4-Fold Cross Validation KNN Accuracy for 2D VASC</p>	<p>Confusion Matrix, k = 1, accuracy = 64.415</p>
10-D	<p>K value vs KNN Accuracy for VASC 10D</p>	<p>4-Fold Cross Validation KNN Accuracy for VASC 10D</p>	<p>Confusion Matrix, k = 15, accuracy = 89.747</p>
20-D	<p>K value vs KNN Accuracy for VASC 20D</p>	<p>4-Fold Cross Validation KNN Accuracy for 20D VASC</p>	<p>Confusion Matrix, k = 1, accuracy = 69.964</p>

Table A.6: VASC Results for 2D, 10D, 20D for Dataset B

Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy	Confusion Matrix
2-D	<p>K value vs KNN Accuracy for VASC 2D</p>	<p>4-Fold Cross Validation KNN Accuracy for VASC 2D</p>	<p>Confusion Matrix, k = 9, accuracy = 84.05</p>
10-D	<p>K value vs KNN Accuracy for VASC 10D</p>	<p>4-Fold Cross Validation KNN Accuracy for VASC 10D</p>	<p>Confusion Matrix, k = 5, accuracy = 86.08</p>
20-D	<p>K value vs KNN Accuracy for VASC 20D</p>	<p>4-Fold Cross Validation KNN Accuracy for VASC 20D</p>	<p>Confusion Matrix, k = 3, accuracy = 86.66</p>

Table A.7: VASC Results for 2D, 10D, 20D for Dataset C

Dim	kNN Accuracy vs k	4-fold cross-validation Accuracy
2-D	 <p>K value vs KNN Accuracy for VASC 2D</p>	 <p>4-Fold Cross Validation KNN Accuracy for VASC 2D</p>
10-D	 <p>K value vs KNN Accuracy for VASC 10D</p>	 <p>4-Fold Cross Validation KNN Accuracy for VASC 10D</p>
20-D	 <p>K value vs KNN Accuracy for VASC 20D</p>	 <p>4-Fold Cross Validation KNN Accuracy for VASC 20D</p>

## BIBLIOGRAPHY

- Jinmiao Chen, Huazhu Fu, Hang XU, Kelvin Chong, Mengwei Li, Kok Siong Ang, Hong Lee, Jingjing Ling, Ao Chen, Ling Shao, and Longqi Liu. Unsupervised spatially embedded deep representation of spatial transcriptomics. 06 2021. doi: 10.21203/rs.3.rs-665505/v1.
- Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nature Communications*, 12, 05 2021. doi: 10.1038/s41467-021-22851-4.
- Eugene Lin, Sudipto Mukherjee, and Sreeram Kannan. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell rna sequencing analysis. *BMC Bioinformatics*, 21(1), 2020. doi: 10.1186/s12859-020-3401-5.
- Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL <https://doi.org/10.1038/s41592-018-0229-2>.
- Gennert Schier Regev Satija, Farrell. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(12):1053–1058, 2015. doi: 10.1038/nbt.3192. URL <https://doi.org/10.1038/nbt.3192>.
- Duc Tran, Hung Nguyen, Bang Tran, and Tin Nguyen. Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*, 2019. doi: 10.1101/799817. URL <https://www.biorxiv.org/content/early/2019/10/10/799817>.
- Dongfang Wang and Jin Gu. Vasc: Dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, Proteomics amp; Bioinformatics*, 16(5):320–331, 2018. doi: 10.1016/j.gpb.2018.08.003.