Aligning Heterogenous Single Cell Assay Datasets

by

Abigail Katcoff

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Author . . . .
Department of Electrical Engineering and Computer Science
May 24, 2019

Certified by . . . .
Caroline Uhler
Associate Professor
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Aligning Heterogenous Single Cell Assay Datasets

by

Abigail Katcoff

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Pluripotent stem cells offer strong promise for regenerative medicine but the pluripotent cell state is poorly understood. The goal of this thesis is the development of methods to analyze how the multiple facets of cell state–including gene expression, chromosome contacts, and chromatin accessibility–relate in the context of stem cells. The variability of each of these characteristics cannot be deduced from population studies, and while recent advances in single-cell transcriptomics have led to the development of a number of different single-cell assays, datasets that collect multiple types of assays on the same cells are rare. In this thesis, we explore the ability of three methods to integrate datasets from different single-cell assays based on an existing paired single-cell dataset of ATAC-seq and RNA-seq for human A549 cells. We then apply these methods to map the variability between three single-cell datasets–ATAC-seq, RNA-seq, and Hi-C–on pluripotent mouse embryonic stem cells and assess the performance of these methods.

Thesis Supervisor: Caroline Uhler
Title: Associate Professor

Acknowledgments

Thank you to Anastasiya Belyaeva for her help advising me, preprocessing data, and more.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pluripotency is the ability of a stem cell to become any cell type in an organism. Pluripotent stem cells including embryonic stem cells have an unlimited capacity to self-renew and can develop into any type of tissue. These abilities give these cells the potential to renew damaged tissues or act as implants, making them important in regenerative medicine. Despite these capabilities, the pluripotent cell state is not well understood [2].

Variations in gene expression are often observed in pluripotent stem cells. Populations of pluripotent stems cells [PSCs] have been observed to have subpopulations, which may allow the cells to respond to environmental cues during self-renewal, differentiation, reprogramming, and apoptosis. To understand the regulatory mechanisms in PSC cell states, analysis must be done at the single-cell level to view the full spectra of gene expression in the population, which would be masked in pooled data in the population [2]. Looking at gene expression at the single cell level reveals variability that cannot be deduced from population studies.

Recent advances in single-cell transcriptomics have allowed collecting single-cell RNA-sequencing (scRNA-seq) data, which allows for the quantification of gene expression at a single cell level [2]. A method called flow cytometry sorting has allowed efficient collection of single-cell chromosomal contact maps: scHi-C (single-cell high-resolution chromosome conformation capture) data [3]. These contact maps represent the counts of contacts between each genomic region in a single cell. Additionally, a high-throughput model allows fast collection of single-cell ATAC-seq data, an assay for transcription of accessible chromatin [21].The goal of this thesis is to explore whether datasets of these types can be used to connect variation in gene expression with variation in chromosome contacts to better understand the relation between gene expression, pluripotency and spatial organization of chromosomes.

Chapter two describes experiments to determine what methods work well to map together single cell datasets for which we know the correspondences. In chapter three we apply knowledge learned in chapter two to unpaired datasets featuring single-cell Hi-C, RNA-seq, and ATAC-seq data. Chapter four discusses conclusions and future work.

1.1 Related Work

Related work can be broken into three sections: work describing current knowledge of the connection between gene expression and chromosome interactions, work on aligning single-cell data, and general methods for aligning datasets used in computer vision.

1.1.1 Gene Expression and Chromosome Interactions

A number of studies have already established the connections between gene expression and chromosome interactions. Fluorescent in situ hybridization and chromosome capture experiments including Hi-C suggest that regions of chromosome intermingling contain a high density of regulatory sequences and epigenetic modifications for repression and activations. These intermingling regions are thought to serve as epigenetic hotspots. Additionally chromosome organization has been found to be cell-type specific [1].

Chromosomes contain regions of genes that are more or less expressed. Regions of chromatin called heterochromatin are characterized by methylation of histone H3 and silencing of gene expression[6]. One view on heterochromatin is that it leads to chromatin compaction resulting in steric exclusion of regulatory proteins, while a newer hypothesis is that formation of chromatin is mediated by phase-separation [6]. We would expect to find more heterochromatin regions in less pluripotent cells, since once cells commit to a lineage, more genes are silenced. Additionally, analysis of Hi-C data has resulted in the discovery of topological domains within chromosomes, highly interacting chromatin domains that limit heterochromatin spread [7]. Active genes tend to interact with each other and inactive genes tend to interact with each other, forming active and inactive domains [8].

1.1.2 Matching Single-Cell Assay Data

Previous work has been done on matching single-cell datasets. A method call MATCHER predicts true single cell correlations between DNA methylation and gene expression without using known cell correspondences based on a single latent variable–pseudotime [13]. However, this method was not applied to single-cell Hi-C data, and one single latent variable may not be enough to capture the joint variability in our datasets.

1.1.3 Image-to-Image Translation Methods

Our problem is similar to the problem of image-to-image translation in computer vision, where one might want to apply methods learned on images in one domain (e.g. computer graphics) to images in another domain (e.g. photographs). Instead of image domains, we have different viewpoints of cells–chromosome contacts and gene expression. In particular, we are interested in unsupervised or semi-supervised methods, since we do not have paired single cell RNA-seq and Hi-C. Therefore, we plan to adapt existing unsupervised methods from this realm. An example of such a method is CycleGAN–a cycle consistent adversarial network that is capable of translating images between domains unsupervised [12]. This neural network learns two functions: G mapping domain X to domain Y, and F mapping domain Y to domain X, and tries to ensure F(G(x))=x and G(F(y))=y.

Two other methods, Unsupervised Image-to-Image Translation Networks (UNIT)

[17] and Multi-Domain Translation by Learning Uncoupled Autoencoders [23] expand on CycleGAN by using paired autoencoders to explicitly learn a latent space between two domains, which would allow us to walk along the differentiation path. Since learning functions between two images might be easier than learning functions between two different types of data collected from cells, we can also incorporate a loss based on points staying in the correct cluster or treatment when transformed from one data type to the other, to semi-supervise the problem.

1.2 Datasets

Single-cell RNA-seq (scRNA-seq) quantifies gene expression at a chromosome level. Mouse embryonic stem cells (mESCs) cultured in fetal bovine serum and cytokine leukemia inhibitory factor (LIF), two factors used to maintain self-renewal of cells in vitro, have shown fluctuations in pluripotency genes Nanog, Dppa3, and Rex1 (Zfp42) [4]. Using serum/LIF cultured mESCs, Kolodziejczyk (2015) found 3 subpopulations in a group of 250 cells. One subpopulation of 39 cells expressed higher levels of differentiation markers and low levels of pluripotency factors. Another subpopulation of 42 cells had slightly lower levels of some pluripotency genes, high expression of other pluripotency genes and some expression of differentiation markers. The remainder of the cells expressed very high levels of plutipotency factors. Kolodziejczyk (2015) concludes these subpopulations represent cells on the differentiation path, differentiation "primed" cells, and a more pluripotent cell groups respectively. The goal of this project is to see what structural organizations are associated with each group of these cells, using mESC scHi-C.

Obtaining thousands of contact maps for individual cells is now possible using a new method that scales up previous Hi-C methods using flow cytometry sorting [3]. Although the breakthrough article featuring this method focuses on cells cultured in another medium called 2i, the data available with the paper also features 859 scHi-C data from diploid cells cultured in serum, making these cells comparable to the scRNA-seq dataset described above. Thus, we can analyze the scHi-C data of these cells together with the data from Kolodziejczyk (2015).

In addition to the RNA-seq and Hi-C datasets, we use data from a new high-throughput single cell method for single cell ATAC-seq, which quantifies chromatin accessibility [21]. We obtained mESC scATAC-seq profiles for 559 of the cells used in the study.

To ensure efficacy of our methods mapping together the scRNA-seq and sc-Hi-C, we make use of paired single-cell assay data from different domains. Angermueller (2016) features 61 mouse embryonic stem cells with paired methylation and gene expression data, utilizing a new method known as scM&T-seq [14]. Cao et. al (2018) features transcriptome and chromatin accessibility profiles for 11,296 adult mouse kidney cells and human lung adenocarcinoma-derived A549 cells after 0, 1, or 3 hours of 100 nM DEX treatment (3,260) [15]. Cao et. al is a big enough dataset to apply deep learning methods.

Chapter 2

Mapping Single-Cell Assay Data

To ensure that we will be able to map unsupervised data together, we first must make sure we are capable of mapping paired datasets. Three methods were tried: Canonical Correlation Analysis, paired autoencoders based on an adaptation of UNIT [17], and Deep CCA [18]. We primarily used a dataset of paired scRNA-seq and scATAC-seq data on A549 cells [15].

2.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is the most basic of methods tried. The method has the benefit of being nonparametric, but it does not capture nonlineari-

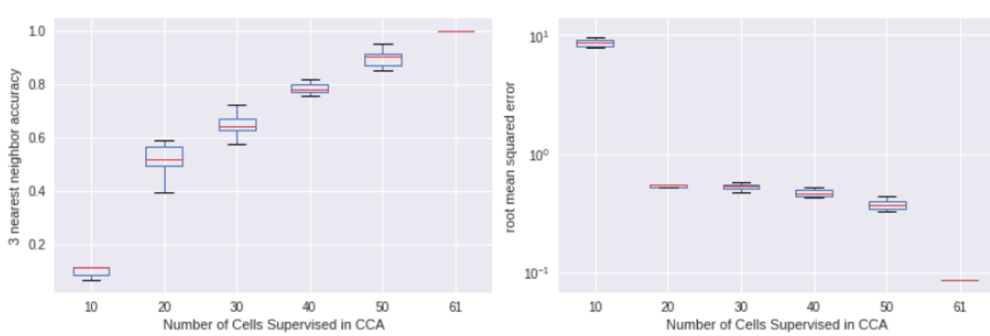


Figure 2-1: Results of CCA and OT method run with different amounts of supervision on the Angermueller dataset

ties in the data. Canonical correlation analysis aims to find weights for each of the dimensions in the two domains such that the outcome is most correlated:

$$(a', b') = argmax_{a,b} corr(a^T X, b^T Y) \tag{2.1}$$

Note that this method requires paired data to find these weights, which is not a realistic setting for our problem since the correspondences between the datasets are the information we are trying to find. After we projected points using the found weights, we aligned the distributions using optimal transport, a method to move one distribution of data onto another using the minimal amount of work. The goal was to see how many data-points we needed to supervise with CCA to obtain adequate results when running entropy-regularized optimal transport on the data transformed by our learned weights.

We used various sized subsets of the data to learn CCA weights; these could then be applied to the unsupervised points to project them to the new space. On the Angermueller dataset, we ran this method with differing number of sample sizes, with 10 different samples for each sample size, mapping gene expression to methylation. We evaluated the results based on 3-nearest neighbor accuracy (probability of a point ending up closest to the correct point or its two closest neighbors) and root mean squared error in the methylation CCA space (Figure 2-1). We also evaluated 10-nearest neighbor accuracy with 4 samples each of differing sample sizes on the Cao gene expression/chromatin accessibility dataset, a much larger dataset (Figure 2-2). The first 10 canonical components were used for each dataset.

These results show that we would need a substantial proportion of points to accurately map points from one domain to another using this method. While this method worked well when highly supervised, we note that this method does not take into account the structure of our unsupervised points when assigning weights via CCA–ideally our transformations would keep the points' relative distances consistent. Therefore, we can surely improve upon this method. Thus, we attempted other unsupervised and semi-supervised methods to align our single-cell assay data, namely

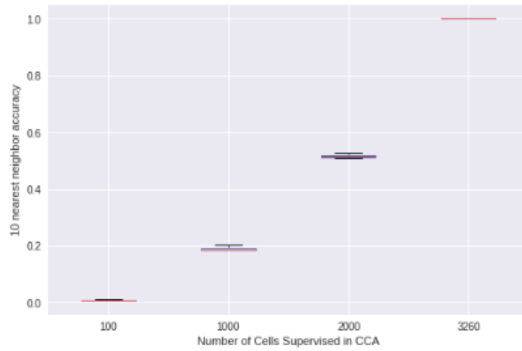


Figure 2-2: Results of CCA and OT method run with different amounts of supervision on the Cao dataset

deep learning methods such as paired autoencoders and Deep CCA.

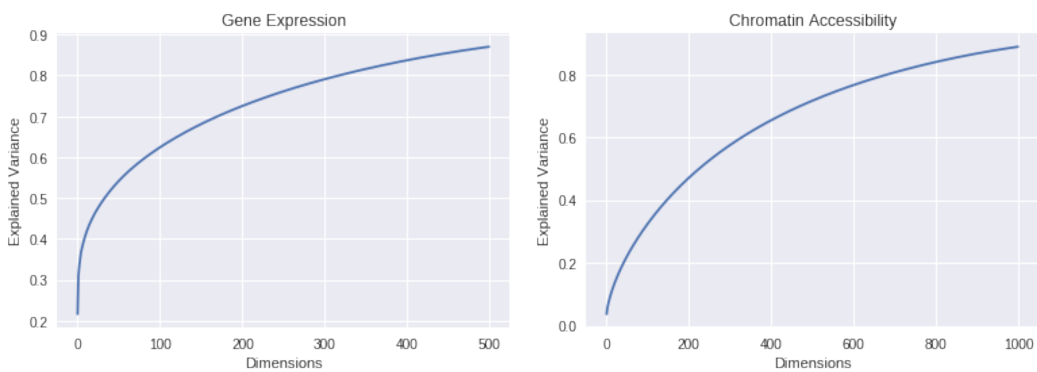


Figure 2-3: Cumulative explained variance with each added dimension of PCA on the Cao dataset

2.2 Autoencoders

The next methods we tried were based on paired autoencoders ([17] [23]). We began by encoding each of the scRNA-seq and scATAC-seq datasets separately to assess the dimensionality required and to understand the difficulty of the problem. We also wanted to confirm that the scRNA-seq and scATAC-seq profiles, when projected into the latent space, were different across treatment times, which we saw in the raw data during preliminary analysis (Appendix A Figure A-1). Before running any autoencoders, we ran PCA on the data to see how many dimensions were required to explain a significant portion of variance in the data (2-3).

The draw-back to the deep-learning methods is that they require substantial data to not overfit, so our evaluation of the methods was based solely on the Cao dataset's thousands of A549 cells, which also have treatment time labels that can be used to semi-supervise learning. The log(1+p) of both the scRNA-seq and scATAC-seq datasets was used instead of the raw values since the data had a large range of counts. Each feature in the datasets was normalized to be zero mean and unit variance. Additionally, to narrow the size of the dataset for easier training, only differentially expressed genes in the scRNA-seq data and differentially accessible regions in the scATAC-seq were included in training.

2.2.1 Separate Autoencoders

To understand the dimensionality required to autoencode the dataset, separate autoencoders were created for the scATAC-seq and scRNA-seq data. Initially, a standard variational autoencoder [22] (with a Kullback-Leibler divergence latent loss and mean squared error reconstruction loss) was attempted, but it was discovered that despite parameter tuning, the autoencoder was unable to accurately recover the full spectrum of values in the original data (Figures A-4 and A-3 in Appendix A). Preliminary analysis (Figure A-2 in Appendix A) of the Cao dataset showed that the scRNA-seq and scATAC-seq data was very zero-inflated, and perhaps the zero-inflatedness caused the autoencoder to be unable to learn higher values.

It was found that a variational autoencoder with Maximum Mean Discrepancy (MMD) loss as the latent loss was able to recover a bigger spectrum of values (Figures A-6 and A-5 in Appendix A). Still, accurately recovering the scATAC-seq and scRNA-seq values proved to be a challenge, which led us to remove the variational component of our autencoders for the next stage and deprioritize reconstructions in favor of generating a meaningful latent space. Fortunately, the latent spaces learned by the autoencoders appeared to maintain information on treatment time, which can be seen in the PCA plots in Figure 2-4.

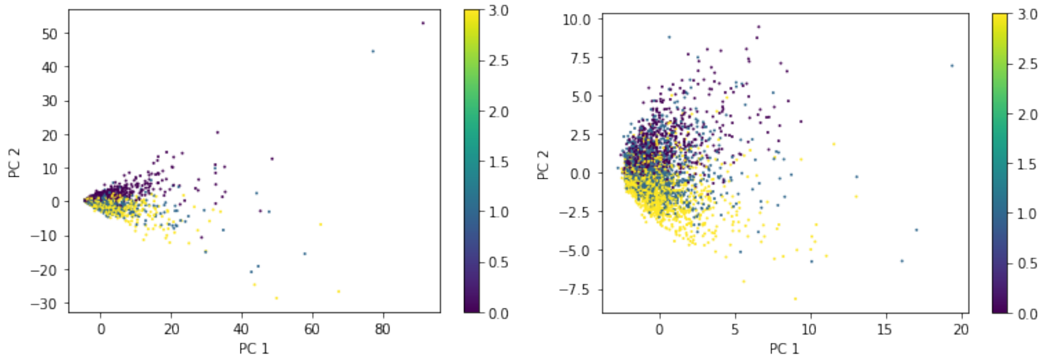


Figure 2-4: Latent spaces found by MMD autoencoders for scATAC-seq (left) and scRNA-seq (right) colored by treatment time

A latent space of size 100 was used for each of the two autoencoders. The best parameters including the network architecture are included in Table B.1 and Figure B-1 in Appendix B.

2.2.2 Paired Autoencoders

The image-to-image translation approach used by [17] was altered for this task. Instead of a separate discriminator for each domain, a single discriminator was used in the latent space, as in [23]. The reason for this choice was that the original discriminators in [17] acted to discriminate between reconstructions and the original data. As we saw in Section 2.2.1, the autoencoders were ineffective at accurately reproducing the original data; thus, the task of discriminating between the original and reconstructed data was simply too easy. Preliminary trials using the original image-to-image network with two discriminators confirmed this problem; the discriminator losses were close to zero and the projections of the two datasets into the latent space did not overlap. Therefore, one discriminator to discriminate between the two datasets' projections into the latent space was used. To simplify the network, we also removed the variational component, which experimentally improved results.

In addition to components in the original network, we added a classifier in the latent space to classify the treatment times of the cells. This additional component was included to help the latent space maintain information on the treatment times.

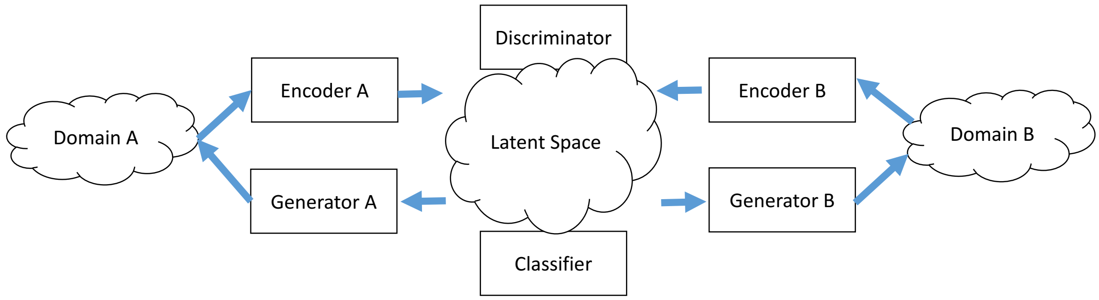


Figure 2-5: Diagram of paired autoencoder network

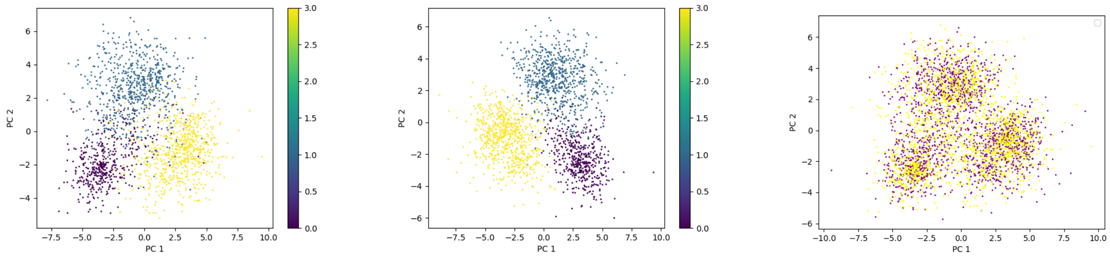
As seen in the preliminary analysis, scRNA-seq/scATAC-seq profiles were different across treatment times (Figure A-1), a trend we sought to retain. The full network is outlined in Figure 2-5.

Initial results with the raw scATAC-seq data did not seem promising; we considered a transformation of the data that physically corresponds better with scRNA-seq data. We acquired a transcription factor motif by cells matrix from the authors. Transcription factors are coded by genes and determine which genes are transcribed, making this data align better with the scRNA-seq data, which quantifies gene expression. Therefore, we used this data instead. However, this matrix contained fewer cells that overlapped with the scRNA-seq data; overall we now had 1874 cells. Both the scATAC-seq transcription factor data and the differentially expressed scRNA-seq data that were fed into the network were logged and normalized to zero mean and unit variance. A summary of all the losses in this network is included in Appendix B (Table B.2).

To evaluate these methods, we created an 80-20 train-test split of the paired data. As a metric, we considered k-nearest neighbors accuracy between the projections of each dataset into the latent space:

$$knn(A, B) = \frac{\sum_i \mathbb{1}(b'_i \in a^k_i)}{n} \tag{2.2}$$

A and B are sets of vectors, with b_i as a vector in B and a_i as its pair in A, and b'_i and a'_i as the encoded versions of b_i and a_i. The set a^k_i contains the k nearest neighbors of a'_i in A', the set of vectors in A projected into the latent space. Note



(a) scATAC-seq (b) scRNA-seq (c) Both datasets

Figure 2-6: PCA plots to visually validate paired autoencoder network latent space. The first two are plots of the a) scATAC-seq and b) scRNA-seq data projected into the latent space colored by treatment time. The third is a plot of both datasets projected into the latent space colored by dataset.

that this metric is not symmetric; $knn(A, B)$ does not necessarily equal $knn(B, A)$. Therefore, both of these values were averaged in further analysis. The training sets and test sets have different numbers of points, so to make the k-nearest neighbors accuracy for both datasets comparable, the sets of possible neighbors considered for both metrics included all 1874 cells.

We wanted to see how many points we needed to supervise to have a well-working model. Points were supervised by adding a supervision loss corresponding to the mean absolute error between paired training points in the latent space. First we fit a model to work best with supervising 100% of the training data (the test data was not supervised). Once we achieved a satisfactory test accuracy with 100% supervision, we applied the same model with differing amounts of training data being supervised (0%, 5%, 25%, 50% and 75%).

Details of the parameters for the best performing network used are supplied in Appendix B (Table B.3, Table B.4). The k-nearest neighbors accuracy from this method can be seen in Figure 2-7.

We also generated plots of the scATAC-seq and scRNA-seq data points projected into the latent space to confirm that 1) the classes are separable and 2) the datasets overlap. The plots can be seen in Figure 2-6.

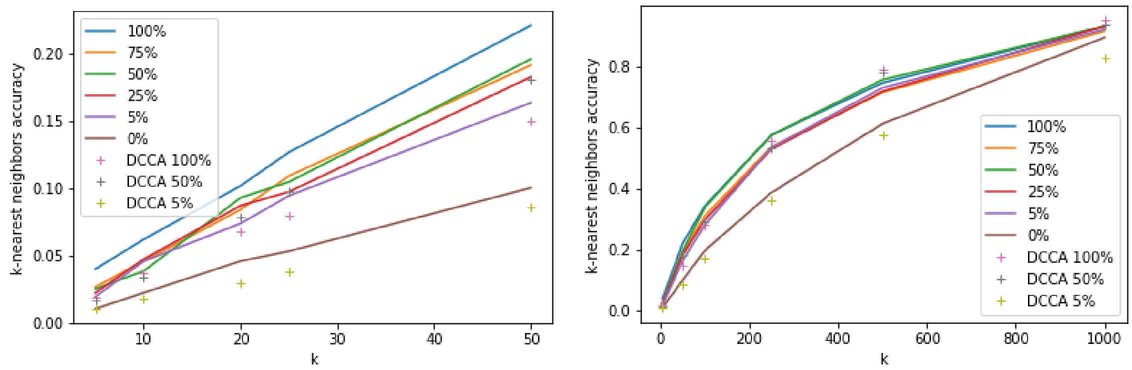


Figure 2-7: K-nearest neighbors accuracy for Paired Autoencoders and Deep CCA with differing percentages of supervised points in the training set. The plot on the left features k values up to 50 and the plot on the right features k values greater than 50 and up to 1000 (more than half the points).

2.3 Deep CCA

We next tried Deep CCA [18], a method that uses deep learning to learn nonlinear transformations on two inputs such that they are highly correlated in the transformed space. This method was considered as it had the benefits of regular CCA while also being nonlinear and was a simpler method to consider than the paired autoencoder model. Deep CCA is compared to the previous method in the Figure 2-7. As before, we evaluated Deep CCA with varying amount of supervised data. Details on the Deep CCA model used can be found in Appendix B Table B.7.

A drawback of a CCA approach is that it requires paired data, which we do not have for the mESC data of interest. Therefore, we wanted to verify how well Deep CCA would work when supplied incorrect pairs. We needed a way to provide Deep CCA paired data that was informative, since supplying completely random pairs would misguide the algorithm. It is a common setting in biology to have class labels (in the mESC datasets we have cell cycle), so we took advantage of our treatment labels when generating the pairs. For each cell, we generated 100 random cells within the same class to use as its partners. Therefore, the number of pairs supplied to Deep CCA was 100x that originally supplied. The results of this experiment are found in Figure 2-8. Interestingly, when k is large (the size of each class or bigger), the

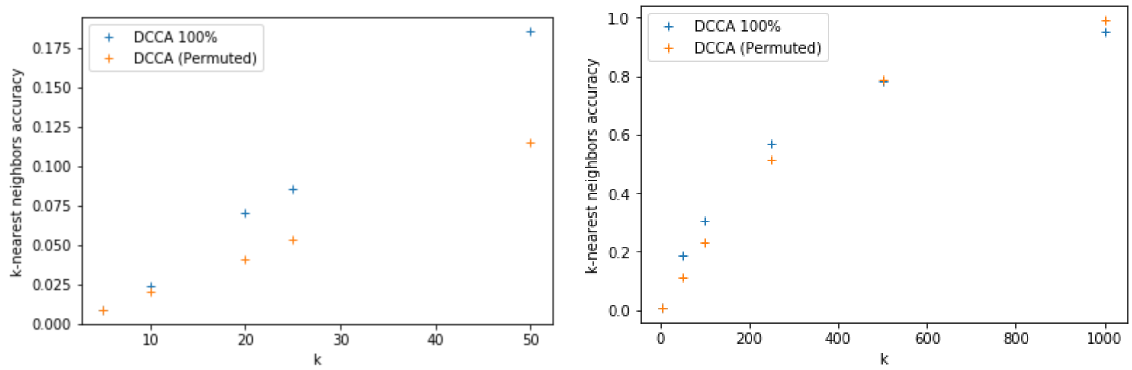


Figure 2-8: Deep CCA permutation experiment results

permuted approach performs as well as the original. The experiment confirms the use of Deep CCA as a potential method to align heterogenous datasets despite lack of paired cells.

Chapter 3

Application to Single Cell Mouse Embryonic Stem Cell Data

In this chapter, we adapt the paired autoencoder method from Chapter 2 to find a shared latent space for the following types of unpaired single cell assay data for mouse embryonic stem cells: scATAC-seq [21], scRNA-seq [4], and scHi-C [3], as described in Chapter 1. Specifically, we trained the method on data from cells which were cultured in a serum medium.

3.1 Supervision

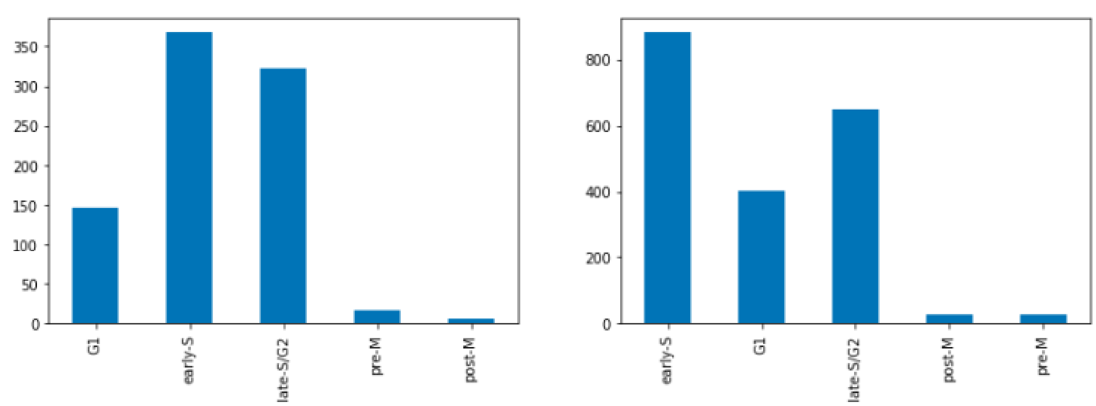


Figure 3-1: Distribution of scHi-C cell cycle labels for serum (left) and 2i (right)

As discovered in the last chapter, supervision helps significantly to align datasets using paired autoencoders. Therefore, we aimed to use prior knowledge in guiding our model to learn a useful latent space. The supervision strategy is summarized in Figure 3-3.

We used a single classifier to predict cell cycle labels for the scRNA-seq and scHi-C datasets (we did not have such labels available for the scATAC-seq data). Cell cycle labels were already available as part of the scHi-C dataset. Figure 3-1 shows the distribution of cell cycle labels in the scHi-C dataset. The scRNA-seq labels were computed with a tool called Cyclone [19]. Cyclone outputs 3 cell cycle phases: G1, S, and G2M; the scHi-C cell cyclone phases were condensed to match these phases: "early-S" was labeled "S," and "G2," "pre-M," and "post-M" were labeled "G2M." The default parameters of Cyclone resulted in most of the scRNA-seq cells being assigned to the S cell cycle phase, a distribution that was not consistent with the cell cycle labels in the scHi-C data. Since both datasets are from the same cell type, the distributions of cell cycle should be similar. Thus the parameters of Cyclone were adjusted to make the distributions of cell cycle phase match. Specifically, cells with both a G1 score of less than 0.11143 and a G2M score of less than 0.12289 were assigned "S," while all other cells were assigned to "G1" or "G2M" depending on which score was higher.

In addition, we wanted to supervise the expected interaction between cells in scATAC-seq and cells in scHi-C. To do this, we calculated the *assortativity* [20] between each of the scHi-C cells and each of the scATAC-seq cells. This method has the following network interpretation: Given a network with bins in the scHi-C data as the nodes and existence of an overlap between the bins as the edges, assortativity measures the Pearson correlation between the ATAC-seq data corresponding to the source and sink nodes. The idea is that areas in the chromosomes that are touching should have similar accessability.

This metric was calculated at the 100kb level. The matrix of reads and their corresponding cells for single-cell ATAC-seq data was obtained from [21] and cells with total number of reads above 500 were kept. In order to obtain a count matrix of 100kb bins by cells, the mouse genome (mm9) was binned into 100kb regions and all reads corresponding to a particular cell were intersected with bins in order to determine which reads correspond to which bins. This process was repeated for each cell to calculate the full bins by cells count matrix. The scHi-C data was also binned at 100kb level using chromosome contact counts. Assortativity was then calculated using only bins that had contacts in 75% or more of the cells in the scHi-C data and were nonzero in 20% or more of the cells in the ATAC-seq data.

To complete the circle, we needed a way to connect scRNA-seq with scATAC-seq. We chose to use a matrix of correspondences between pairs of cells, based on the correlations between the expression of transcription factors in the scRNA-seq data and the accessibility of the transcription factors' motifs in the scATAC-seq level. To do so, for each cell, we used the scATAC-seq data binned at transcription factor motif level. The single-cell ATAC-seq data was processed to obtain a transcription factor (TF) motifs by cells matrix, quantifying the accessibility of each transcription factor motif in each cell with respect to expected accessibility of the TF motif. ChromVAR [24] was used to obtain the TF motifs by cells matrix from the raw reads collected for each cell. ChromVAR was run with peaks obtained from [25] and using the mouse collection of transcription factor motifs. The z-score deviations from expected accessibility for that TF motif were used in downstream analysis.

For each transcription factor, we selected the motif that was most variable across the cells. Thus we generated a cell by transcription factor matrix. For the scRNA-seq data, we generated a cell by transcription factor matrix based on the scRNA-seq counts for each transcription factor. We took the Pearson correlation between each scATAC-seq cell's transcription factor counts and each scRNA-seq cell's transcription factor counts to generate a matrix of scATAC-seq cells versus scRNA-seq cells.

It was not immediately obvious how to incorporate the generated correspondence matrices between scRNA-seq and scATAC-seq and between scATAC-seq and scHi-C as loss functions in the neural network. Our initial thought was that cell pairs that are highly assortative or correlated should be close together in the latent space. Therefore, the first candidate loss function was to use the Kendall Tau correlation between pairwise distances in the latent space and each correspondence matrix (flattened into vectors):

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{3.1}$$

where n_c is the number of concordant pairs of indices [1] and n_d is the number of discordant pairs of indices between pairwise distances and correspondences, and n is the dimension of the vectors.

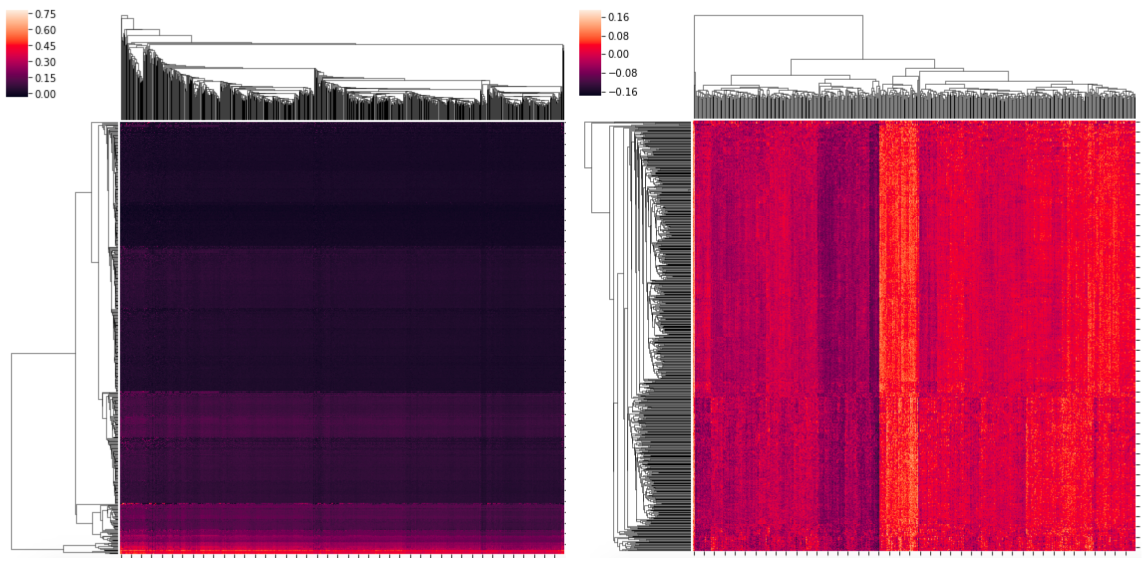


Figure 3-2: Hierarchical clustering of values in scATAC-seq/Hi-C assortativity matrix (left) and scATAC-seq/scRNA-seq correlation matrix (right)

This loss function, however, was problematic in the following ways. First, values in the distance matrices are limited by the triangle inequality while the values in the correspondence matrices are not. The result is that rankings of values that are possible in each correspondence matrix are not possible in each distance matrix. It was also observed that the correspondence matrices were high rank while distance matrices cannot have rank much more than the dimensionality of the space. Secondly, some cells in the correspondence matrices were highly assortative/correlated with all cells while many had low assortativity/correlation with all cells (Figure 3-2). If

[1] Denoting the vector of distance matrices d and vector of correspondences c, a pair of indices (i, j) is concordant if $d_i > d_j$ and $c_i > c_j$ or $d_i < d_j$ and $c_i < c_j$

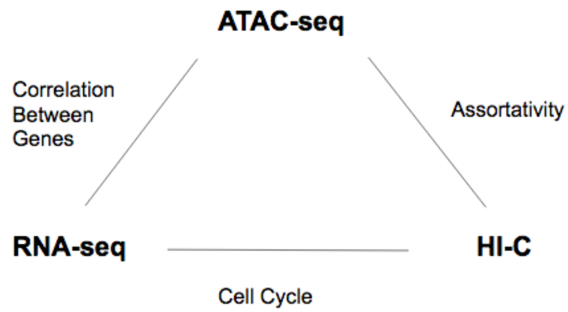


Figure 3-3: Summary of plan to supervise training in the mESC autoencoder network

the rankings of values in the distance matrices were to reflect the rankings in the correspondence matrices, the smallest distances would be between a few cells of type A and all cells of type B and the largest distances would be between a few other cells of type A and all cells of type B, which does not meet our goal of having a latent space in which the distributions of cells of different types overlap. When attempting to use this loss function in the network, it hovered around 0 when ideally it would be negative for both the scRNA-seq/ATAC-seq and scATAC-seq/Hi-C relations, even when the weight for those losses were 100000 times the weights of all other losses.

Therefore, we needed a more flexible loss function that would be more consistent with the geometry of the latent space. Because of the challenge of aligning arbitrary values to distances in the latent space, we chose to use a metric that did not use pairwise distances. We took pairwise Pearson correlations of encoded values, resulting in a scATAC-seq cell by scRNA-seq cell matrix and a scATAC-seq cell by scHi-C cell matrix. We then took the negative Pearson correlations between these matrices and their corresponding correspondence matrix, resulting in two losses, one for each relation.

3.2 Network

To adapt the network from the last chapter, we added a third autoencoder for encoding the scHi-C data and a second discriminator for discriminating the new dataset from one of the others. We established cycle consistency losses for all possible com-

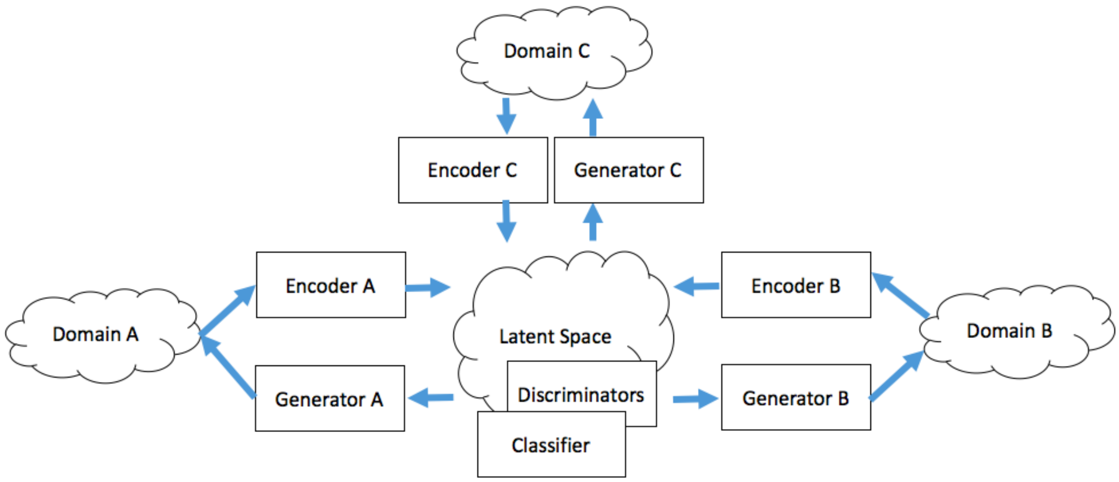


Figure 3-4: Diagram of the autoencoder network adapted for the 3 mouse embryonic stem cell datasets

binations of cycles (6) in the new system. Only one classifier was used. A diagram of this adapted network can be seen in Figure 3-4. A detailed list of all the losses can be seen in Table B.5 in Appendix B. Details on the parameters used in training this network can also be found in Appendix B (Table B.6).

As input to the network we used scRNA-seq gene counts normalized by cell (250 cells, 38658 features), scATAC-seq data binned at the 100kb level (559 cells, 18899 features), and 1Mb scHi-C bin pairs that were non-zero in 25% of cells to reduce input size (858 cells, 10323 features).

3.3 Results

To validate our results, we used a combination of the loss function values, visual images, and the 2i data that was not used for training.

The classification losses of the model were extremely low; as seen in Figure 3-6, the scRNA-seq and scHi-C data are separated by cell type in the latent space. In particular, it is important to note that each of the cell cycle phases between the two datasets overlap. As seen in Figure 3-6c, the G2M scHi-C cells (purple) are near the

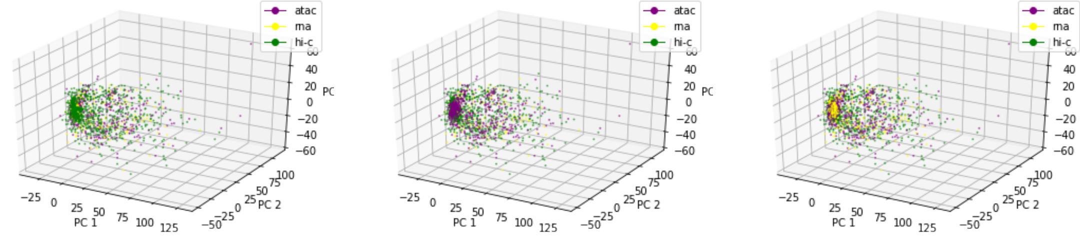
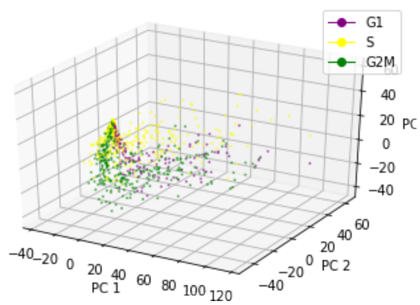
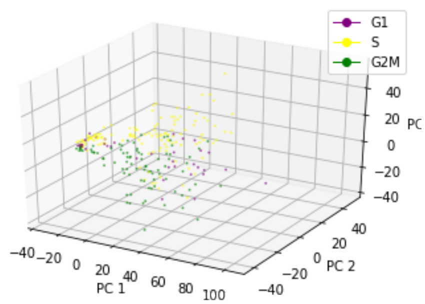


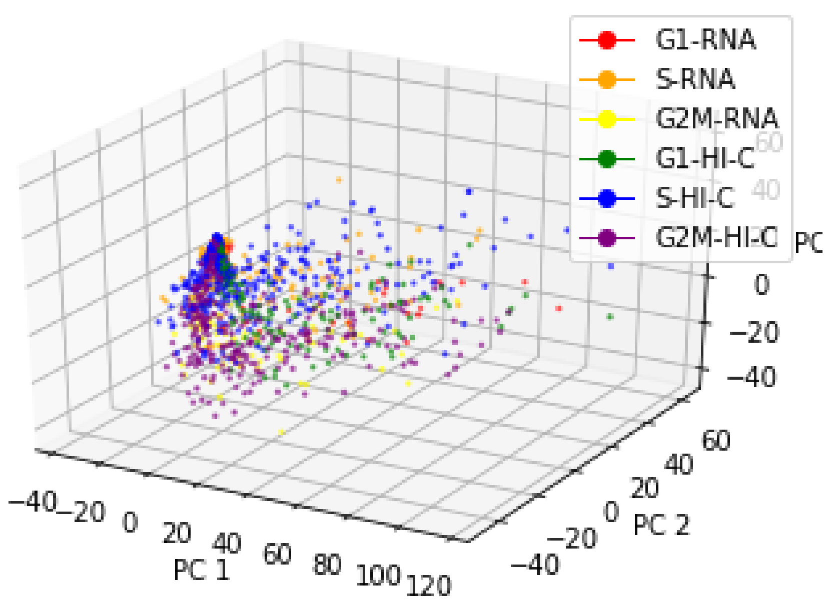
Figure 3-5: PCA plots of mouse embryonic stem cell data projected into found latent space, plotted in different orders



(a) Hi-C



(b) RNA-seq



(c) RNA-seq and scHi-C

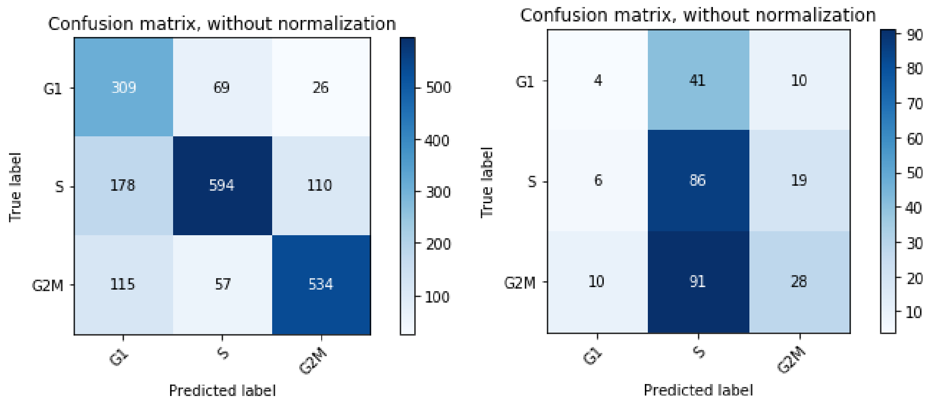
Figure 3-6: PCA plots of mESC data projected into the latent space colored by cell cycle phase

G2M scRNA-seq cells (yellow), the G1 scRNA-seq and scHi-C cells (green and red) are close, and the (blue and orange) S phase cells appear together. Additionally, Figure 3-5 shows that each domain was successfully mapped together, with the distributions of cells in the latent space overlapping.

The correlation losses to associate the scRNA-seq/ATAC-seq cells and scATAC-seq/Hi-C cells were negative, and the losses' decreasing over time shows the network successfully used the data to arrange the latent space. However, the magnitude of the correlation was low. The loss pertaining to the correlation between scATAC-seq/Hi-C assortativity and correlations in latent space was about -0.04 and the loss pertaining to correlation between scATAC-seq and scRNA-seq was about -0.01. During training, the losses seemed to contradict each other; one would go and the other would go down, indicating that perhaps the distances between the scATAC-seq cells implied by the scRNA-seq/ATAC-seq matrix were not in agreement with the distances between the scATAC-seq cells implied by the scATAC-seq/Hi-C matrix. A better set of associations between these pairs of datasets and the losses that go with them is needed. Possible options are included in the Conclusion as future work.

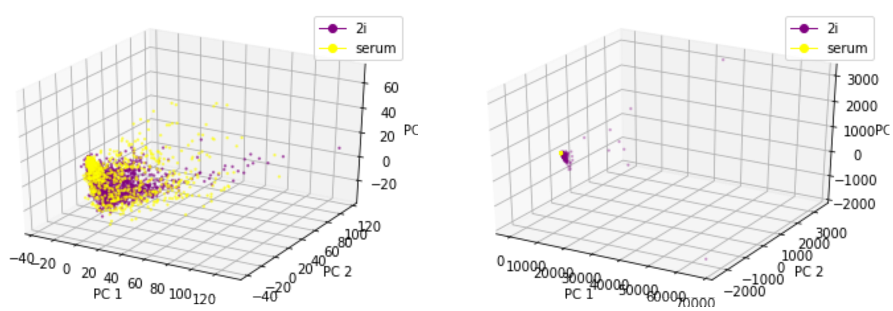
Recall that although we only trained this network on mouse embryonic stem cells cultured in serum, we also had at our disposal scRNA-seq and scHi-C data from cells cultured in a 2i medium. We used these cells to validate our newly found latent space in the following manner: 1) to evaluate our classifier's accuracy at predicting the class labels for each of the 2i datasets and 2) to establish that the 2i data overlaps with the serum training data in the latent space.

Application of the network to data from cells cultured in a 2i medium yielded mixed results, with the scHi-C data performing definitively better. The accuracy of the predicted cell cycle labels for the 2i data is depicted in confusion matrices in Figure 3-8. Overall, the scHi-C data had an accuracy of 72.1% and the scRNA-seq data had an accuracy of 40%. Figure 3-8 shows the 2i scHi-C and scRNA-seq data projected into the latent space. For the 2i scHi-C data, the 2i cells' variability when projected in the latent space is contained in the serum cells', while the scRNA-seq cells contained significantly more variability in the latent space than the serum cells



(a) Hi-C (b) RNA-seq

Figure 3-7: Confusion matrices for predicted cell cycle labels for the cells cultured in 2i medium



(a) Hi-C (b) RNA-seq

Figure 3-8: PCA of latent space when applying the network to 2i cells

(Figure A-7 in Appendix A).

Possible explanations for the poor performance for the scRNA-seq data include: 1) Overfitting. We only had 250 cells for scRNA-seq with a latent space size of 50. In comparison, we had 858 scHi-C cells. 2) Different variation. As discussed in the original paper for this dataset [4], the serum cells had a more even distribution of the cell cycle genes than the 2i cells.

Chapter 4

Conclusion

In this chapter, we discuss findings and future work.

4.1 Findings

In this thesis, we were able to confirm that we can use methods such as paired autoencoders and Deep CCA to map heterogenous single cell datasets, using paired scATAC-seq and scRNA-seq data. We then applied one of these methods to unpaired datasets, and preliminary findings show the acquired latent space displays some biological trends.

4.2 Future Work

While we were able to map the different mESC datasets together using a network of autoencoders, more work needs to be done to confirm the latent space contains relevant biological information relating the datasets. For example, if we can obtain cell cycle labels for the scATAC-seq cells, we can check if the cells originating from the scATAC-seq dataset are correctly classified in the latent space. With binned scRNA-seq data, we can also calculate a similar assortativity metric for the scRNA-seq and scHi-C data as we did for the scATAC-seq and evaluate this new matrix's agreement with the found latent space.

Additionally, more work can be done in calculating matrices of correspondence between the domains to supervise training and the losses applied to them. For example, the scATAC-seq and scHi-C assortativity matrix could be calculated at the 1Mb level instead of the 100kb level to avoid problems related to the sparsity of the scATAC-seq data. The scRNA-seq/scATAC-seq correspondence matrices could be calculated based on the expression of all genes coded for by the transcription factors instead of the expression of the transcription factors themselves, which would be more consistent with the accessibility of the transcription factor motifs. Different types of correlations and distances, instead of Pearson correlation which was used, can be applied in calculating the matrices and the losses.

While the method being used now to map the datasets together seems promising, more work can be done to find the optimal method of doing so. This method may be the autoencoder network with different parameters, Deep CCA, or a completely different method that we have not studied in the scope of this thesis.

Once a good latent space is found and validated, it can be used to explore potential connections between scRNA-seq, scATAC-seq, and scHi-C data, potentially helping unlock mechanisms behind the interplay between gene expression, pluripotency, and chromosome interaction.

Appendix A

Dataset Plots

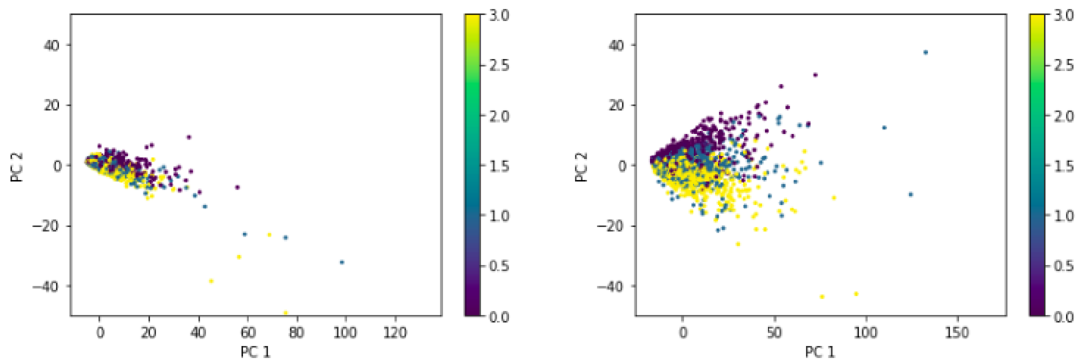


Figure A-1: PCA plot of ATAC-seq (left) and RNA-seq (right) datasets for the Cao dataset

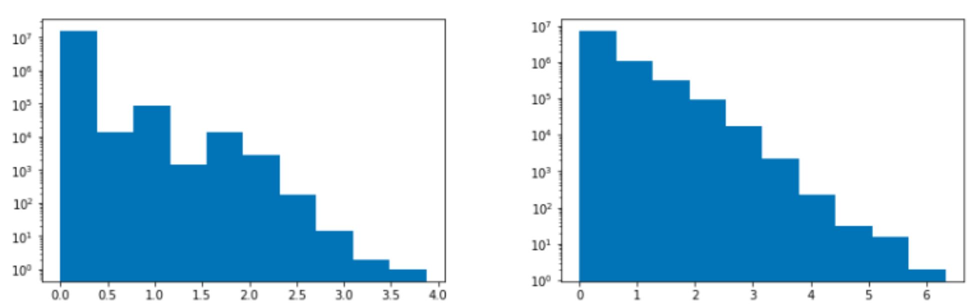
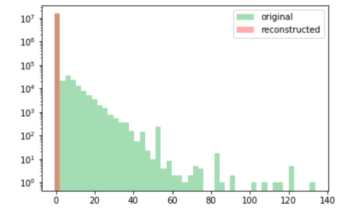
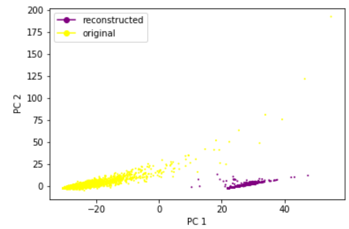


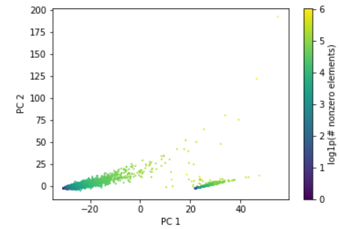
Figure A-2: Log-log histograms of the values in the ATAC-seq (left) and RNA-seq (right) datasets (using just differentially accessible regions and differentially expressed genes respectively). The plots show that both datasets are very zero-inflated.



(a) Histogram of original vs reconstructed values

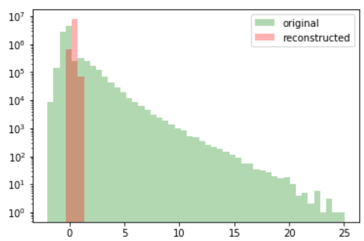


(b) Reconstruction vs original datapoints

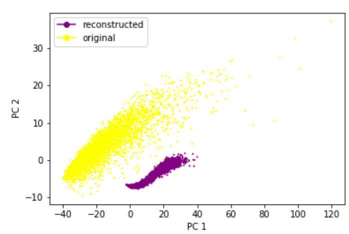


(c) Same as A-3b but colored by sparsity

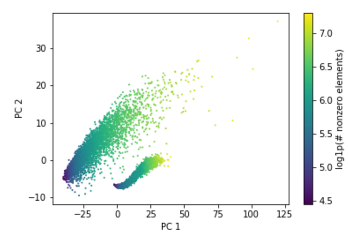
Figure A-3: KL loss autoencoder (ATAC-seq) results



(a) Histogram of original vs reconstructed values

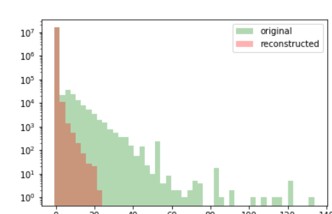


(b) Reconstruction vs original datapoints

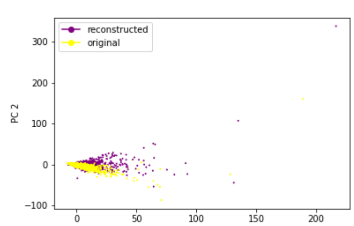


(c) Same as A-4b but colored by sparsity

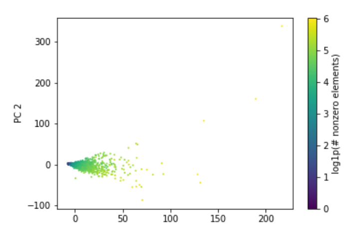
Figure A-4: KL loss autoencoder (RNA-seq) results



(a) Histogram of original vs reconstructed values

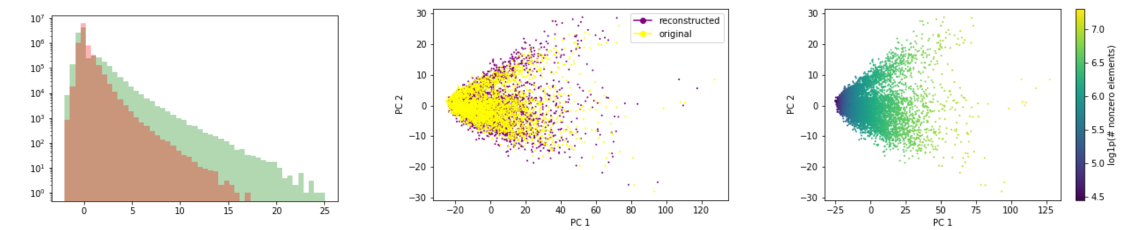


(b) Reconstruction vs original datapoints



(c) Same as A-5b but colored by sparsity

Figure A-5: MMD loss autoencoder (ATAC-seq) results



(a) Histogram of original vs reconstructed values (b) Reconstruction vs original datapoints (c) Same as A-6b but colored by sparsity

Figure A-6: MMD loss autoencoder (RNA-seq) results

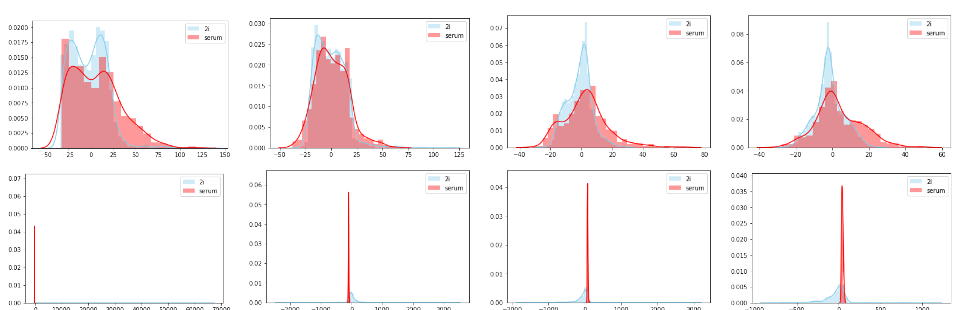


Figure A-7: Distributions of first four principal components of scHi-C (top) and scRNA-seq (bottom) 2i and serum values projected into the latent space

Appendix B

Losses and Parameters

	Epochs	Learning Rate	Batch Size	Lambda
RNA-seq	13	$1 * 10^{-5}$	1	.01
ATAC-seq	5	$1 * 10^{-5}$	1	.01

Table B.1: Autoencoder parameters for the A549 dataset

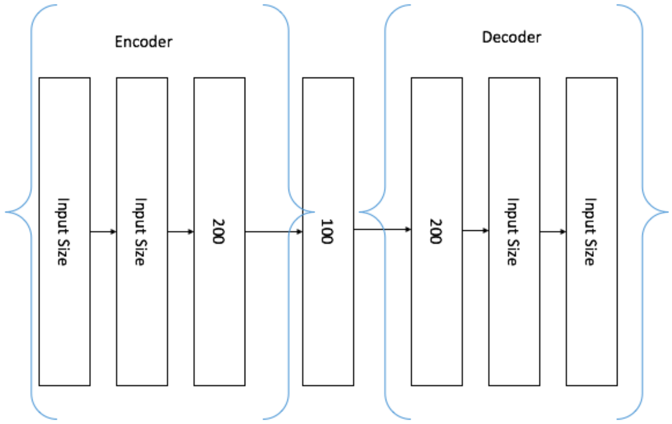
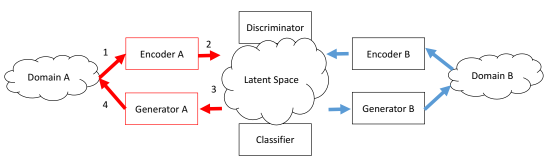
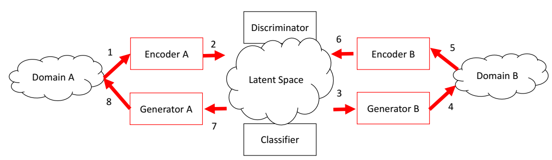


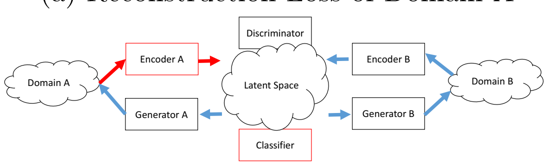
Figure B-1: Architecture of autoencoders which were trained on scRNA-seq and scATAC-seq data seperately. Input size was 2613 for scRNA-seq and 4763 for scATAC-seq.



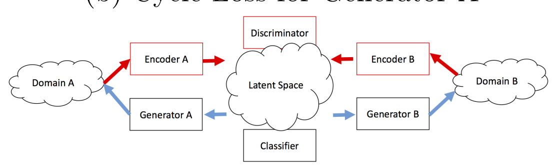
(a) Reconstruction Loss of Domain A



(b) Cycle Loss for Generator A



(c) Class Loss of Domain A



(d) Generative Loss for Generator A

Figure B-2: Illustrations of the data used in each the losses in the autoencoder network. Networks used to transform the data are highlighted in red. Red arrows and numbers are used to indicate the path of the data through the network.

Loss	Description	Type
	Class Losses	
loss_class_a	Treatment time classification loss for the ATAC-seq data	cross entropy
loss_class_b	Treatment time classification loss for the RNA-seq data	cross entropy
	Cycle Consistency Losses	
loss_cycle_a_b	Cycle consistency loss for ATAC->RNA->ATAC cycle	mean absolute error
loss_cycle_b_a	Cycle consistency loss for RNA->ATAC ->RNA cycle	mean absolute error
	Reconstructive Losses	
loss_gen_recon_a	Reconstruction loss for ATAC-seq	mean absolute error
loss_gen_recon_b	Reconstruction loss for RNA-seq	mean absolute error
	Generative Losses	
loss_latent_a	Generative loss for ATAC-seq	mean squared error
loss_latent_b	Generative loss for RNA-seq	mean squared error
	Discriminative Losses	
loss_dis_latent	Discriminative loss	mean squared error

Table B.2: Losses used in A549 autoencoder network

learning rate	epochs	batch size	supervision loss weight	weights for all other losses	latent space size
0.0001	4000	32	0.1	10	50

Table B.3: Parameters used for paired autoencoder networks with supervision

	Input Size	Hidden Layer Size(s)	Output Size
Classifier	50 (latent space)	N/A	3 (class probabilities)
Discriminator	50 (latent space)	50, 100	1
Generator A	50 (latent space)	100, 815, 815, 815	815 (ATAC-seq TF)
Generator B	50 (latent space)	100, 2613, 2613, 2613	2613 (RNA-seq)
Encoder A	815(ATAC-seq TF)	815, 815, 815, 100	50 (latent space)
Encoder B	2613 (RNA-seq)	2613, 2613, 2613, 100	50 (latent space)

Table B.4: Network architecture for paired autoencoder network for the A549 cells. The discriminator, generators, and encoders also had Leaky ReLU activations after each layer.

Loss	Description	Type
Class Losses		
loss_class_hic	Cell cycle phase classification loss for the Hi-C data	cross entropy
loss_class_rna	Cell cycle phase classification loss for the RNA-seq data	cross entropy
Cycle Consistency Losses		
loss_cycle_from_atac_to_hic	Cycle consistency loss for ATAC->Hi-C->ATAC cycle	mean absolute error
loss_cycle_from_atac_to_rna	Cycle consistency loss for ATAC->RNA>ATAC cycle	mean absolute error
loss_cycle_from_hic_to_atac	Cycle consistency loss for Hi-C->ATAC->Hi-C cycle	mean absolute error
loss_cycle_from_hic_to_rna	Cycle consistency loss for Hi-C->ATAC->Hi-C cycle	mean absolute error
loss_cycle_from_rna_to_atac	Cycle consistency loss for RNA->ATAC->RNAcycle	mean absolute error
loss_cycle_from_rna_to_hic	Cycle consistency loss for RNA->ATAC->RNA cycle	mean absolute error
Reconstructive Losses		
loss_gen_recon_atac	Reconstruction loss for ATAC-seq	mean absolute error
loss_gen_recon_hic	Reconstruction loss for Hi-C	mean absolute error
loss_gen_recon_rna	Reconstruction loss for RNA-seq	mean absolute error
Generative Losses		
loss_latent_atac	Generative loss for ATAC-seq	mean squared error
loss_latent_hic	Generative loss for Hi-C	mean squared error
loss_latent_rna	Generative loss for RNA-seq	mean squared error
Correlation Losses		
loss_class_relation_atac_hic	Kendall Tau correlation between (flattened) assortativity matrix and pairwise distances between ATAC and HIC cells in the latent space	kendall tau correlation
loss_class_relation_rna_atac	Kendall Tau correlation between (flattened) correlation matrix and pairwise distances between ATAC and RNA cells in the latent space	kendall tau correlation
Discriminative Losses		
loss_dis_atac_rna	Discriminative loss between ATAC-seq and RNA-seq	mean squared error
loss_dis_rna_hic	Discriminative loss between RNA-seq and Hi-C	mean squared error

Table B.5: Losses used in mESC autoencoder network

learning rate	epochs	batch size	correlation loss weight	weights for all other losses	latent space size
0.0001	160000*	64	1000000	10	50

Table B.6: Parameters used for mESC autoencoder network. *Epochs are counted over the length of the smallest dataset.

batch size	epochs	learning rate	regularization	input size 1	input size 2	hidden layer size 1	hidden layer size 2	output layer size
1024	25	1e-3	1e-9	815	2613	815	2613	50

Table B.7: Parameters used for Deep CCA

Bibliography

[1] C. Uhler and G. V. Shivashankar. Chromosome Intermingling: Mechanical Hotspots for Genome Regulation. CellPress. 2017

[2] Natarajan, K. N., Teichmann, S. A., Kolodziejczyk, A. A. (2017). Single cell transcriptomics of pluripotent stem cells: reprogramming and differentiation. Current opinion in genetics development, 46, 66-76.

[3] Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., ... Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature, 547(7661), 61.

[4] Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., ... Marioni, J. C. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell stem cell, 17(4), 471-485.

[5] Flyamer, I. M., Gassler, J., Imakaev, M., Brandão, H. B., Ulianov, S. V., Abdennur, N. Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature, 544(7648), 110.

[6] Strom, A. R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., Karpen, G. H. (2017). Phase separation drives heterochromatin domain formation. Nature, 547(7662), 241.

[7] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., ... Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398), 376.

[8] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., ... De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture?on-chip (4C). Nature genetics, 38(11), 1348.

[9] Maaten, L. V. D., Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

[10] Ramani, V., Shendure, J., Duan, Z. (2016). Understanding Spatial Genome Organization: Methods and Insights. Genomics, Proteomics and Bioinformatics, 14(1), 7?20. http://doi.org/10.1016/j.gpb.2016.01.002

[11] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10

[12] Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", in IEEE International Conference on Computer Vision (ICCV), 2017.

[13] Welch, J. D., Hartemink, A. J., Prins, J. F. (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics.ăGenome biology,ă18(1), 138.

[14] Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., ... Kelsey, G. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.ăNature methods,ă13(3), 229.

[15] Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., ... Steemers, F. J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells.ăScience,ă361(6409), 1380-1385.

[16] Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., ... Stegle, O. (2018). scNMT-seq enables joint profiling of chromatin

accessibility DNA methylation and transcription in single cells.ăNature communications,ă9(1), 781.Chicago

[17] Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." Advances in Neural Information Processing Systems. 2017.

[18] Andrew, G., Arora, R., Bilmes, J., Livescu, K. (2013, February). Deep canonical correlation analysis. In International conference on machine learning (pp. 1247-1255).

[19] Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., ... Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods, 85, 54-61.

[20] Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B. M., Juan, D., Fraser, P., Spivakov, M., ... Rico, D. (2016). Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome biology, 17(1), 152.

[21] Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., ... & Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. Nature communications, 9(1), 3647.

[22] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[23] Yang, K. D., & Uhler, C. (2019). Multi-Domain Translation by Learning Uncoupled Autoencoders. arXiv preprint arXiv:1902.03515.

[24] Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods, 14(10), 975.

[25] Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., ... & Li, W. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature, 534(7609), 652.